ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΑΞΙΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ ΤΟΥ ΠΡΑΓΜΑΤΙΚΟΥ ΚΟΣΜΟΥ ΜΕ ΕΦΑΡΜΟΓΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Στυλιανός Σωτήρχος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς

Ιούνιος 2024

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμό. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό ό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Μπερσίμης Σωτήριος, Καθηγητής (Επιβλέπων)

- Γεώργιος Τζαβελάς, Αναπληρωτής Καθηγητής

- Πολυχρόνης Οικονόμου, Επίκουρος Καθηγητής

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

Περίληψη

Είναι γεγονός ότι στον 21ο αιώνα, τα δεδομένα αποτελούν θεμέλιο της σύγχρονης κοινωνίας, ενσωματώνοντας την ουσία του ψηφιακού κόσμου και των διαφόρων αλληλεπιδράσεών μας με αυτό. Η συλλογή δεδομένων διαδραματίζει καίριο ρόλο στην ανάλυση και την εξαγωγή πολύτιμων πληροφοριών. Σήμερα, οι τελευταίες τεχνολογικές εξελίξεις έχουν δημιουργήσει τεράστιες βάσεις δεδομένων με την ικανότητα να ανανεώνονται αυτόματα σε σύντομα χρονικά διαστήματα, σε ένα ευρύ φάσμα επιστημών, μεταξύ αυτών και της υγείας. Πιο συγκεκριμένα τα δεδομένα του πραγματικού κόσμου και οι τεχνικές μηχανικής μάθησης διαδραματίζουν ολοένα και σημαντικότερο ρόλο στον τομέα της βιοπληροφορικής, επαναστατώντας στις πρακτικές υγειονομικής περίθαλψης, την έρευνα και ανάπτυξη φαρμακευτικών δοκιμών. Η παρούσα εργασία εξετάζει το πολυδιάστατο τοπίο των εφαρμογών αυτων των δεδομένων στο κλάδο της υγείας, διερευνώντας τις ευκερείες και τους κινδύνους που μπορεί να ελλοχεύουν. Εξερευνώντας τις προκλήσεις που παρουσιάζουν τα πραγματικά δεδομένα, αναλύουμε στρατηγικές για την αξιοποίηση τους, μέσω από οργανισμούς που προσπαθούν καθημερινά να αναπτύξουν την διατήρηση και χρήση αυτων των βάσεων δεδομένων. Επιπλέον, εξετάζουμε πώς τα πραγματικα δεδομένα διευκολύνουν την ανάπτυξη φαρμάκων μέσω καινοτόμων τεχνικών μοντελοποίησης και πώς επιτρέπουν την εξατομικευμένη ιατρική μέσω μεθόδων όπως η κοινωνική ακρόαση. Επίσης, η εργασία συζητά την μελλοντική πορεία της ενσωμάτωσης αυτων του είδους δεδομένων, τονίζοντας τον ρόλο του στη βελτιστοποίηση της κλινικής ανάπτυξης και στη μετάβαση από τα δεδομένα σε εφαρμόσιμες πραγματικές γνώσεις παρουσιάζοντας πληροφορίες από τη βιβλιογραφία της φαρμακευτικής έρευνας, υπογραμμίζει τη σημασία της μηχανικής μάθησης στην εξαγωγή πολύτιμων πληροφοριών, αναλύοντας διάφορες μεθοδολογίες από την ταξινόμηση έως τη μείωση διαστάσεων. Συνολικά, η εργασία αυτή ρίχνει φώς στο μεταμορφωτικό στην αξία των πραγματικών δεδομένων και της μηχανικής μάθησης και στη διαμόρφωση του μέλλοντος της υγειονομικής περίθαλψης.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

EXPLOITING REAL WORLD DATA APPLYING MACHINE LEARNING TECHNIQUES

By

Stylianos Sotirchos

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
June 2024

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Sotiris Bersimis, for his invaluable guidance, support, patience and encouragement throughout the course of my research and thesis writing. Their expertise and insight have been instrumental in shaping this work, and I am profoundly grateful for their mentorship.

I would also like to extend my sincere thanks to the members of my thesis committee, for their constructive feedback, suggestions, and for taking the time to review my work. Their contributions have greatly enriched this thesis.

My heartfelt appreciation goes to my family for their support and understanding throughout my academic journey. To my parents, Stamatis & Vaso, thank you for always believing in me and providing me with the foundation to pursue my dreams.

To my friends, thank you for your companionship, and for the joy and motivation you brought into this journey. I am particularly grateful for your feedback and support over the nights, which has been invaluable in refining my work and pushing me towards the finish line.

A special thanks to my sister, Matoula, whose support has been exceptional. Your encouragement, patience, and faith in my abilities have been a constant source of strength and motivation for me. I couldn't have achieved this milestone without your love and support.

Thank you all.

Abstract

In today's era, data stands as the cornerstone of modern society, encapsulating the essence of the digital landscape and our myriad interactions within it. The collection of data marks a pivotal role in the analysis and extraction of invaluable insights. Today, technological advancements have birthed vast databases across a spectrum of disciplines, among them healthcare, each reservoir of information fostering the growth of knowledge and innovation. Real-World Data and Machine Learning techniques are increasingly pivotal in the field of bioinformatics, revolutionizing healthcare practices and pharmaceutical research and development. This paper delves into the multifaceted landscape of RWD applications, navigating through its distinctions from real-world evidence and its regulatory implications. Exploring the challenges and opportunities presented by RWD, it scrutinizes strategies for harnessing its potential, such as the Observational Medical Outcomes Partnership framework and Health Technology Assessment. Furthermore, it investigates how RWD facilitates drug development through innovative modeling techniques, and how it enables personalized medicine through methods like Social Listening and Quantitative Systems Pharmacology. Additionally, the paper discusses the future trajectory of RWD integration, emphasizing its role in optimizing clinical development and transitioning from data to actionable real-world insights. Presenting insights from library literature in pharmaceutical research, it underscores the significance of ML in extracting valuable insights from RWD, detailing various methodologies from classification to dimensionality reduction. Overall, this paper illuminates the transformative potential of RWD and ML in advancing bioinformatics and shaping the future of healthcare.

Table of Contents

1 CHAPTER

Introduction

1.1 Initial Framework

The present thesis aims to investigate and analyze the role of Real-World Data (RWD) and the applications of them to improve the modern healthcare sector. The study begins with an overview of the current landscape of machine learning and its intersection with health sciences, setting the stage for a comprehensive analysis of how these technologies can revolutionize patient care. RWD refers to health-related information collected outside the context of randomized controlled trials, encompassing a wide range of sources such as electronic health records, electronic medical records, medical claims, and patient registries. While Real World Evidence (RWE) is derived from the analysis of RWD and provides insights into patient experiences and treatment outcomes in real-world settings. This distinction is crucial as RWE informs clinical and regulatory decisions, enhancing the relevance and applicability of healthcare research.

The thesis elaborates on the roles of major regulatory bodies, including the European Medicines Agency and the Food and Drug Administration. These organizations leverage RWD to support drug development and approval processes, ensuring that new therapies are both safe and effective. The study also introduces the Observational Medical Outcomes Partnerships and Health Technology Assessment initiatives, which aim to harmonize health data and facilitating more robust and reliable knowledge of evidence. Additionally, the European Health Data & Evidence Network is discussed for its efforts in integrating diverse health data sources across Europe. Also, one of the critical applications of RWD examined in this thesis is its contribution to the development of pharmacokinetic, pharmacodynamic, and models. These models are instrumental in predicting drug behavior and optimizing therapeutic strategies, ultimately leading to more personalized and effective treatments.

Beyond the foundational concepts, this thesis includes a comprehensive review of the existing literature in pharmaceutical research and development. It explores various machine learning techniques and their applications in advancing healthcare. Specific examples are provided within the fields of arthroplasty, acoustics, cardiology, and the COVID-19 spectrum. These case studies illustrate how machine learning models can be utilized to enhance research outcomes, optimize treatment strategies, and improve patient care. By presenting these examples, the thesis demonstrates the practical implications and transformative potential of integrating machine learning with real-world data in diverse medical domains.

Subsequently, the thesis examines the intricate relationship between machine learning and statistical methods, highlighting how machine learning serves as an overarching framework that incorporates and enhances traditional statistical techniques. This chapter provides an in-depth analysis of the primary categories within machine learning: supervised learning, which includes classification and regression methods, and unsupervised learning, which encompasses clustering methods and dimensionality reduction techniques. The discussion covers some of the most widely

used methodologies, such as logistic regression, linear regression, and support vector machines. Additionally, the chapter explores artificial neural networks, detailing their architecture and applications in processing complex datasets. By thoroughly examining these techniques, the thesis underscores their significant role in advancing healthcare research and development.

Building on this theoretical foundation, the thesis progresses to the practical application of these methodologies. In a dedicated chapter, we apply our knowledge to a dataset of diabetes participants, utilizing various machine learning techniques and artificial neural networks to predict the likelihood of diabetes. Diabetes stands as one of the most prevalent metabolic disorders, casting its shadow over millions worldwide. The term "diabetes" traces its origins back to the Greek physician Aretaios, who coined it from the Greek verb "διαβαίνω», meaning "to pass through." Aretaios observed that individuals affected by this condition experienced excessive urination, causing fluids to pass through the body undiluted, leading to the characteristic symptom of frequent and copious urination. Thus, the name "diabetes" aptly captures the essence of the disease as a condition where fluids pass through the body in an unrestrained manner, highlighting its defining feature in ancient medical terminology. This empirical analysis aims to demonstrate the effectiveness and comparative performance of different approaches in identifying individuals at risk of diabetes, thereby providing valuable insights into their practical utility and potential impact on healthcare outcomes.

Finally, the thesis presents the general outcomes of the analysis and discusses how RWD can revolutionize modern healthcare. By exploring various machine learning techniques, the study underscores the potential of these methods to enhance early disease detection and improve healthcare outcomes.

2 CHAPTER

Introducing the Real-World Data

2.1 Defining Real-World Data

Nowadays, it's no secret that the most popular trend in the health and life science companies is known as Real-World Data (RWD), or someone may even hear it as Real-World Evidence (RWE). The ability to extract information and expertise from new types of data, mining technologies or electronic health records, is now possible and it gains more and more popularity. With the use of them, we can detect side effects, or long-term outcomes, as well as the costs, advantages, and risks of medical treatments, by Togo and Yonemoto (2022).

The pharmaceutical companies are under unprecedented pressure mostly because of austerity measures and price reductions. Manufacturers are required to give data on clinical and financial value in addition to safety, appropriate usage, and effectiveness. Although Randomized clinical trials (RCTs) are widely regarded as the cornerstone of clinical tests, Eduardo Valencia (2017), the generalizability of results from RCTs is constrained by things like different responses to a drug in real life, failure to finish a prescription, or taking unapproved medication before or during the trial.

One of the most widely cited definitions of RWD originates from the field of Pharmacoeconomics Garrison et al., 2007, the ISPOR (International Society for Pharmacoeconomics and Outcomes Research) defines as:

«Data used for decision making that are not collected in conventional randomized controlled trials. »

In the spectrum of bioinformatics, we frequently employ the terms "Real World Data" and "Real World Evidence" interchangeably, yet it is essential to note that they do not carry identical meanings, the term "Data" refers to «factual information, serving as raw material», while "Evidence" refers to «that the organization intends to utilize the information to formulate a conclusion or judgment» Eduardo Valencia (2017). According, to the United States Food and Drug Administration (FDA), RWE is the proof obtained from combining and analyzing real-world data parts. On the other hand, RWD is information gathered from sources other than conventional clinical trials. As a result, we set up the analytical procedures that enable us to turn data into evidence.

FIGURE 2.1 : *Explaining the RWD&RWE connection (source: https://www.meaningcloud.com/blog/real-world-evidence)*

The term RWD is frequently used in the healthcare industry to describe patient-level data acquired outside of the typical clinical trial environment. Such information may be created during routine clinical treatment, the processing of administrative claims, or it may come directly from patients. Illustrative instances encompass information sourced from a variety of outlets such as patient charts, clinical reports, prescription refills, history of patients treated both on- and off-label, data derived from multiple clinical trials, survey responses, and inputs from mobile health devices. Additionally, other pertinent data from established secondary sources are utilized to inform decisions related to safety, quality, care coordination, coverage, and reimbursement, in

The amount, sources, and use of RWD have increased dramatically because of evolutionary breakthrough in technology, data science and healthcare policies, with the gathering of bigger and a wider variety of data sets. The increasing adoption of digital technologies, such as mobile devices, wearables, sensors, social media and online patient communities, has not only introduced new data sources but has also enhanced the methods for capturing, storing, and analyzing longitudinal RWD pertaining to patients. Massive variety and complexity characterize the present RWD landscape as we can see and from the (Figure 2.2) below.

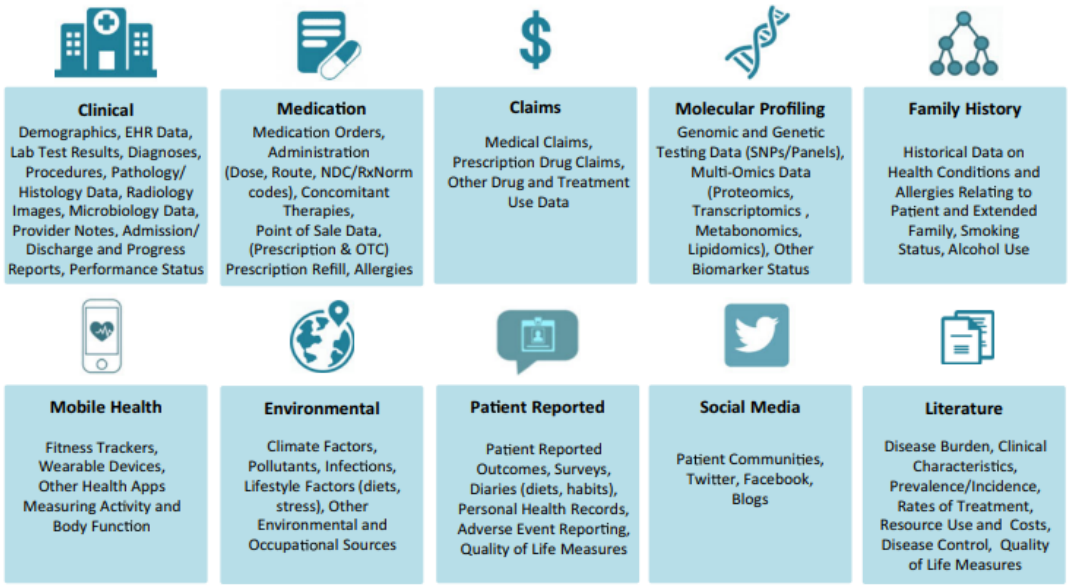

FIGURE 2.2 : Types of RWD (source: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01768-6)

It includes both organized and unstructured data from a variety of heterogeneous sources, going beyond conventional sources like chart reviews, prescriptions, or claims data. Also, patients themselves can fill reports about the therapy results and prospective registries. Additional unique sources of previously inaccessible patient-level data can be found in mobile health devices or, wearable apps. These technologies that we further examine later, provide capabilities for continuous monitoring, data collecting, and real-time transmission that are uncommon in ordinary clinical care as we knew it, Izmailova (2018).

2.1.1 Electronic Health Records and Medical Records

In the healthcare industry, healthcare information systems (HIS), often known as electronic health or e-health, are used to improve patient data, individual patient experiences, and drug information. The conversion of paper-based health and medical data into electronic formats marked only the beginning of the digitization of healthcare services. But thanks to technological breakthroughs like big data, the Internet of Things, and 5G networks, patient empowerment is now a crucial component of e-health, where patients actively engage in their healthcare decisions and activities. Electronic Health Records (EHRs) are digital versions of patients' paper charts and are real-time, patient-centered records that make information available instantly and securely to authorized users. They contain the medical and treatment histories of patients and are designed to go beyond standard clinical data collected in a provider's office and can be inclusive of a broader view of a patient's care. In the following paragraphs, we will examine further the terms of EHRs and Electronic Medical Records (EMRs), which often cause confusion when referring to these two e-health formats Anshari (2019).

To begin with, EMRs are digital medical records that include patient information, medical conditions, histories, checkup reports, medications, and treatment records. Although they are easily stored, the data in EMR does not always easily leave the organization. It's possible that the person's record would have to be copied and sent to experts or other organizations. Thus, EMR is not so different from common paper files.

On the contrary, EHRs are digital health information of the person as it contains much more than what is already included in EMR, Anshari (2019).. To be specific, EHRs are a superset of EMRs containing retrospective, concurrent, and prospective patient data in digital form, stored securely so it can be sharable to authorized users to support integrated health.

Overall, EHRs aim to gradually increase efficiency, improve the quality of care, promote evidence-based medicine, and predict new relationships between patients and healthcare. Although we should mention that access to EMRs and EHRs by patients is limited in many cases, as they are not primarily intended for customer accessibility. The upcoming shift in healthcare aims to change this as it wants to start treating patients as partners in care, and healthcare organizations need tools and strategies to manage and empower these relationships.

2.1.2 Real World Evidence

To get back on topic, the outcome of «real-world data analysis» is known as «real-world evidence» and it is used to develop insights utilizing suitable research design and scientific methodologies to assist healthcare stakeholders in making decisions. Therefore, producing evidence from RWD requires not only gathering "big data" but also skillfully fusing these various and frequently unrelated kinds of data to provide insightful conclusions. For instance, a smartphone could count the meters traveled by a user to examine the fitness level of the user. In contrast, Miksad & Abernethy, 2018 asserted to exclude data from clinical research settings such as in EHRs from their definition of RWE. In the same way, the definition of real-world evidence was most

recently expanded by the FDA in a publication, which stated that any data "generated from any study design including RCTs as long as the data source is from routine care and the design is highly pragmaticT, U.S. FDA (2018), meaning the trial design and conduct closely approximate the eventual use of the product in clinical practice" is now considered RWE.

As we previously mentioned the majority of RWD are not always gathered for research objectives, the data collection is episodic, reactive, and sometimes provides partial information at best. Real-world evidence is therefore typically messy and sparse, necessitating the use of rigorous and reliable statistical techniques to clean the data and fix errors. Precision treatments in oncology, where vital information regarding molecular biomarkers or end-points data can frequently be lacking, require careful data curation employing both structured and unstructured data. Linking to additional data sources may be needed to complete any missing information in the data.

Analysts need to pinpoint and account for confounding factors, including demographics, socioeconomic, insurance, severity, concomitant medications, and genetic predispositions to specific conditions, prior to conducting additional analyses, Swift (2018). RWE is always susceptible to selection bias, given that cohort selection and treatment decisions in clinical practice are non-random. Consequently, adhering to established guidelines for the design and validation of RWE studies can mitigate some of the sources of bias and inconsistencies.

All major markets are heavily emphasizing the use of RWE to assist and enhance healthcare decisions. Nevertheless, the acceptance and utilization of RWE for decision-making vary globally. The FDA has a longstanding interest in gaining more insights about medical items, particularly drug safety, using healthcare data produced in the real world, Breckenridge (2019). Considering this, the FDA introduced the Sentinel project in 2008, which enables RWD from sources like EHRs, insurance claims data, and registries to be used for monitoring the safety of FDA-regulated products. To offer insight into recent developments, we present a visual representation (Figure 2.3) highlighting key moments that underscore the significance of RWE over the past few years.

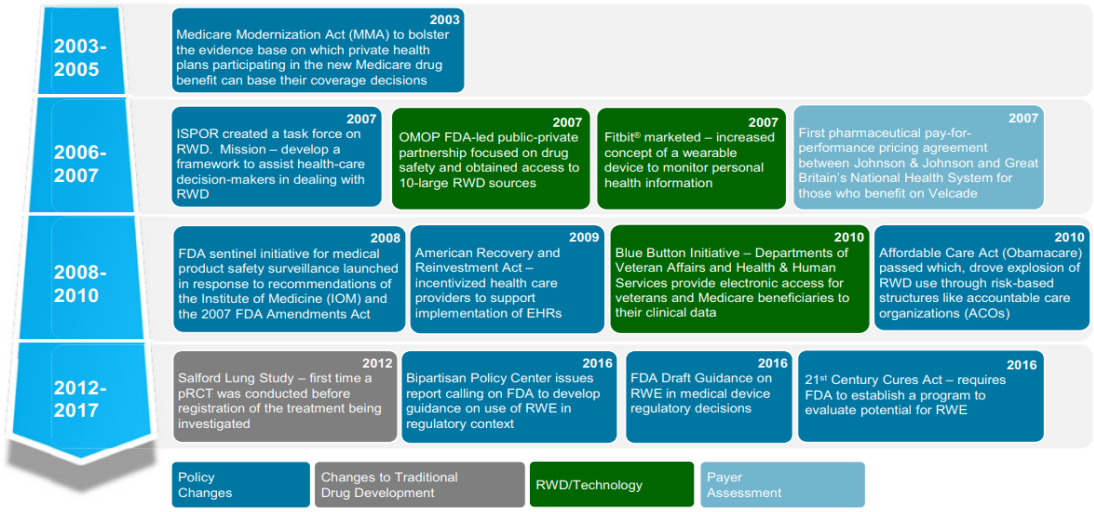

FIGURE 2.3: Most important breakthroughs in the RWE in the recent-past *(source: https://pubmed.ncbi.nlm.nih.gov/29768712/)*

2.1.3 Regulatory approvals in the healthcare industry

In today's rapidly evolving healthcare landscape, regulatory oversight plays a pivotal role in ensuring the safety, efficacy, and accessibility of medical products Togo and Yonemoto (2022). The importance of regulatory agencies cannot be overstated, as they are tasked with evaluating and monitoring medicinal products, pharmaceutical innovations, and novel therapies. As we delve deeper into the evolving landscape of healthcare regulation, we'll explore the steps that have already been made for the legal and transparent development of therapies and medicines, and how these changes hold promise for the future of healthcare in general.

A. The role of European Medicines Agency

The European Medicines Agency (EMA) serves as a decentralized agency within the European Union, primarily tasked with the scientific evaluation, supervision, and safety monitoring of medicinal products across the EU. It was established in 1995 and is headquartered in Amsterdam, Netherlands. The EMA (formerly EMEA) played a significant role in the regulation of biosimilar medicines by creating the legal framework and regulatory approval pathway in 2005. When the bio-originators' patents and exclusivity periods end, the creation of medications can expand the biotherapeutic industry and enhance patient outcomes by making biological therapies more accessible. And one year later, in 2006 by approving the first biosimilar, Omnitrope (somatotropin). Biosimilars (BS) are biological medicines that are extremely like an already-approved biologic, known as the reference product (RP) Gherghescu & Delgado-Charro (2021). They are known as similar biotherapeutic products (SBPs) by the World Health Organization (WHO), which defines them: «highly similar to an original biotherapeutic product».

The Pharmaceutical industry has witnessed a surge in the process of BS, and various regulatory bodies worldwide have authorized these medicines. To be specific, EMA approved 55 BS over 13 years, while the FDA only approved 26 in the past 5 years, Gherghescu & Delgado-Charro (2021). The developmental pathway for BS encompasses comparative Phase I and case-by-case Phase III clinical studies, aiming to establish the similarity of the two molecules with respect to their pharmacokinetic and pharmacodynamic properties. It’s important to note that, EMA works by evaluating the scientific data submitted by pharmaceutical companies in support of their applications for marketing authorization for new medicinal products while also monitoring the safety of medicinal products after they have been marketed, to protect public health if any safety concerns arise. Yet, the genuine economic and clinical advantages of these medications will only be known when they achieve broader availability in the market and become seamlessly integrated into clinical practice.

Furthermore, for BS to make a meaningful difference in patient healthcare and healthcare systems, they need to become widely available and gain the trust of doctors who prescribe them. The growing BS market, along with the pharmaceutical industry's strong interest in developing these medicines and changing regulations, looks promising for the future of BS. Nevertheless, additional research is necessary to meticulously evaluate the tangible impact of these medicinal products on the broader

public health landscape and patient outcomes, particularly as BS market penetration experiences ongoing global expansion.

B. The role of Food and Drug Administration

The value of RWD and RWE for the FDA and the health sciences sector has increased since the «21st Century Cures Act» was passed into law in 2016. The Cures Act prioritizes RWE- and RWD-driven decision-making to hasten the development and innovation of medical products. RWE is defined by Congress as information about a drug's use and any potential advantages or dangers from sources other than clinical trials for instance, randomized trials and observational studies.

RWD, as defined by Congress, is information about patient health status and care delivery that is regularly gathered from EHRs, claims and billing, patient-generated data, and other sources. With the increased usage of computers, mobile devices, and wearables, RWE and RWD are expanding in both volume and depth. They are also becoming more useful as new analytics capabilities, such as AI and machine learning, offer more individualized and useful insights.

In 2018, the FDA took a proactive stance by publishing the Framework for FDA's Real-World Evidence Program. This document intricately outlines the utility of RWD in trials for new therapies. The evolving focus of RWD is shifting towards an innovative approach, encompassing pragmatic trials and synthetic cohorts, Breckenridge (2019). The adoption of pragmatic trials is particularly noteworthy as they have the potential to significantly reduce both costs and timelines. Life science companies that adeptly embrace these methodologies are poised to emerge as frontrunners in their respective fields.

As several organizations seek to enroll patients in both conventional research and cutting-edge studies that use data-driven methodologies, key RWD & RWE challenges are emerging surrounding access to health systems and patients, Stukpa (2019). Also, life science businesses must ensure that breakthroughs in clinical development assist health systems reach goals, which means they must do more than just use data to their advantage, they must also clearly demonstrate value for both patients and providers. Overall, clinical strategies that comes together with clinical trials standards of care, will certainly lead us to the correct direction. Lastly, selecting trials that both the patient and the health industry could benefit from, would be beneficial in the future.

2.2 Strategies for confronting the dangers of Real-World Data

In the previous section, we presented some important terms of RWD, RWE and the roles of regulatory bodies such as the EMA and FDA in shaping healthcare decisions. Building on this foundation, we now examine HTA which extends the principles of RWE, as a comprehensive form of policy research that evaluates the social, economic, and ethical impacts of health technologies. This section will explore its role in guiding policy recommendations and its integration into regulatory frameworks. Additionally,

we will address the challenges and dangers associated when dealing with RWE and RWD, and the efforts to harmonize health data using frameworks like OMOP. Through this exploration, we aim to provide a deeper understanding of how HTA contributes to the sustainability and improvement of healthcare systems globally.

2.2.1 Health Technology Assessment

It is well known that throughout history, physician assessment has had an important role in the selection of therapy of the patient. The term HTA was used by the U.S. Congressional Office of Technology Assessment (OTA) in 1972. The general definition of technology assessment used was: ‘‘a comprehensive form of policy research that examines the short- and long-term social consequences of the application or use of technology’’. In the health field, OTA established that assessment would symbolize ‘efficacy’, since the goal of health care is to improve health. The scope of this definition highlights how pervasive technology is and how it can be beneficial as we merge it in our everyday life with the right way. HTA can be seen as an extension of RWE, as it utilizes data to assess the effectiveness and safety of health technologies. RWE provides a rich source that reflects the actual experiences of patients and healthcare providers, enabling a more comprehensive and realistic evaluation of health interventions. Furthermore, HTA can be also considered an umbrella term of Health Economics studies as it encompasses economic evaluations, such as cost-effectiveness analyses, and a broader analysis of the social and ethical implications of health technologies. Through this approach, HTA helps policymakers make informed decisions that promote efficiency, equity in healthcare and address the growing costs of it by identifying the most effective and efficient treatments, ultimately contributing to the sustainability of health systems globally.

In its early reports, OTA invested plenty of time and work on defining the field and principles of HTA. Simultaneously, OTA conducted numerous evaluations of health technology, starting with short case studies, and progressing to longer assessments. Since its beginnings, the primary use of HTA has been in the formulation of policy recommendations. The problem of medicine regulation provided an attractive model for official policymaking. For instance, to market an innovative medicine in the US, a business must apply to the FDA, which then authorizes it to conduct human drug testing Banta, (2002). The results are submitted to the FDA when the experiments, which traditionally have been well-controlled studies, are finished.

After evaluating the data, the agency determines if the medication can be marketed or not. Thus, assessment is integrated into the regulatory and policy-making process. It might be utilized, like blood products, vaccinations, and medical devices even if most of these areas haven't been subject to as strict regulations as the pharmaceutical industry. Nevertheless, the approach applies less well to medical and health care practice. The primary substitute has been to connect payment decisions with assessment. Increasing research, for instance, demonstrates that health insurance decision-making is becoming more connected to official evaluation in both the United America and Europe.

In recent years, there appears to be a growing interest in social and ethical issues, according to the latest HTA reports. HTA is a wide concept with a variety of features

and blurry borders from country to country both in its focus and method. Probably a significant part of the variations in HTA by country has depended on the goals of societal groups by Banta, 2002:

1) **Policymakers:** Broad problems, like the value for income perspective.
2) **Insurers:** Overriding worries for expenditures and their management.
3) **Clinical physicians:** Generally interested in quality, not much attention to costs or other public policy issues.
4) **Epidemiologists and other researchers:** Interest in the poor state of research and the ways that can evolve it, including attention to systematic reviews and dissemination of information.
5) **Industry:** Overriding worries for profits, however, competition forces growing interest in efficacy and cost-effectiveness.
6) **The general public:** Access to personal care of acceptable quality.

Differences in HTA from country to country hamper its development internationally. An HTA can be a technical assessment of a medical device completed for regulatory purposes, a pharmacoepidemiologic study of a drug conducted or funded by industry with the primary goal of receiving reimbursement, an academic study of the potential health effects of a specific medical practice, such as a randomized trial, or a systematic review of any or all facets of a specific medical practice conducted by an HTA agency. Although this variety has advantages, it also makes generalization hard, and hampers change.

The study that follows assessed different challenges, related to this increasingly complex environment of new health technologies, make the acceptance of RWD most likely, Hogervorst et al.2022. Thirty-three HTA organizations that are members of EUnetHTA were given the questionnaire, with the twenty-two answering, (67%) total, from twenty-one different countries. The questions centered on approved data sources, conditions that made RWD acceptability possible, and obstacles to acceptability. To begin with, acceptance of RCTs was reported by all. The distribution of organizations across Europe was even, with the Nordic region slightly overrepresented. Moreover, pharmaceutical assessments were the responsibility of twenty-one responding organizations (95%) out of which nine (41%) assessed just pharmaceuticals. The questionnaire devoted to RWD covered mainly three topics: the general willingness to use and accept RWD, probability to accept RWD challenging circumstances and the obstacles to approve RWD.

First, respondents could highlight whether they experienced the need for broader systematic use of RWD and whether they experienced a willingness among assessors, among decision makers or even both, as both affect the ultimate reimbursement conclusion Hogervorst et al. (2022). Second, the respondents indicated the categories of data sources that their organization approved for evaluation in a binary form (yes/no). Third, using a 5-point Likert scale, respondents stated how likely they were to accept RWD sources in specific difficult situations. Respondents were allowed to explain any further situations they came across that would enable the acceptance of RWD in HTA in the fourth and only open question, guaranteeing the questionnaire's confidential nature. Finally, a list of obstacles to accepting RWD for HTA was ranked by the respondents.

Regarding the willingness to use and accept RWD sources in HTA, the results showed that 18 out of 22 representatives indicated that there is a need for wider systematic use of RWD in HTA decisions than in the current stage. Furthermore, sixteen representatives said they sense willingness from both assessors and decision-makers, while two said they sense willingness from assessors alone, and one only from decision-makers. The remainder indicated no knowledge about the willingness at all. The top three approved data sources for HTA in Europe were found to be the traditional sources like “meta-analyses”, “systematic reviews”, and “RCTs”. It’s important to note that all three were embraced by all participating organizations. On the other hand, the least popular RWD sources were “case reports”, “unpublished data”, “editorial and expert opinions”, which were all accepted via only one third of the organizations.

We then look at how likely people are to embrace RWD under challenging circumstances. With a score of 4.3 on a 5-point Likert scale, the evaluation of “orphan drugs or other treatments with small patient populations” was shown to be the scenario most likely to accept RWD sources. Organizations would be less inclined to embrace RWD in HTA if the data originated from countries other than their own, even if it represents the sole available data source. Among “orphan drugs” and employing “RWD from outside its country's region”, the first and last rated situations we can see small difference in probabilities of accepting RWD all above 3 (the scores varied from 3.2 to 4.3). Based on the open question mentioned earlier, seven companies reported additional situations in which RWD may be approved. Several stated that RWD would not be received as the sole source of evidence, though could be supplementary to traditional RCT evidence. Overall, the findings show that, in these situations, there is a greater tendency for positive than opposing views regarding RWD acceptance.

After examining the section concerning “Barriers to accepting RWD”, the responses from the HTA organizations this time displayed significant diversity. As shown in (Figure 2.4), an organization's average mean rank for the most significant obstacle to embrace RWD in HTA was “lacking necessary RWD sources”. Also, this was followed by “existing policy structures or information governance,” which had a mean rank of 3.5. On the contrary, “lack of statisticians or other relevant analysts” and “financial reasons” came in last. The exception is “lack methods to use RWD,” where there is an uneven distribution of rankings. Additionally, “No possibility to interpret or verify data, or that it was challenging to do so” has the smallest variance among all reasons.

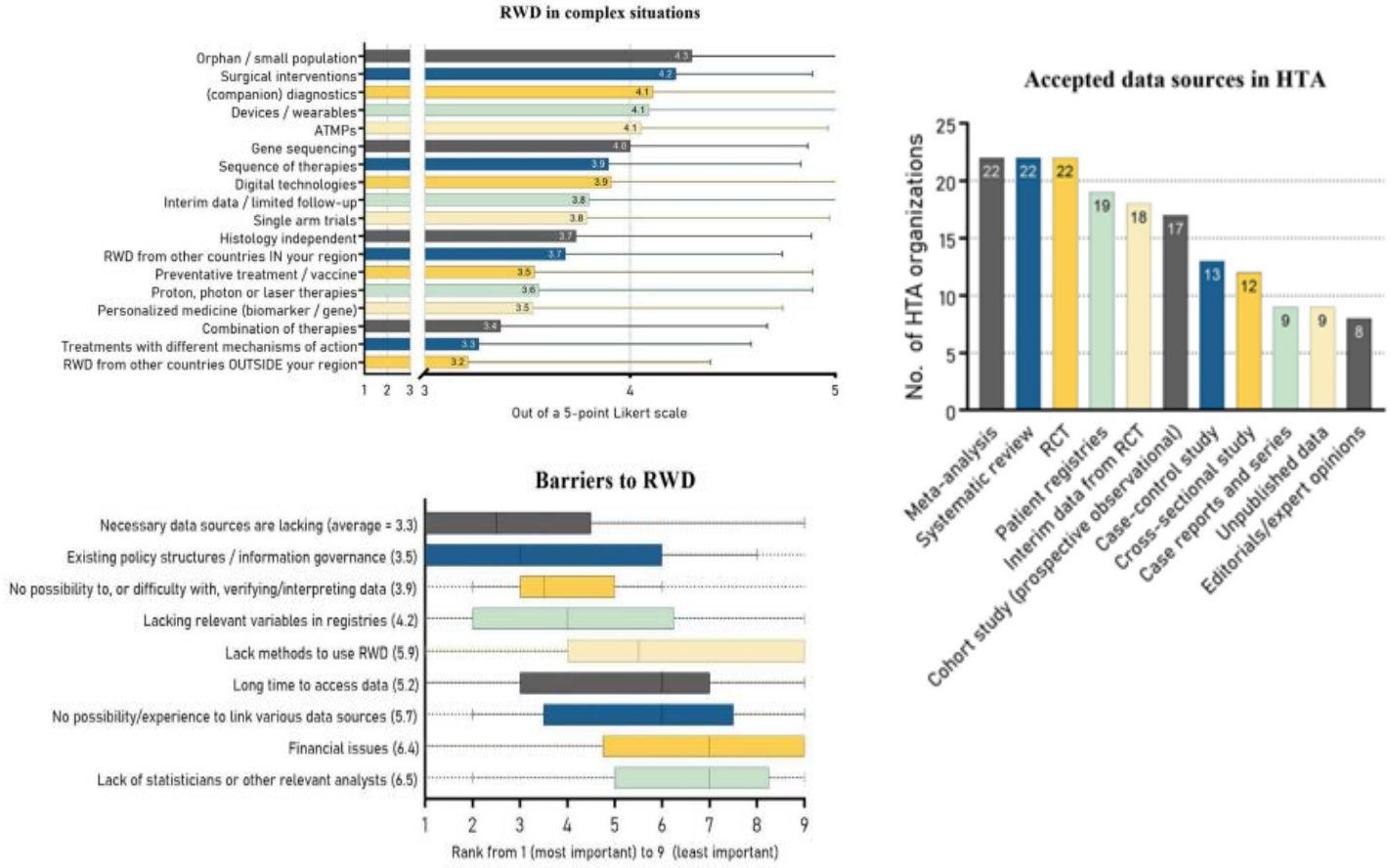

FIGURE 2.4: (A) Accepted sources in HTA, N = 22 (B) Mean Likert-scores, of the possibility to accept RWD in emergency circumstances (C) Boxplots of the barriers to accepting RWD in HTA, in order of their median score (source: https://www.frontiersin.org/articles/10.3389/fphar.2022.837302/full)

2.2.2 The obstacles emerge when deriving RWE from RWD

After explaining the vital role of EMA and FDA we will try to answer the one question that comes to all of us when we previously talked about the link of RWD with RWE. For a comprehensive grasp, it is crucial to provide insight into the business-related risk exposure associated with RWD. The risks related to RWD can be classified into three main groups as it is simplified in (Figure 2.5). These aggregated risks are related to "Compliance Controversies," "Registration Failure," and "Business Model Disruption," as examined throughout time and ranging from short to long term, Schneider (2019).

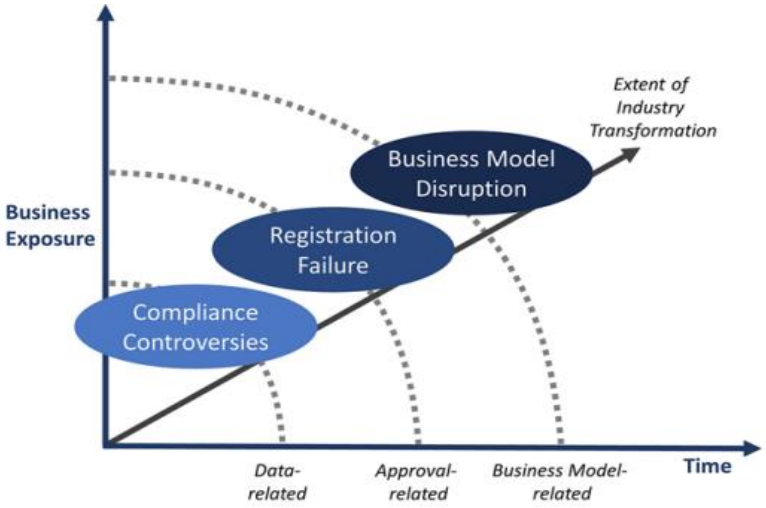

FIGURE 2.5 : *Risk fields in the context of RWD (source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8339486/)*

The «data-related risks» that could result in compliance violations, for instance when specific requirements cannot be met during an audit, are addressed in the inner layer of the image. The model's intermediate layer focuses on "approval-related" risks associated with the development of new medicines and therapies, for instance when certain process improvements created using RWD are insufficient. The outer layer of risk categorization is "business model related," and it deals with the potential of RWD to disrupt the life science industry, for instance, when new competitors enter the market with superior RWD, Grimberg, et al. (2021) for instance, from health apps combined with wearables, like smartwatches.

The grouping of RWD hazards into three "layers" helps to comprehend how RWD can be exposing the health sciences industry to an ever-increasing business exposure over time. However, this abstracted model does not intend to overlook or omit risks that may arise approval-related risks may be one of the firsts, following later in time multiple risks connected to data clarity.

We will now examine and categorize the main difficulties risk managers are likely to experience when dealing with RWD. Next, we provide a criteria schema (Figure 2.6), in a form of a Radar, that will assist in our comprehension, Grimberg et al., 2021. The foundation of the RWD Challenges Radar rests on three pivotal categories, delineating the landscape within the information systems discipline, recognized as «confluence of people, organizations and technology» Grimberg, et al. (2021). Each category has several subcategories that were determined through the literature search. In fact, the size of each subcategory is meant to represent the importance or range of the current discussion surrounding that topic, the broader the field, the more significant the topic seems to be.

As we already know nowadays, risk is a part of business, and in a world where tremendous amounts of the data are being processed at increasingly rapid rates, we can safely say that identifying and minimizing risks is a challenge for any organization Robert B. Hirth Jr. (2017). Since the radar is an abstract illustration of a very specific part of RWD risk, this specific focus will be helpful to unveil the challenges associated.

FIGURE 2.6: *RWD Challenges Radar (source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8339486/)*

Additionally, based on the underlying observations of the Radar, a RWD Challenges Cockpit can be developed. Such a dashboard-type solution would automatically capture, classify, assess, and visualize the quality of certain RWD. The use of the RWD Challenges Radar fits the different stages of the drug development process and will allow the RWD users to be fully aware of the challenges and risks related to the data while they fully utilize the potential of RWD. In the following sections, three main categories and related parameters of the RWD Challenges Radar are discussed:

- Expertise:

To rely on evidence from RWD, one must first understand the data, analyze it and extract vital information that can be found useful in a decision-making process. However, research shows that these skills are not only in "abundant supply within the pharmaceutical industry"measure , but in domain knowledge, healthcare information technology, and methodological and technical knowledge as well Wise et al. (2018). An additional survey where interviews were healthcare stakeholders, stated that there is a shortage of expertise in the RWD analysis domain, giving the example of "innocent misinterpretation" in which analysts misunderstood relationships as causality. It is crucial to note that an excellent understanding of accessible databases supports the assumptions of the validity of these databases. Research also mentions the absence of elite education on data analytics and the insufficient research in the challenges that RWD may be facing. However, several projects try to combine information technology

and medical skills to examine various databases and promote productive communication across other HCPs.

❖ Bias:

Accurate evidence generated from RWD must be high quality and without any form of bias. The selection bias is still recognized as the most well-known and difficult danger that is preventing the adoption of RWD, Grimberg, et al. (2021), even when the quality is confirmed, and privacy concerns are resolved. Previous studies revealed proof of reporting bias in several disease areas, such as depression, bipolar disorder, and several others through refusing study data of drug manufacturing and regulatory bodies. As a result, bias has been a problem in data analysis for many years.

In other words, additional types of bias, such as information bias, may become apparent in observational studies, in addition to choose and reporting bias. The FDA has stated within this context that randomization serves as the preventive measure against bias by ensuring a balance in study groups concerning risk factors for the intended outcome.

❖ Complexity:

The diversity in data formats among different sources and nations poses aslso a technological challenge that impedes the progress of RWD advancement. We already talk about the FDA recognizing the importance of having a common data model, along with the standard representations like coding schemes and common terminologies, to maximize the utility of RWD. Several companies, like the «Institute for Clinical and Economic Review», have already begin asking the organizations to provide RWD in certain structure with the goal to maximize the integration of different types of data. In comparative research, combining data formats from observational databases can be helpful to find the causes of observed effects as well as give answers to related issues. The Observational Health Data Science and Informatics has introduced a common data model called Observational Medical Outcomes Partnership (OMOP), which enables a distinct database to be carefully analyzed.

2.2.3 Harmonizing health data for the future.

RCTs, or explanatory trials, as we already have mention, generally measure efficacy, that is, the benefit a treatment produces under ideal conditions, often using carefully defined subjects in a research clinic and aims primarily to further scientific knowledge Mahajan (2015). RWD can influence RCT design from a scientific, what might be measured, for example, and logistical, where can we locate the patients to participate in the trial, perspective. As a hypothetical scenario, if one wanted to test a candidate therapy to address a particular medical need, they would need to find potential subjects who had that need, perhaps within a particular age range, with inclusion and exclusion criteria based on medical history and current medications. To make it easy to understand, RCTs is research in which various similar people are randomly assigned to two, or more groups to test a specific drug, treatment, or other intervention, Wise, et al.

(2018). One group, the experimental group, has the intervention being tested, the other, the comparison or control group, has an alternative intervention, a dummy intervention, placebo, or no intervention at all. On the other hand, Practical trials measure effectiveness, or the advantage a treatment has in everyday clinical practice. The design of a pragmatic trial reflects the variations between patients that occur in real clinical practice and aims to inform choices between treatments.

Nevertheless, evaluating the effectiveness of different therapy combinations would require multiple research efforts if RCTs were employed. To analyze all conceivable combinations of just five therapies, 32 distinct study arms would be required. The costs would be unaffordable under such a strategy. 'Platform trials' allow numerous treatments to be evaluated simultaneously and 'offer flexible features such as dropping treatments for futility, choosing one or more treatments superior, or adding brand-new treatments to be tested throughout the course of a trial Mahajan (2015).

The significance of reusing health data for research has been clearly shown by the IMI project EH4CR. The EHR4CR project (Figure 2.7) has developed a robust and scalable platform that can make use of deidentified data from hospital EHR systems, in full compliance with the moral, legal, and data safety regulations and requirements of every single involved nation. The EHR4CR created the 'The Champion Programme', which sets out to establish the value of RWD for clinical research and has also developed i~HD its sustainability model is in the European Institute for Innovation through Health Data.

FIGURE 2.7 : *The Electronic Health Records Systems for Clinical Research (source: https://www.imi.europa.eu/projects-results/project-factsheets/ehr4cr)*

Also, the European Health Data and Evidence Network (EHDEN) aspires to become the reliable system of observational research to support improved health care, Almeida

(2018). The goal is to establish a new standard for the gathering, analyzing, and sharing of health data across Europe by creating a huge, sustainable, federated network of sources that are standardized to a specific data model. At the main domains of EHDEN first is the OMOP Common Data Model (CDM), to unify data structure and semantics and facilitate cross-source analysis and second encourage the use of analytical tools developed by the Observational Health Data Sciences and Informatics (OHDSI), Quiroz (2022) which could start a huge collaboration in the field.

OMOP is a framework for conducting observational studies and evidence-based research using large and diverse health datasets, including EHRs, administrative data, clinical registries, and more. EHDEN's challenge is to harmonize 100 million anonymized health records from multiple sources. The adoption of services and tools that carry out data standards, enable data discovery, and carry out analytical pipelines, as well as facilities that promote the sharing of study results, support this effort, Almeida (2018). Efficiency, transparency, reproducibility, and scalability are probably the main factors influencing the conversion of various health data types to OMOP CDM. This will help, cut expenses, and produce breakthroughs faster than anticipated. It also makes it possible to reuse analytical tools, compare results across different data sources without exchanging raw data, and conduct studies across numerous locations.

Although there is an urgent need to convert datasets to OMOP, doing so takes a lot of time and resources, hence the attention of researchers demands tools for mapping data to OMOP the best way possible. Furthermore, it is imperative to highlight that, owing to challenges related to data dictionaries and structural disparities, the utilization of OMOP has not been as extensive as its potential merits. This constraint is compounded by the prevalent issue of data loss that often occurs during the conversion process, Quiroz (2022).

It is certain that more work is required to build confidence in its use and encouragement by the industry to motivate more examples for it to be effective in 100% of its potential. With a standard data format, data interoperability and analytics across numerous data sets become more effective, allowing new worries to be resolved. Lastly, both the FDA and EMA have already expressed a great deal of interest in OMOP and CDM, and regulatory expectations of enforcement may present one chance to expand its use.

2.3 How the Real-World Data contribute to the art of drug development

Its widely recognized particularly as highlighted in by Stupka (2019) that drug development process spans 8 to 15 years, with costs reaching up to $11 billion. Typically, it heavily relies on an expensive clinical trials process, yielding sparse and costly data, often claims data, and above all that still fails to offer a comprehensive view of patient health. For instance, while claims data may reveal a filled prescription, it lacks insights into side effects, or other information that may be crucial.

By collaborating with a healthcare transformation firm to access and utilize expanded RWD and RWE, life science companies may enhance this process and reduce

costs while better understanding the populations utilizing their products and their outcomes. Today's healthcare sector has a significant opportunity since, for the first time, major industry actors are aligned on the same crucial objectives. Thus, the urgency to provide the appropriate treatment to the appropriate patient, as determined by real-world results and monitoring, is being driven by regulatory, financial, and reimbursement demands. This means that overcoming identical problems such as identifying patients with lower risk and highest benefit from treatment "X" or managing patient population to drive overall balance between clinical and financial outcomes can be beneficial for manufacturers, payers, and providers equally, Stupka (2019). Moreover, predict, identify, minimize, monitor, and measure drug safety problems or even ensure high standards of treatment adherence, patient education, support, can maximize outcome potential. That's why we will examine next the role of some models and techniques which can be ideal for this role. Moreover, we discuss about the new terms that we all keep hearing by the name of Precision Medicine and Social Listening and in what way the new technologies can play a significant role in the evolution of this technique.

2.3.1 Pharmacokinetic–Pharmacodynamic–Pharmacoeconomic models for early predictions

As we already know traditional clinical trials and regulatory approval processes target primally on "does the drug actually work?" under an ideal design. This is reasonable, but it might not give enough details about how the medication performs in various situations (such polypharmacy or comorbidities) and with numerous patient subpopulations. As a result, there is now a greater emphasis on including RWD in healthcare choices as well as in the creation and marketing of new medications. Pharmaceutical companies are also under constant pressure to show the value of new treatments in the context of their habitual use due to the increasing attention on value-based pricing.

In the subsequent paragraphs we discuss mainly for PK-PD and PE models although these terms are usually use together, they are not the same. Moreover, Pharmacokinetic models (PK) are defined by how a compound is absorbed, distributed, metabolized, and excreted, while Pharmacodynamic models (PD) are all about the measure of a compound's ability to interact with its intended target leading to a biologic effect. In addition to the above, Pharmacoeconomic models (PE) are usually sets of equations that identify the economic factors that influence the prices and sales of imports and competing domestic products in the industry. It is more crucial than ever, to evaluate the worth of new medications early in the research using cutting-edge predictive techniques like (PK-PD-PE) models. Through the use of PK-PD modeling to guide go/no-go decisions by integrating adherence rates from RWD as we show in the next (Figure 2.8), dose and response data from RCTs were modeled using the PK-PD modeling approach, Swift et al. (2018).

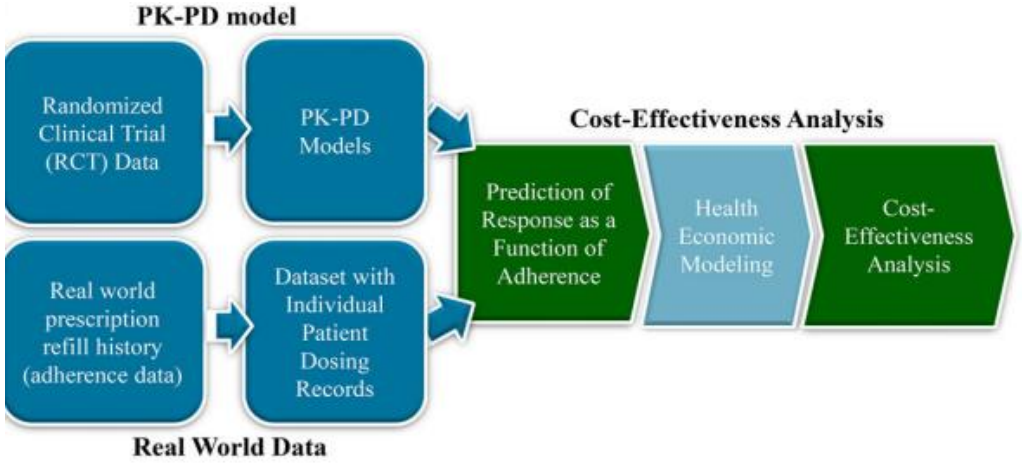

FIGURE 2.8 :*Illustrative application in the context of medication adherence (source:* https://pubmed.ncbi.nlm.nih.gov/29768712/*)*

Separately, RWD on patient adherence from a large database with prescription refill history was turned into a patient-level database at the individual level by making some assumptions to bring it into the same format as the PK-PD database to enable simulations using the PK-PD modeling program. To replicate the clinical responses under various levels of adherence in the real-world scenario, the PK-PD model based on RCT was then applied to the individual patient database built from the prescription refill history. In order to determine the level of adherence improvement that would lead to clinically significant improvements in clinical outcomes that are also cost-effective, these response forecasts were input into a health economic model, Swift (2018).

These methods can be used to evaluate the demand for further information as well as the economic benefit of therapeutic measures aimed at enhancing adherence. Before a medicine is made available to a larger population, predictive analytics may be the only way to predict how it will operate in actual-life scenarios. To promote realistic future value during clinical development, such model-based outputs for future performance can be efficiently used.

Furthermore, the ability to collect data in both a healthy and diseased condition could enable generation of new hypotheses during clinical trials. For instance, real-time tracking of activity, vitals, sleep, speech, and other factors can be help us to better understand both beneficial and harmful effects of the selected medications. It can also be used to demonstrate how different medications differ from one another. Pharmacometric (quantitative pharmacology) methods are used in model-based clinical drug development to optimize compound development processes. As we have already mention, this can be done through the combination of PK, PD, and clinical data using empirical or mechanism-based modeling to predict efficacy and safety outcomes from simulated clinical trials. With the aim of lowering late-stage failure and increasing the effectiveness of drug development, such methods have been used to explore the effects of various dosing regimens, consider particular populations, adjust for nonadherence and dropout, and provide useful insights for upcoming studies.

Linking with PE models (Figure 2.9) that take resource restrictions of payers of health care into account is a natural extension to pharmacometrics analyses, utilizing the structural relationship between dose and response and accounting for statistical uncertainty. These models can all be categorized as PK-PD-PE models. The incremental costs per quality-adjusted life year (QALY) obtained for a particular intervention are evaluated by PE models. If the cost per QALY is below a predefined threshold (often £30,000; €50,000; $100,000 per QALY in the United Kingdom, EU and USA, respectively) then the intervention is considered cost-effective. Someone can find more interesting facts in this field if just have a look at, Pink et al., (2012).

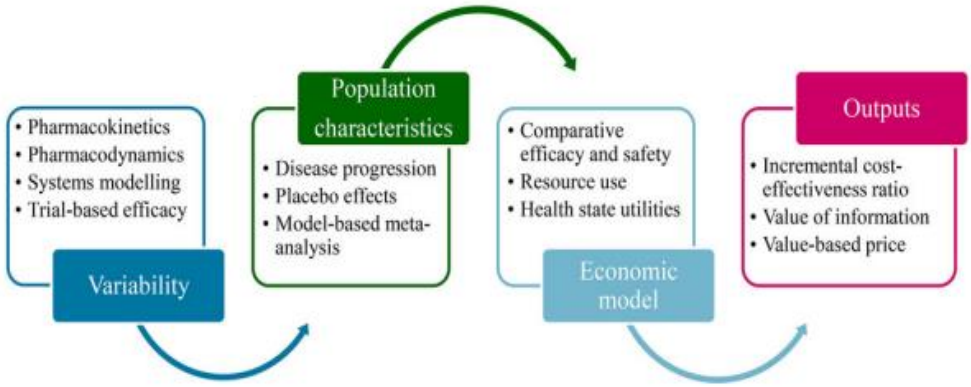

FIGURE 2.9 : Simple *skeleton of a pharmacometric-pharmacoeconomic model (source:* https://pubmed.ncbi.nlm.nih.gov/29768712/*)*

This evaluation methodology offers advantages over conventional (empirical) PE models during phase II and phase III of clinical drug development and has the potential to enhance approaches for strategic, clinical, and pricing choices. If there is no chance of attaining a value-based pricing, it may be appropriate to stop the development of a medicine. For example, the assessment of a value-based price will indicate whether continued development is commercially viable. As an alternative, one might perform a value of information analysis while utilizing the PK/PD uncertainty and economic parameters generated by RWE. This can assist in go/no-go decisions, providing insights into whether undertaking a trial is justified by evaluating the expected net trade-off between the trial's benefits and costs.

PK-PD-PE models, exemplified by applications like rituximab for lymphoma and guided warfarin dosing, demonstrate potential benefits. While still in its early stages, the use of pharmacometrics in pharmacoeconomic evaluation shows promise, especially in early-stage assessments using RWE. This approach offers early insights into cost-effectiveness, guides future research, assesses subgroups, informs strategic decisions, and estimates the cost-effectiveness of complex interventions like pharmacogenetics testing.

However, overcoming different modeling paradigms in pharmacometrics and health economic evaluation, the requirement for increased acceptance of model-based drug development through pricing and reimbursement to inform critical-stage decision-making, and the need for additional evidence on the validity and reliability of complex and computationally intensive models are key obstacles to the further advancement of PK-PD-PE model development and application. These obstacles could be removed by

fostering closer integration and cooperation between the disparate fields of clinical pharmacology, health economics, and outcomes research. This could be done by colocating the relevant experts and providing them with the necessary training. In the long run, this should result in a broader acceptance of model-based drug development, notably RWE and PK-PD-PE integration in the drug development process.

2.3.2 QSP: Bridging the gap between Biology and Pharmacology

We continue our research by exploring the quantitative systems pharmacology, or else QSP modeling, that aims to help us understand the biological system and disease, facilitate early and more thorough in silico testing of drug candidates, and lastly support rational decision making to cut both development cost and time. In other words, QSP is an umbrella term for modeling approaches that combine a mathematical representation of the biological system with pharmacological information regarding a medicine of interest to facilitate enhanced understanding of human pharmaceutical reactions, Wise, et al. (2018). What is the difference though with the PK-PD-PE modeling we previous talk about?

PK-PD-PE modeling is a set of interconnected models that focus on different aspects of drug development and utilization. PK deals with how drugs move within the body, including absorption, distribution, metabolism, and excretion. PD focuses on the relationship between drug concentration and its effects on the body. PE involves evaluating the economic aspects of drug use, such as cost-effectiveness and health outcomes. QSP on the other, is a modeling and simulation approach that focuses on understanding the underlying biological and physiological processes in a quantitative way. It targets to integrate data from different sources to build comprehensive models of drug actions and disease progression. QSP models are often used to predict the effects of drugs on complex biological systems and can be beneficial in optimize drug development strategies.

QSP aims to enrich the process to detect and examine targets, reveal possible biomarkers, support drug design, tell dose and regimen selection, and aid proactively identify responders and non-responders. In more straightforward terms, QSP is used both for drug discovery and development, with a focus on understanding the mechanisms of action, optimizing dosing regimens, and predicting clinical outcomes. Although, it is very promising, QSP is a potent translational science tool that requires quantified patient data.

However, RWE, derived from RWD, at the moment mostly providing contextual input, with qualitative insights into disease, epidemiological patterns, and effects of therapy. Some people in the industry already envision more quantitative data sets to come directly from patients in the future via electronic devices, namely wearables (like the IMI RADAR-CNS project). Because this type of data is quantitative and can be utilized for modeling and simulation in complex illness models, it will considerably expand what can be accomplished with RWD. This kind of data is required for QSP modeling in both preclinical and clinical pharmacology in order to them to flourish.

2.3.3 The importance of Social Listening and Precision Medicine

The practice of health care is advancing from classifying and then treating patients based to coarse-grained and traditionally defined illnesses to a stratified medicine, or cohort-based medicine, or more typically but somewhat misleadingly titled ‘personalized medicine’, which is based on identifying subgroups of patients with distinct mechanisms of disease, or responses to treatments, Wise, et al. (2018). Precision medicine, ‘an arising approach for disease treatment and avoidance that takes into consideration human being variation in genetics, surroundings, and lifestyle for each, is gradually being viewed as the strategic path for medical care. Because of underlying variations in their genes, different patients respond to the same medications differently.

RWD, along with its associated methodologies of QSP can be exploited better to understand these vital molecular biology differences among patients. If the biopharmaceutical industry can deploy classified medicine and precision medicine approaches that prove to the regulator and the consumer a therapeutic worth proposition that identifies those patients who will respond to a treatment and, conversely, those patients who will not respond then the effectiveness and the safety will upgrade its value, and the potential of a listing in the pharmacopoeia enhanced.

To identify individuals who will respond well to specific medications, the Royal Marsden Hospital in London is examining the viability of employing genetic sequencing, Wise et al. (2018). Up to 200 patients with advanced gastrointestinal tumors, including as stomach, pancreas, intestine, and esophageal cancers, are being recruited for its FOrMAT research. Additionally, this experiment intends to build the technical and logistical infrastructure necessary for the routine easy identification of patients in this manner. Genomics England's 100,000 Genomes Project is an illustration of how RWD/RWE is applied to further knowledge in both science and clinical practice that way patients with rare diseases or cancers have their genomes investigated. A closer understanding of this data can result in more accurate diagnoses, which in turn can lead to more precise treatment approaches.

To make sure that data stored in EHRs which by nature itself can be chaotic can be mined effectively, two approaches are now being promoted Wise et al. (2018). The first is expanding the structure in such documents while using controlled vocabularies while the second has to do with using Natural Language Processing (NLP) to unstructured or semi structured health data. In this part we ought to make clear that improvements in NLP depend on the amount and quality of biomedical data sets use to train such algorithms. For instance, smart watches, Fitbits, and other wearable devices can collect a plethora of patient generated RWD without requiring the patient to enter a clinical care setting. The capacity to combine data from many wearables and identify trends has already altered the way we gather data, but there are still no clear standards that would enable seamless interoperability and data exchange for such wearables.

The mining of social media, or social listening, has the potential to give beneficial information into a patient’s real-life use of therapeutics and combine all of these. Biopharmaceutical companies, regulators, and others are now actively researching social listening as a unique and extra technique of pharmacovigilance. Social listening is frequently used by commercial organizations to acquire brand insight. These data's accessibility and immediate nature could offer information on negative outcomes and

non-compliance. For instance, social listening can be used to monitor the most prevalent causes of patients changing treatments, Risson et al., 2016.

Overall, this method can with no-doubt described as a powerful, low-cost, real-time data source, but there are still some restrictions on how the information may be utilized. Although there are several problems like data validation, source verification, managing the complexities of large unstructured datasets, discerning meaningful signals amidst noise, and interpreting free text that includes elements like misspellings, and local dialects. While these challenges are not insurmountable, they do call for robust applications of natural language processing and advanced AI-tools. Nevertheless, to fully leverage the substantial benefits offered by social media data, the establishment of universally accepted practices concerning privacy and regulatory guidance becomes imperative.

2.4 Leveraging Real-World Data to maximize the clinical development results

For the majority of the previous decades, clinical drug development inefficiencies resulted in higher costs for the pharmaceutical industry from research to launch as well as worse overall outcomes for patients who were subjected to a rigid, artificial clinical trial environment (e.g., being placed in a placebo arm for the trial's design's sake rather than the patient's sake or being excluded from promising trials due to the complexity of trial designs) by Elia (2019). Life science firms can add significant value to their pipeline (Figure 2.10) by utilizing extended RWD/RWE, which enables them to spend less, increase the speed of clinical studies and time to market, develop a digital marketing plan that is more intelligent or even improve drug matching with patients. To get the point, we attach a visual representation of the subject matter at hand.

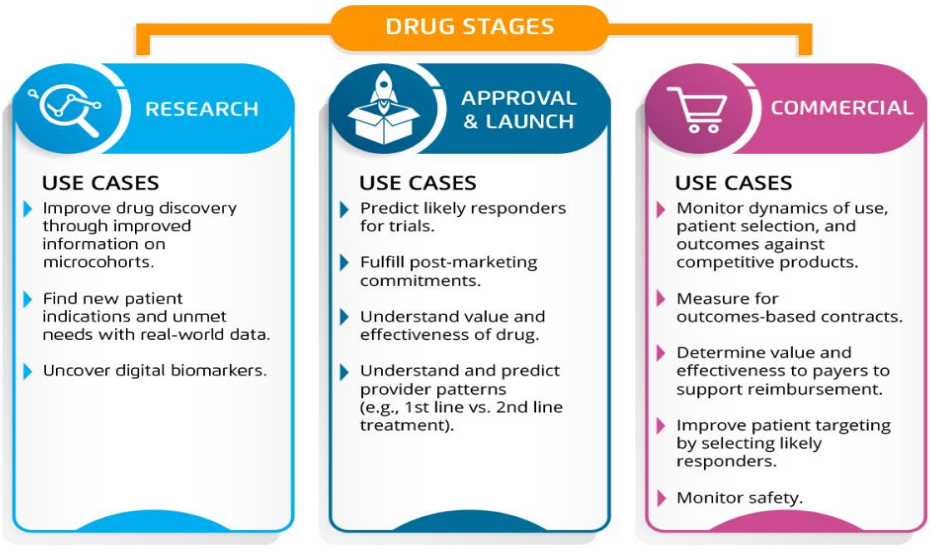

FIGURE 2.10 : Data uses across the life science pipeline *(source: https://www.healthcatalyst.com/insights/real-world-data-chief-driver-drug-development)*

RWD presents an opportunity to innovate in the traditional drug development paradigm changing from the RCT, to gain regulatory approval to an all-encompassing collection of real-world evidence in the context of a therapeutic solution Karamehic (2013). Other future perspectives for leveraging RWD, is in the traditional drug discovery/development model. In the conventional model, obtaining regulatory permission for a "pill-in-a-bottle" is the first step, and then real-world factors are considered (Figure 2.11). The suggested new approach makes use of big data innovations like sensor devices, technologies, imaging, and other pertinent data from the health ecosystem, wellness applications, social networking to obtain the RWE while still obtaining the data required for regulatory approval. Also, pharmaceutical companies may be able to transform the product identity from simply selling pills in bottles to providing an all-inclusive therapeutic solution thanks to the new sorts of data. In other words, the new goal of drug developers should be to develop an integrated therapeutic strategy that takes into consideration real-world usage of a therapeutic solution in the context of digital devices, behavioral interventions, and other therapeutic options, which is equally possible to include a pill or exclude it.

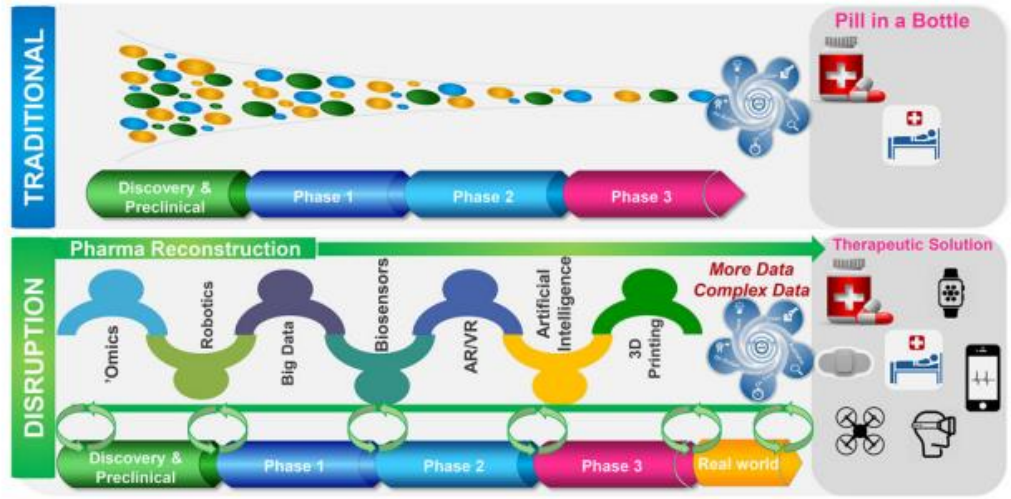

FIGURE 2.11: *Comparison the old and the new stages of* drug discovery/development model *(source:* https://pubmed.ncbi.nlm.nih.gov/29768712/*)*

However, a major portion of the early data was generated from shallow, big claims datasets. These databases pose several difficulties for clinical trials such as, claims conversion into visits and a clear patient history or recognizing the datasets completeness or incompleteness. In the same context how service locations and providers are categorized (provenance tracing) and how choosing the most beneficial measurements for expenditures and usage may pose a challenge for clinical trials. Particularly for some diseases like cancer, deeper EHR-derived databases from specialist businesses or from specific locations, payers came up. However, most of these databases are not deep and broad enough, and they are frequently just loosely linked. As it makes sense healthcare analytics vendors are now expanding their offerings to meet the demand for integrated data from numerous sources (e.g., labs,

consumer behavior, etc.) that capture the breadth of patient health. Offerings have started to grow around specific therapeutic areas.

FIGURE 2.12 : Life science companies are collecting now a complex of data (source: *https://www.healthcatalyst.com/insights/real-world-data-chief-driver-drug-development)*

2.4.1 The Future: Real-World Data to Real-World

As we know from healthcare transformation companies, only 8 percent of the needed data resides in the HER, Stupka (2019). This means that the industry has to gain access in the rest 92% of data remaining outside of the EHR to complete understand the patient condition. Having access to significant claims data and a portion of EHRs represents just the initial step in the broader outcomes' measurements for both population and personalized health. Oncology and cardiology specialty EHRs, for example, fill in certain marketable gaps but leave out the information that influences public health, such as costs, patient satisfaction, or lab findings. Extended RWD/RWE calls for broader sources to completely comprehend patients' life science firms can have access to this information by collaborating with an established healthcare transformation firm.

For a healthcare company to produce extended RWD its crucial to describe her the following capabilities, by Stupka, 2019:

1. Ability to evaluate outcomes in the actual world.
2. Ability to direct and secured trusted relationships with patients across the continuum of health, versus one area.
3. Ability to make the transition from practical insight to practical action (for instance, through patient- and population-level treatments).
4. Ability to address issues and provide practical answers, one needs access to the spectrum of expertise spanning data, and clinical trials.

Overall, it's wise to say that RWD and RWE is only just the begging, life science corporations must possess the capability to translate those kinds of data into provider-level actions that benefit both patients and the corporations. This involves implementing initiatives such as conducting important clinical studies, patient education programs, adherence, or safety programs, and more. By integrating businesses with various healthcare systems, patients, and the data that decides whether hypotheses are operational and lead to improvement, the correct healthcare transformation organization enables life sciences to achieve real-world action. For instance, a company can forecast drug response, identify patterns, and generate insights with shallow data, but it cannot fully comprehend the impact of the drug until it is used in a hospital environment.

By working together, both the life science industry and the right healthcare transformation companies can drive change, monitor, and measure drug performance. This concludes the real-world action circle, with five key points that are mentioned detailed in the following figure.

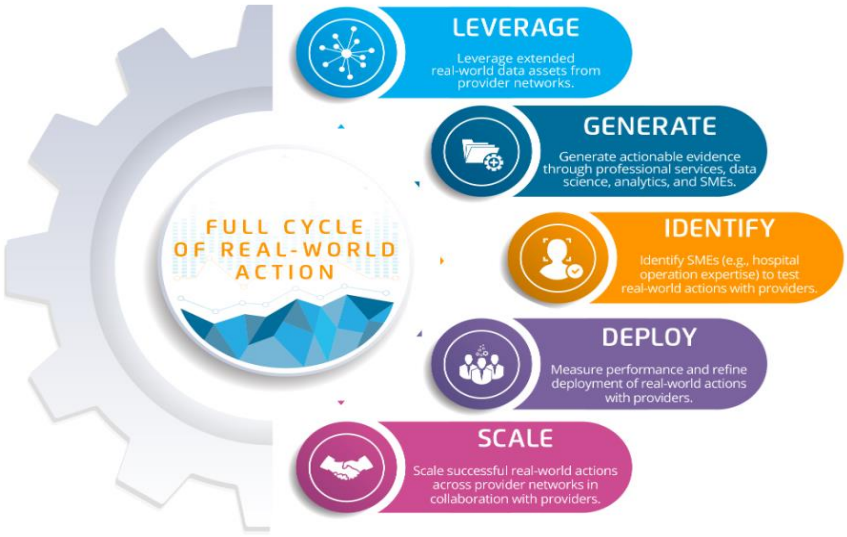

FIGURE 2.13 : The five key goals that will make life science companies thrive (source: *https://www.healthcatalyst.com/insights/real-world-data-chief-driver-drug-development*)

As a result of this alliance life science companies can improve in two major fields. First in core capabilities because as we said a leading healthcare transformation company is always looking ways to improve outcomes, by providing essential capabilities to the drug procedure, from start to finish, the three fundamental pillars of best practice, analytics, and adoption not only contribute to outcomes improvement but also wield a significant influence on the effective framework in drug development. And finally, we can have improvement in the Strategic consulting and professional services offerings skills around statistics, analytics and data manipulation that have special data to quickly find solutions, leverage technologies for data ingestion, visualization, and

complement it with deep operational expertise to contextualize the human factors and processes that drive success, Stupka (2019). By ensuring a trusted network of providers, life science companies can merge data and professional skills and refine a solution until it can function and scale across providers.

The next step for the life science sector is to scale those insights into actions, much like how payers and regulators have recognized the usefulness of extended RWD for critical decisions involving regulatory clearances and payments. The life science industry is accustomed to using data to provide specific insights. Life science companies can adopt meaningful outcomes-driven strategies for the development, regulation, reimbursement, and monitoring of new therapies through partnerships with organizations dedicated to healthcare transformation. This collaborative approach involves the utilization of extended-RWD, insights from real-world scenarios, trusted provider networks, population health management, and the establishment of clear definitions alongside real-time measurements of real-world outcomes.

3 CHAPTER

Literature Review in Pharmaceutical Research and Development

3.1 Exploring different ways Real-World Data can improve our health-system.

In the next chapter we try to provide insights through the entire spectrum of biological research. We already talked about how the real-world data encompasses a broad canvas such as, spanning DNA sequences and protein structures to clinical records and epidemiological datasets. These data repositories are not static archives but dynamic narratives of life itself, unfolding the intricate interplay of genetics, diseases, and evolutionary processes. As we venture into the literary landscape exploring real-world data in bioinformatics, we embark on a journey to decipher the hidden techniques in the domain of pharmaceutical research and development, with the utilization of Artificial Intelligence, Deep and Machine Learning (Section 3.1).

In our exploration, we'll traverse through a series of compelling case studies (Section 3.1) that demonstrate the profound impact of RWD on diverse aspects of bioinformatics. Additionally, we'll delve into the intriguing synergy between artificial intelligence and arthroplasty (Section 3.2), highlighting how real-world data-driven AI technologies are revolutionizing orthopedic surgery by optimizing patient outcomes and prosthetic implant designs. Turning our attention to the broader spectrum of AI applications, we'll explore how RWD and ML are addressing acoustic problems in the everyday environment (Section 3.3). With the impactful aid of hearing aids and wearables, AI algorithms are reshaping the way we collect and interact with audio data, enhancing the quality of soundscapes and communication systems.

Moreover, in (Section 3.4) we'll focus on how a large real-world registry TREVO 2000 aids in the treatment of acute ischemic stroke, where real-world data-driven machine learning models are proving instrumental in early diagnosis and treatment planning, potentially saving countless lives. Finally, we examine a study which aimed to predict COVID-19 severity (Section 3.5). The research focus on the occurrence of acute respiratory distress syndrome (ARDS) within four months of the initial diagnosis using different machine learning techniques. In each of these domains, RWD emerges as a critical catalyst, propelling bioinformatics into an era of unprecedented discovery and innovation. With these powerful tools at our disposal, we find ourselves at the precipice of a new era in medicine a future where data driven insights empower us to improve healthcare outcomes, ultimately benefiting patients worldwide.

Even though, the growth of this news techniques represents major technological breakthroughs, the results that we gather could be misleading if we are not able to separate the confounding factors, use the correct algorithms, examine the correct data, and fully comprehend the clinical questions behind the endpoints. It is vital to train ML algorithms accurately to have reliable performance in practice using multiple data scenarios. Furthermore, not all research questions can be answered by ML and AI especially when there is unpredictability or bad quality data, under-representation of certain patient groups, or even flawed trial design. Under-representation is a problem that should be taken seriously into account because it may result in systematic bias.

3.1.1 Harnessing Deep Learning for insights into protein structure

The functional mechanism of a protein is dictated by its three-dimensional configuration, which is intricately encoded within its linear sequence of amino acids. Utilizing insights into protein structures aids in comprehending their biological functions and contributes to the exploration of novel therapeutic approaches for either inhibiting or activating proteins to address specific diseases. Anomalies in protein folding play a pivotal role in various disorders, encompassing type II diabetes, Alzheimer's, Parkinson's, and amyotrophic lateral sclerosis. Also, the disparity between the one-dimensional amino acid sequence of proteins and their intricate three-dimensional structures, has substantial merit in devising accurate methods for predicting these structures. Such advancements not only facilitate drug discovery but also enhance our comprehension of diseases linked to protein misfolding.

An AI network called AlphaFold, created by DeepMind (Google), uses a protein's amino acid sequence to identify the protein's three-dimensional form Senior et al. (2020). It applied a DL method to estimate the protein's structure based on its sequence. The central aspect of AlphaFold is a convolutional neural network that was trained on the Protein Data Bank structures to forecast the distances among every pair of residues in a protein sequence, with a probabilistic distance map of a 64 × 64 region. To create a protein structure that complies with the distance predictions, these sections are then tiled together to produce distance predictions for the full protein.

The first results come in 2020, when AlphaFold released the structure predictions of five understudied SARS-CoV-2 targets including SARS-CoV-2 membrane protein, Nsp2, Nsp4, Nsp6, which will hopefully enlighten us in the domain of under-studied biological systems.

Moving next to Molecule Transformer-Drug Target Interaction (MT-DTI), which mention also by Beck et al., 2020 and in a few words can described as a deep learning-based drug-target interaction prediction model that predicts binding affinities based on chemical and amino acid sequences of a target protein, without their structural information. It can be used to find powerful FDA-approved medications that may inhibit the functions of SARS-CoV-2's proteins by Beck et al.(2020), computationally discovered several known antiviral substances, such as atazanavir, remdesivir, efavirenz, ritonavir, and dolutegravir, which are estimated to demonstrate an inhibitory potency against SARS-CoV-2 3C–like proteinase and can be possibly repurposed as candidate treatments of SARS-CoV-2 infection in clinical trials.

3.1.2 The role of machine learning for developing predictive biomarkers

Various case studies have lately been presented that the biomarkers derived by the ML predictive models were applied to stratify patients in clinical development. Predictive models were created to see if they might be used to foretell patient response to the treatments erlotinib, which is used to treat non-small cell lung or pancreatic cancer, and sorafenib, which is used in kidney, liver, and thyroid cancer. The models

use the IC50 values as a dependent variable and gene expression data from untreated cells as independent. The training dataset was the whole-cell line panel, and the testing dataset was the gene expression data gathered from tumor samples of patients who had taken similar medication.

The training data for the drug sensitivity predictive models did not include any information from the testing dataset. Furthermore, one must consider that to evaluate how well the drug sensitivity predictive models developed using cell line data, the BATTLE clinical trial data was used as an independent testing dataset. In that way we could determine the IC50s that define the model-predicted drug-sensitive and drug-resistant groups and after choosing the best models. Moreover, by B. Li et al., 2015 employed a predictive model to stratify patients in the erlotinib arm from the BATTLE trial. The median Progression-Free Survival (PFS) for the erlotinib-sensitive patient group was predicted by the model to be 3.84 months, while the PFS for the erlotinib-resistant patient group was predicted by the model to be 1.84 months. This indicates that the erlotinib-sensitive patients predicted by the model had a PFS advantage of more than twice as much as erlotinib-resistant patients.

In a comparable manner the model-predicted sorafenib-sensitive group had a median PFS benefit of 2.66 months over the sorafenib-resistant group with (p-value= 0.006 and hazard ratio=0.32). For the model-predicted sorafenib-sensitive and sorafenib-resistant groups, the median PFS was 4.53 and 1.87 months, in each case.

3.1.3 When to choose each method to optimize pharmaceutical research

According to research, by Hwang (2016) phase 3 studies with innovative treatments failed in clinical development in 54% of cases, with 57% of those failures being related to insufficient efficacy. Failure to correctly identify the target patient population with the appropriate dosing regimen, including the proper dose levels and combination partners, is a very significant reason for this. A possible solution approach could be a systematic model utilizing ML applied to (a) build a probabilistic model to forecast odds of success (b) identify subgroups of patients with a higher chance of therapeutic benefit. This will make it possible to match patients with the proper therapy in the best possible way, maximizing both patient benefit and resource benefit.

The training datasets are restricted to the same class of medications and may comprise all current early-phase and published data. Determining endpoints that can be used to calculate therapeutic effect most effectively is a significant difficulty in developing the probabilistic model. Early-phase clinical trials particularly in oncology use various primary efficacy endpoints in contrast with confirmatory pivotal trials due to a relatively shorter monitoring and need for faster decision-making. For instance, typical oncology objectives are overall response rate or complete response rate in phase I/II and progression-free survival (PFS) and/or overall survival both measured over an extended period benefit in pivotal phase III studies.

For instance, if we focus on Phase I/II trials in oncology we frequently use single arm settings to establish proof of concept and generate the treatment benefit hypothesis,

whereas pivotal trials, particularly randomized phase III trials with a control arm, are intended to show superior treatment benefit over currently prescribed therapies. This modification in the targeted endpoints from the very early phase to late phase makes the estimation of POS in the pivotal trial, using early-phase data, quite difficult at times. Training datasets using previous trials for medications with a similar procedure and/or evidence can encourage determining the relationship among the short-term and long-term endpoints, which eventually defines the success of drug development. Additionally, unsupervised learning can be used to cluster patients. Case in point, nonparametric Bayesian hierarchical models by executing the Dirichlet process enables patient grouping, without pre-specified number of clusters, with key predictive or prognostic variables, to symbolize various benefits. In conclusion this strategy is certain, that will increase efficiency in the clinical development of precision medicine over the next few years.

To sum it up, now that we take a first taste of several techniques and where they can be beneficial, we have to mention that knowing when Deep Learning, Artificial Intelligence, Machine Learning, and traditional inference are most effective in pharmaceutical research and development is also crucial. We try to suggest in the (Figure 3.1) that follows below based on the dataset's dimensionality.

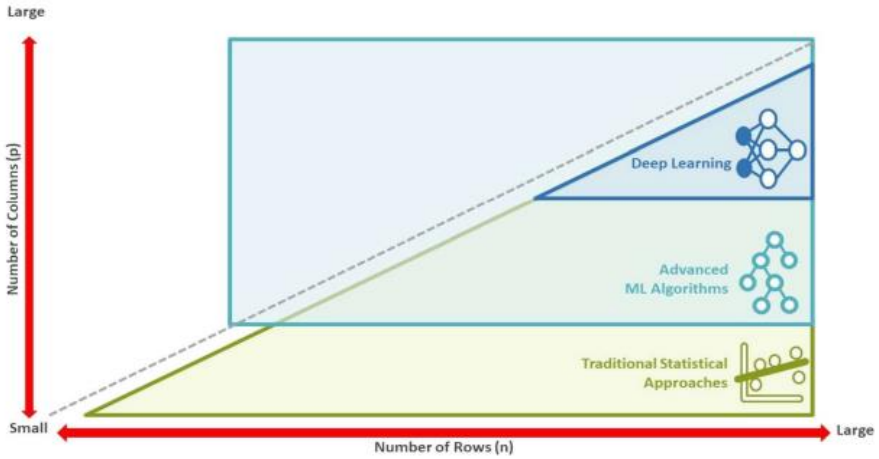

FIGURE 3.1: *Choosing the most suitable technique based on the size of the data (source: https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/34984579)*

In a similar way we try to distinguish the methods (Figure 3.2), but this time it is based on numerous factors of medication development. Even though many ML algorithms are capable of handling big data with the “Large p, Small n” issue, the increased number of variables/predictors continues to be a difficult task. To be specific, as the irrelevant variables increases, the weight of the noise becomes greater, resulting in the reduced predictive performance of most ML algorithms Kolluri et al. (2022).

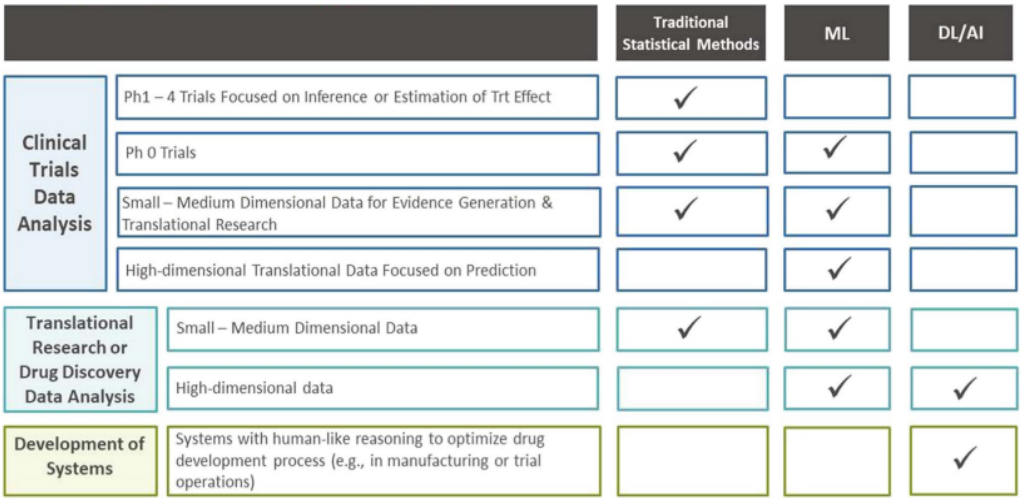

FIGURE 3.2: *Choosing the most suitable technique based on various aspects of drug development(source: https://www.ncbi.nlm.nih.gov/pmc/articles/pmid/34984579)*

3.2 How can we elevate arthroplasty using Real-World Data

In this unit we will examine how we can harness the information from big data and computing power, so we can upgrade the knowledge in the field of Arthroplasty. To be specific scientists have tried to establish a Machine Learning Arthroplasty Laboratory (MLAL) to examine the usage of AI to musculoskeletal medicine. Furthermore, in the next paragraphs, we focus on the two core objectives of the MLAL as they relate to the practice and progress of orthopedic surgery: (1) patient-specific, value-based care and (2) human movement.

To illustrate in the field of orthopedics the success of a procedure can be determined not only by the anatomical restoration on an X-ray or the improved motion of a joint, but also by the patient's subjective experience of the procedure. This has caused a paradigmatic change in orthopedic practice and prompted an organized effort to gather information on patient-reported outcomes. Over the past two decades, arthroplasty research has used numerous outcome metrics and registries. Machine learning aids to that cause, through the use of algorithms can be trained to assist humans with little to no human continuous effort.

3.2.1 The role of MLAL and how to utilize it

In 2018 the MLAL was established to make computer-based algorithms demonstrate the primary sustainable way for the future of orthopedic surgeons and take advantage of all available data to find the best possible outcomes for patients, Ramkumar et al., 2019. Orthopedic care and the MLAL operate on 2 crucial planes: system-based and practice based. At the system level, results and expenses are the two main determinants

for value-based care. However, what some patients consider to be highly valuable could not apply to other people. This is more obvious when comparing total hip arthroplasty (THA) patients who want to run a marathon with those who just want to have a walk in the park. Since "value" in medicine varies from patient to patient, machine learning gives the opportunity to take these patient-level factors into account and provide value-based care that is tailored to the patient's needs. In conclusion, the major objective of MLAL is to identify and implement machine-learning solutions that enhance the normal practice of orthopedic medicine by prioritizing the patient, supporting the doctor, and benefiting key stakeholders (e.g., hospitals, institutions, and payers).

For instance, to test the viability of predicting LOS (length of stay) and inpatient payments, a Naive Bayesian classifier algorithm was used to a statewide administrative database of over 260,000 primary total hip and knee arthroplasty (THA) and (TKA) patients. Representing a rudimentary form of machine learning, the Naïve Bayesian classifier is capable to analyze a large dataset, examine patterns based on the outcome variable of interest (ie, cost and LOS), and estimate what predetermined "bucket" to identify a new patient outside the studied dataset would likely resemble (i.e., <$12,000, $12,000- 24,000, >$24,000 or <3 nights, 3-5 nights, or >5 nights) based on patterns from the earlier introduced dataset.

This had as a result for primary TKA patients, reimbursement tiers warrant increases of 3%, 10%, and 15% for moderate, major, and extreme comorbidities. While for primary THA patients, reimbursement tiers warrant increases of 3%, 12%, and 32% for moderate, major, and extreme comorbidities. Nonetheless, the limitation of this model centered on the use of just one database population, creating homogeneity bias, and the inability of a Naïve Bayesian model to output a specific value rather than an LOS or cost "bucket."

Moving on from the model we just discussed, we shift our interest in simple Naïve Bayesian approaches, which fall under the category of "supervised learning." With this method, larger human participation is needed than "unsupervised learning," in order artificial neural network (ANN) function well. Such ANNs provide the chance to increase algorithm accuracy and include external data in various forms. As an illustration, ANNs are a subtype of machine learning that can analyze a database full of radiographs labeled with implant designs, try to find a correlation between the radiograph patterns and associated label, and then, if the implant has already been "learned," recognize the implant from a new radiograph.

In simple speaking, these ANNs represent a microcosm of experience-based learning and are even schematically organized after the individuals with several processing "nodes" densely linked in an axonal fashion. Like a neuron, one node may receive information from several other "dendritic" nodes while only sending information in one direction. The weight of the entering variable must be high enough to stimulate more nodes and create a correlational relationship before a node can "fire" or send data. When an ANN is being trained, all weights and thresholds are originally set to random values. Training data are provided to the input layer, and it passes through the succeeding layers, getting multiplied and added together in complex ways, until it eventually arrives, radically transformed, at the output layer. Weights and thresholds are continuously changed throughout training until training data with identical labels consistently produce results that are similar.

By applying a cohort study of 175,042 primary TKA patients with 15 preoperative input parameters, the ANN predicted LOS, charges, and discharge disposition with a discriminatory power of 74.8%, 82.8%, and 76.1%, respectively, based on the area under the curve. For moderate, significant, and severe comorbidities, the model showed increased reimbursements of 2.0%, 21.8%, and 82.6%, respectively. Similarly, an ANN developed for primary THA demonstrated area under the curves of 82.0%, 83.4%, and 79.4% for LOS, charges, and disposition, respectively, with charges increasing by 2.5%, 8.9%, and 17.3% for moderate, major, and severe comorbidities, respectively. These ANNs are capable of further learning and changes when new data is acquired in the future, which will enhance their predicting powers.

3.2.2 Remote patient monitoring through mobiles

It is well known that any huge dataset can be processed using ML. Apart from the extensive outcome datasets stored in registries, our mobiles record and store vast quantities of "small data," which also requires examination for clinically relevant insights. Smartphones and other mobile gadgets, including wearables, are now popular for this particular use. Except for the instant connectivity offered by cellular networks and the Internet, these devices also serve as sensors capable of maintaining enormous amounts of personal health data “mHealth”.

Of course, all EMR that rely on remote servers, maintaining Health Insurance Portability and Accountability Act compliance with standard regulation supervision must be guaranteed before clinical adoption. Once the “small data” of a given individual’s minute-by-minute step count or heart rate are correct measured they transform into big data. Furthermore, the user interface needs to be simple and employ real-time feedback to encourage bilateral involvement between the patient and doctor. To address this, the MLAL has teamed up with Focus Motion, a developer of proprietary data-driven orthopedic solutions, in Santa Monica, California, to develop a remote patient monitoring system that uses open architecture to harness the power of mHealth data, AI algorithms to "learn" human movements, and real-time feedback.

For the system to "learn" a movement, an activity must be labeled (for example, "straight leg raise"), carried out while using the wearable, and then all positional signals from the sensors must be processed and "taught" that a specific movement belongs to this action. The algorithm starts to recognize and give feedback on an activity after enough variations and repeats of that activity. This remote patient monitoring system is broadly scalable, freely available, and interoperable with any consumer mobile device in contrast to other platforms.

By providing a knee sleeve that pairs to the patient’s smartphone we managed to collect data from TKA patients about home exercise plan compliance, daily step count (i.e., activity level), daily knee range of motion, weekly patient-reported outcome scores, and opioid use. To be more specific we illustrate an example of a patient in the next (Figure 3.3).

FIGURE 3.3: *Summative dashboard data of a patient recovering from TKA who found the remote patient monitoring platform "highly motivating (source: https://pubmed.ncbi.nlm.nih.gov/31280916/)*

No patient had uninterrupted data collection at the time the trial was completed 90 days after surgery, proving an outstanding connectivity. All 22 of the 25 patients who were available for follow-up interviews also considered the system to be inspiring and interesting. Daily home exercise program compliance with automatic notification was 62% within the first 90 days postoperatively. This platform among numerous mobile apps being used across the globe to perioperatively assess and communicate with TKA patients. Opioid use typically stopped by postoperative fifth day, and mean mobility returned to baseline at approximately 6 weeks. This research tackles a critical obstacle in the captivity of results and therapy compliance data that have been previously limited by patient access, discontinuous data, high overhead cost, and capable technology.

By constantly studying RWD and importing them into clinical workflow, we might accomplish "high performance medicine." soon. This calls for continuing to be at the cutting edge of information regarding the benefits and drawbacks of these developing technologies for orthopedics, and large volume subspecialties like arthroplasty as we examined. Finally, allowing automation should not automatically trigger suspicions because some time-consuming processes, such as "clicks" in the electronic medical record, may in fact call for automation.

3.3 Improve our acoustics everyday with the power of Real-World Data

By utilizing data collected from hearing aid we might evolve the customization of hearing aid processing for each user according to hearing circumstances they face everyday life. Prior studies describing hearing aid users' auditory environments have examined mean sound pressure levels and proportions of environments relying on classifications. The diversity of auditory environments experienced by hearing aid users will be quantified in the following paragraphs (3.3.1 & 3.3.2) by introducing entropy as an extension of these methodologies. Participants from four different groups wore

research hearing aids and answered evaluations on a smartphone for one week. The smartphone was set up to offer an ecological momentary evaluation every 40 minutes and to sample the hearing aids' processing state every 10 minutes.

Next, we'll delve into a second, even more captivating study (3.3.3 & 3.3.4) that explores the correlation between the everyday acoustic environment (the real-world data once more has been gathered from hearing aids and wearables) and human heart rate. The focus of the following study is to examine the short-term association among multidimensional acoustic characteristics of everyday ambient sound and continuous mean heart rate.

3.3.1 The sequential steps for investigating our acoustic environments

In today's era we have grown a huge interest in understanding the everyday soundscapes or auditory environments that hearing aid users encounter. To begin with we could combine the auditory environments users as well as their unique hearing needs in certain environments with the information from hearing aid selection, signal processing, counseling, and aural rehabilitation.

For instance, hearing aid users who lead more active lifestyles may benefit more from cutting-edge hearing aid technologies than listeners who lead fewer active lifestyles since they are more likely to experience a variety of auditory settings. Technological improvements have enabled new techniques for acquiring real-world data for example the use of hearing aids that can collect data about the environment, including the sound pressure level (SPL), hearing aid environment classification, and the hearing aid processing state. However, how to utilize these data the most to characterize the environments users experience and draw conclusions is still a challenge, Jorgensen et al., 2023. The common approach has been to outline averages and proportions, typically average sound pressure levels and proportions of environment styles. This method though offers a restricted glimpse into the auditory habits of hearing aid users.

The ability of contemporary hearing aids to adapt to environments, and even, with the help of machine learning, to modify their processing, is one of their primary features hearing aids can offer. Thus, it is of interest to find data that describes how different a hearing aid user's auditory environments are, how they shift over time, and what lifestyle variables might forecast these metrics. In that study, younger individuals with normal hearing and older participants with hearing loss living in urban or rural settings, were compared in terms of auditory environments and the activation of hearing aid features.

Furthermore, entropy values for sound pressure levels, environment classifications, and ecological momentary evaluation reactions were calculated for each member to measure the diversity of auditory environments encountered over the course of the week. Entropy can be described as a measure of the number of ways a system can be arranged, often taken to be a measure of "disorder" (the higher the entropy, the higher the disorder). Additionally, to validate the use of entropy as a measure of auditory

environment diversity, entropy measured from hearing aid data will be compared to self-report data from ecological momentary assessments (EMA) surveys taken on a mobile device during each day. Only EMA responses where the subject proved that they were actively listening were included in the evaluation. With the participants' permission, GPS coordinates were also added to EMA surveys (Figure 3.4).

Question	Response options
Q1. What did your active listening involve?	1. Conversation, live
	2. Conversation via electronic device
	3. Speech/music listening live
	4. Speech/music listening, media
	5. Environmental sound listening
Q2. (if Q1 = 1 or 2) Were you talking with more than one person?	1. Yes
	2. No
Q3. (if Q1 = 3 or 4) What kind of sounds were you listening to?	1. Speech
	2. Music
Q4. Were you in wind?	1. Yes
	2. No
Q5. Was there music in the background?	1. Yes
	2. No
Q6. Were there people around you talking in the background?	1. Yes
	2. No
Q7. How loud were the background environmental sounds?	1. Very loud
	2. Loud
	3. Medium
	4. Soft
	5. Very soft
Q8. (if Q1 = 1 or 2, or Q3 = 1) The speech of interest was _____ when compared to all other sounds.	1. Much louder
	2. Somewhat louder
	3. Equally loud
	4. Somewhat softer
	5. Much softer

FIGURE 3.4: *EMA questions and potential answers (source: https://www.frontiersin.org/articles/10.3389/fdgth.2023.1141917/full)*

The final point to equally examine has to do with the data set used for this research. To be specific, 46 participants were enrolled in that study (data collection took place from 2017–2019), divided as we said into four groups: younger listeners with normal hearing from an urban area (YNH-U), younger listeners with normal hearing from a rural area (YNH-R), and older listeners with hearing loss from both urban and rural areas (OHL-U).The agricultural area was eastern Iowa, focused around Iowa City, and the urban area was the greater San Francisco Bay Area, centered around Berkeley, California. Older was arranged as the age of 35 and beyond while participants with normal hearing had to show audiometric thresholds less than 25 dB HL at all audiometric frequencies. Users with hearing loss had to have acquired, mild-to-moderate sensorineural hearing loss and be experienced hearing aid users. Additionally, OHL groups although were retired, they participated in various volunteer, social, religious, or community groups or held part-time employment also, the YNH consisted of students and working professionals, with a majority expressing involvement in a diverse array of social and community activities.

3.3.2 Conclusive findings and research techniques from the study

This study focusses on the use of entropy to quantify auditory environment diversity through SPL and environment classification data from hearing aids. It is well known that entropy can be calculated in a straightforward manner and can be validated from hearing aid data, as a measure of auditory environment diversity by comparing SPL and environment class between multiple listeners and comparing these differences to EMA results. To illustrate, SPL and environment class entropy was significantly higher for the YNH than the OHL participants, with the largest differences observed between the YNH-U and OHL-R groups. Similarly, the YNH participants had significantly higher EMA entropy than the OHL participants. Finally, this research aimed to compare entropy measured from hearing aid data to entropy from EMA. Significant, moderate correlations were observed among SPL and environment class entropy and between SPL-EMA entropy, providing further evidence for the validity of entropy as a measure of auditory environment diversity.

When considered collectively, the results of this study indicate that younger listeners experience a greater variety of auditory environments than older listeners, that this variety can be captured using hearing aid data and measured using entropy, and that entropy calculated using objective hearing aid data broadly corresponds with entropy measured from self-report EMA data. We should take also into account that the research indicates that age is a more robust predictor of auditory environment diversity than geographic location, and with a relatively small sample, the clearest differences emerged when groups were combined along age. Overall, entropy could be a vital metric for a hearing aid to examine the auditory environment diversity of different participants with different needs and make processing changes based on individual users' auditory environment diversity.

Finally, as it comes to the methods used, we briefly mention that parts of the analysis were performed comparing all groups and other part for groups that were combined according to age and hearing ability. One-way Analysis of Variance (ANOVA) was used to examine group differences. Significant omnibus statistics were followed when appropriate by a priori pair QSP comparisons with Tukey p-value corrections for multiple comparisons. Model assumptions were analyzed by visually studying the data distribution and residuals, and no evidence of violating model assumptions was detected. Pearson-product moment correlation was used to evaluate the correlations between the different entropy measures.

3.3.3 Connecting different acoustic environments with heart rate

In this paper the sound environment is defined through four different criteria: sound pressure level (SPL), sound modulation level (SML), signal-to-noise ratio (SNR) and last soundscape class, demonstrating distinct characteristics of the slight sound immersion. SPLs are the most widely used measure of sound wave strength and describe sound intensity. Moreover, it is essential to emphasize SPL and the loudness that individuals perceive are highly correlated. On the other hand, SMLs are used to

describe temporal amplitude modulation, which is the degree of oscillation in the sound wave amplitude across brief time intervals that is present in speech and music. In other words, SML represents the sound wave's short-term dynamics. Moving next to the SNRs someone can say that, represent a spectral dimension of the sound by differentiating among the level of background sound relative to the level of the signal in decibels, Christensen, et al. (2021). In more straightforward terms, a more positive value indicates less noise relative to the signal. Finally, soundscapes are a qualitative dimension of the acoustic environment assumed to relate to how effortful it is to listen to speech-like sources in the presence of different levels of background noise.

To give a quick glimpse of the dataset used for this study, 1.115.332 acoustic environment data logs and 522.715 heart rate logs were collected, which represents almost 9.000 h of bilateral hearing aid use and 61.000 h of data from wearables, respectively. Only data logged during 06.00 and 24.00 were considered legitimate to prevent confounds from night-time logs that were possibly gathered when neither the hearing aids nor wearables were being used Christensen et al., 2021.

Data have been studied in two separate phases. The auditory environment was first described using all observations, independent of temporal overlap with HR logs, to make the most of the data that was available. Second, the data were pre-processed to guarantee complete overlap between auditory variables and HR logs for further statistical modeling. During pre-processing, time frames of 5 minutes before each HR log were chosen, and the arithmetic average of each acoustic variable was calculated within that window. As a result, the data records with totally overlapping data have each acoustic data variable's value calculated from the same time window as the running mean HR logs. To prevent potential confounds from low-incident HRs, the records under 5th and above 95th percentile of the group mean were excluded. These eliminations ensured that the residuals in statistical modeling were normal, while also did not influence the regression coefficients' order or the statistical significance of the included statistical models. Also due to the unbalanced samples per individual and hierarchical multi-level nature of the data documents, associations across variables were estimated using linear mixed-effect (LME) models. The 'nlme—Linear and Nonlinear Mixed-Effects Models' package (v. 3.1) was used for the application of mixed-effects models.

3.3.4 Linking sound to heart rate and exporting final conclusions

To begin with LME models were executed separate for associating mean heart rate with either the categorical soundscape or the acoustic data SPL, SML and SNR, Christensen, et al. (2021). The random effect’s structure has been modified for baseline offsets in heart rates caused by time of day, week nested within individuals, and unique variations in baseline sensitivity to each fixed effect (i.e., random intercepts and slopes). In separate models, modification for the movement was used by including an additional nested random effect equal to the estimated movement quantified into 10 equal-sized bins (deciles).

The movement we already mention, was estimated within the database to examine the possible confounding effect on the relationship between HRs and acoustic

environment data and later extracted for analysis. Specifically, the distance in meters between two consecutive latitude and longitude coordinates was computed using the haversine method. Using this method, for each pair of subsequent latitude (φ1, φ2) and longitude (λ1, λ2) coordinates the distance between them in meters, d, is calculated as:

$$d = 2\ r \arcsin\left(\sqrt{sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

with r being equal to the radius of the earth. Movement in m s−1 was then computed by dividing d with the 1 min time-window among each observation and averaging across the 5 min time-window preceding each heart rate ratio. To ensure that only movement through physical activity were included, data that exceeded cycling speed (10 m s−1) were not take into consideration.

To conclude the study findings, indicate that variations in heart rate were highly correlated with parameters of the auditory environment. Higher SPL and SML values were linked to higher HRs, but more favorable (higher) SNR values were linked to lower heart rates. Moreover, the documented associations among acoustic data and heart rate were higher in simple listening soundscapes (such as 'Quiet' and 'Speech') compared with soundscapes classified as containing noise, while marginal means revealed that heart rate moderation by SMLs and SNRs were distinct, depending on the decibels. Also, a positive but smaller association between HR and SML appeared. Sound that is highly modulated is commonly recognized by rapid oscillations in SPLs, serving as a characteristic sign of either speech or music. One could consider that the positive association between SMLs and HRs are a result of conversational task demand. That is, in highly modulated noise environments, speech and listening needs are more demanding, which leads to increased sympathetic autonomic nervous system activity.

This study is the first to use longitudinal data to investigate the relationship between actual human heart rates and various aspects of the ambient auditory environment. According to the study, heart rates and ambient sound intensity are positively correlated. Additionally, the study results indicate that lower HRs are linked to real-world ambient signal-to-noise ratios, which suggests that the human cardiovascular system is less burdened by sound conditions that reduce auditory perceptual load and listening effort. This finding is empowered by a documented effect of soundscape on the strength of the association between acoustic characteristics and autonomic nervous system reactions. In other words, auditory properties have the strongest correlation with changes in heart rate under favorable listening settings.

In closing, but not any less crucial, our results indicate a mixed influence of everyday sounds on cardiovascular stress, and that the relationship is more complicated than is seen from an evaluation of sound intensity alone, Christensen, et al. (2021). Our results also demonstrate that data logging with commercially available devices can be used to study how ecological everyday acoustic environments affect human physiological reactions and emphasize the significance of including exposure to ambient sound in models predicting human physiology.

3.4 TREVO 2000: Real-World Data registry in cardiology

Recent RCTs demonstrate the usefulness of thrombectomy for stroke patients with major vascular blockage. Real-world data contribute to evaluating reproducibility of results outside of clinical trials. A multicenter, worldwide, prospective research called the Trevo Retriever Registry was created to evaluate patient outcomes over a huge cohort of patients Binning et al. (2018). It is no secret that stroke accounts for 9% of the total deaths around the world and is the second-leading killer after ischemic heart disease. Strokes can have an ischemic origin and be caused by embolic or thrombotic etiologies in up to 87% of incidents. In patients who report within 4.5 hours of the onset of stroke symptoms, intravenous thrombolysis with tissue plasminogen activator has historically been the first choice of treatment Binning et al. (2018). However, recent multicenter, RCTs have shown that selected patients with large vessel arterial occlusions are observed to have higher recanalization rates and better results when intravenous thrombolysis with tissue plasminogen activator is applied in conjunction with mechanical thrombectomy.

3.4.1 Data insights from the mechanical thrombectomy research

The Trevo stent retriever (Stryker Neurovascular) is a third generation mechanical thrombectomy machine used to incorporate and discard arterial thrombus in patients. The Trevo Retriever Registry is a prospective, real-world registry that gathered data from platforms that performed thrombectomy based on their regional protocols. The largest data collection on patients undergoing mechanical thrombectomy utilizing a stent retriever as the first-line device is in this cohort Binning et al. 2018). In other words, the Trevo Registry is a prospective database of individuals with large vessel occlusion treated with the Trevo as the first device. Revascularization based on the modified Thrombolysis in Cerebral Infarction score is the major end point, while other end points include the modified Rankin Scale at 90 days, mortality at 90 days, neurological worsening at 24 hours, and adverse events related to the device or surgery.

A total of 2008 patients were enrolled during the enrollment period from 76 centers internationally. From the intention-to-treat population, the 1365 (68%) were enrolled in the USA whereas 643 (32%) were enrolled outside USA. In the following (Figure 3.5), more patient characteristics are presented.

Characteristic	Intention to Treat (n=2008)
Age, y (mean±SD)	68±14
Sex: female	51.8% (1041/2008)
Atrial fibrillation	36.1% (722/2000)
Diabetes mellitus	23.8% (477/2002)
Coronary artery disease	22.2% (443/1999)
Congestive heart failure	14.2% (285/2002)
Baseline glucose >150 mg/dL	24% (446/1859)
Prestroke mRS	
0	70.9% (1372/1972)
1	14.7% (290/1972)
2	7.6% (151/1972)
3	4.1% (80/1972)
4	2.1% (42/1972)
5	0.5% (10/1972)
Baseline NIHSS (mean±SD)	15.5±6.8 (1991)
IV t-PA delivered	52.3% (1041/1990)
Pretreatment ASPECTS (core lab adjudicated)*	
0 to 5	13.4% (176/1309)
6 to 10	86.6% (1133/1309)

FIGURE 3.5: *Patients Characteristics (source: https://pubmed.ncbi.nlm.nih.gov/30561262/)*

For the statistical analyses, patient baseline characteristics and procedural data were examined at first glance using meters such as frequency, mean, SD, and median. Then, t test or Wilcoxon sum test was used to compare groups, while Fisher's exact test was used to compare dichotomous variables. Also, Clopper–Pearson CIs were constructed for inferences of key results. Additionally, both, uni- and multivariate logistic regression was executed on the intention-to-treat cohort to determine predictors of good results. Lastly when used chi-square p-value was set at 0.05. SAS software was used to perform these analyses and in patients with missing 90- day mRS, last observation carried forward was utilized.

3.4.2 TREVO 2000: Study findings and concerns about the future

Median admission National Institutes of Health Stroke Scale was 16 (interquartile range, 11–20). To talk with numbers, occlusion sites included the internal carotid artery 17.8%, middle cerebral artery 73.5%, posterior circulation 7.1%, and distal vascular locations 1.6%. In 92.8% of procedures, a modified Thrombolysis in Cerebral Infarction 2b or 3 was achieved, and at three months 55.3% of patients attained a modified Rankin Scale 2. At the same time, three months, patients who underwent thrombectomy and met the updated 2015 American Heart Association (AHA) criteria, had a modified Rankin Scale 0 to 2 of 59.7%, as opposed to 51.4% of patients who received treatment in violation of the AHA guidelines. Also, the risk of symptomatic cerebral bleeding was 1.7%.

On the other hand, we should mention that the analysis is subject to noteworthy limitations that warrant some attention and consideration. In total, we will enumerate and discuss five of these limitations. To begin with, this single-arm registry, which does not include a control arm, represents a sizable cohort of stent-retriever patients gathered

in the real world. Secondly, thrombectomy cases were not consecutively gathered at each hospital. It would not have been possible to obtain sequential patient data during this time because there were open clinical trials and competitive device registries. Furthermore, permission was received within 7 days of the treatment, and there is a chance that patients with less-than-desirable outcomes won't be enrolled because it would be awkward to contact families after the procedure. Moreover, local sites or investigators reported 3-month mRS, this assessment may have led to a greater number of patients who had good or exceptional clinical outcomes.

Last but certainly not least, there were inconsistencies in the imaging data. The fact that our results fall within the same range as the HERMES (Highly Effective Reperfusion Evaluated in Multiple Endovascular Stroke Trials) pooled analysis, with equivalent mortality rates (14% versus 15.3%) and autonomous results (55.7% versus 46%) at 90 days despite a slightly smaller stroke severity (median baseline National Institutes of Health Stroke Scale 15.5 versus 17), minimizes fears regarding selection bias, which are a natural byproduct of the structure of any registry. In order to decrease the bias for outcomes related to reperfusion and hemorrhagic complications, the Trevo registry also maintained a central core lab for reviewing angio- and radiographic imaging.

To sum it up the Trevo Retriever Registry represents real-world data through stent retriever. The registry indicates similar reperfusion rates and results in the community when compared with rigorous centrally adjudicated clinical trials. Results from clinical trials for stroke appear to be reproducible in the real world for a variety of occlusion sites, stroke severity, onset times, and patient comorbidities, according to outcome data. Keeping in mind also that, future subgroup analyses of this cohort will help identify potential study topics.

3.5 Leveraging big data to forecast COVID-19 severity cases

The COVID-19 pandemic has placed enormous pressure on worldwide medical systems, with varying degrees of symptom severity among infected individuals. machine learning is being used as a powerful tool for forecasting and controlling the severity of COVID-19 cases. ML techniques, such as XGBoost, Artificial Neural Networks, Random Forest, and deep learning, have been utilized to identify risk factors as well as create predictive models for COVID-19 severity, including hospitalization, ICU admission, ventilation needs, and mortality. These models frequently make use of a range of clinical data sources, including EHR and Medicare/Medicaid data, to give helpful information for healthcare professionals in managing and allocating resources.

This study is about to examine Lazzarini et al., 2022 introduces a robust Machine Learning model for predicting COVID-19 severity, specifically defined as the development of acute respiratory distress syndrome (ARDS) within four months of initial infection. This model was created utilizing a large dataset of almost 290,000 COVID-19 patients, encompassing over 800 diagnosis codes, in contrast to several prior studies with limited data. This extensive dataset guarantees a realistic representation of both severe and non-severe cases, leading to a more reliable predictive model. The model's performance was later validated on an independent test set, and its

predictions were compared to five clinicians. Additionally, advanced interpretability techniques were employed to identify key risk factors influencing ARDS development in COVID-19 patients.

3.5.1 Evaluating machine learning models for COVID-19

To begin with, this study, aimed to predict COVID-19 severity, specifically the occurrence of ARDS within four months of the initial diagnosis. It is important to mention that the cohort consisted of patients who were first diagnosed with COVID-19 in April 2020, ensuring they had no prior COVID-19 diagnoses or ARDS cases between October 2015 and January 2020. This cohort consisted of 289,351 patients as we said, with 10,793 progressing to severe COVID-19 and 278,558 not experiencing severe outcomes.

To construct a feature space for each patient, we delved into their claim history within a predefined "lookback" period, extending from the introduction of the ICD-10 code system in October 2015 to January 2020. Furthermore, ICD-10 code system refers to the tenth edition of the International Classification of Diseases, which is a medical coding system chiefly designed by the WHO to catalog health conditions by categories of similar diseases. This lookback period was essential for capturing comorbidities and medical history. Also, a gap of several months was included between the end of the lookback and the patient selection period in April 2020 to avoid incorporating features correlated with the COVID-19 infection that might have manifested earlier. For each patient, was generated 817 boolean comorbidity features, representing the presence or absence of specific diseases within the lookback window for example age and gender were included as input features.

The dataset was meticulously divided into training, validation, and test sets, with proportions of 80%, 10%, and 10% each, always making sure that the original positive-to-negative patient ratio of 1:26 was preserved in all subsets. To maximize the power of the predictive models, three widely used machine learning algorithms were selected: Logistic Regression, Random Forest, and LightGBM. Moreover, Bayesian optimization, facilitated by Hyperopt, was employed to fine-tune hyperparameters and maximize the Area Under the Curve of Precision-Recall (AUCPR), a comprehensive metric that balances precision and recall across various decision thresholds. Finally, the credibility of the model was strengthened by comparing its predictions with assessments made by five experienced clinicians, thus ensuring real-world relevance and practicality.

Additionally, it is crucial to recognize that the model performance was evaluated using mainly two metrics, the Receiver Operating Characteristic (ROC) curve and the Precision-Recall (PR) curve. ROC curve captures the balance between the true positive rate and false positive rate across various probability thresholds, whereas the PR curve emphasizes the compromise between the true positive rate (recall) and the positive predictive value (precision). To further quantify performance, examination of AUC the (Area Under the Curve) for the ROC curve and AUPRC (Area Under Precision Recall Curve) for the PR curve, was enlighten. Also, it is important to notice that, due to the imbalanced nature of the dataset, where negative samples were predominant, ROC

curves were recognized as potentially misleading due to their sensitivity in this imbalances Saito & Rehmsmeier, (2015). In contrast, precision remained a more robust performance metric independent of the dataset balance.

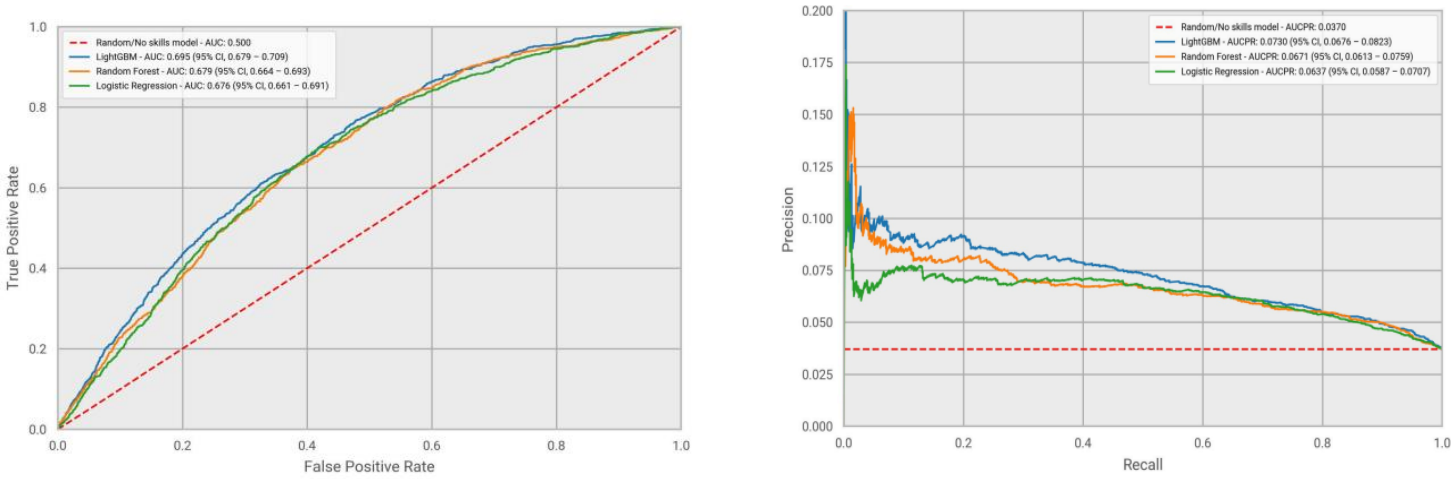

FIGURE 3.6: *Performance evaluation of the three machine learning models measured using a ROC curve (left) and (right) a PR curve (source: https://pubmed.ncbi.nlm.nih.gov/35901089/)*

Hyperopt was employed to optimize the hyperparameters for the models in this study. The chosen hyperparameters were those that yielded the highest performance on the training dataset, evaluated using a standard 5-fold cross-validation approach. Subsequently, the final models were constructed using the entire training dataset, and their performance was assessed on a separate test dataset, which was not utilized during the model training process.

In (Figure 3.6), the performance of each model was compared using both ROC curves and PR curves because of this analysis, the LightGBM model shown superior performance both times. The result above, aligns with existing literature, which has consistently shown that Gradient Boosted Decision Trees excel in healthcare situations. To speak with numbers, first in ROC curves, the LightGBM model reach an AUC of 0.695, leaving behind the values of 0.679 and 0.676 obtained by the Random Forest and Logistic Regression models, respectively. It is also essential to mention that all models shown values significantly above the baseline of AUC which is 0.5. Similar results were observed in PR curves, where the LightGBM model again emerged as the top-performing model, with an AUPRC of 0.0730. The AUCPR was again higher than the other models: 0.0671 for Random Forest and 0.0637 for Logistic Regression. The PR curve results imply that the selected model successfully extracted meaningful patterns, enabling accurate predictions of ARDS in patients from the test dataset.

It is also worth noting that the prevalence of ARDS in the study was lower than that reported in most published related works. This variation can be attributed to the inclusive method mentioned, which encompassed a broader spectrum of COVID-19 patients using IQVIA's claims data, as opposed to focusing solely on a specific subset. However, it is important to emphasize one more time the 1:26 positive-to-negative class ratio to mitigate the introduction of bias in the metrics. This ensured that the distribution of patient types in our test set remained consistent with the real-world. Taking into

consideration the performance of the LightGBM model, it would be a mistake to not concentrate exclusively on this model for the remainder of our analysis.

3.5.2 Conclusions and promises for the future

In the research, we embarked on an extensive exploration of machine learning models for predicting COVID-19 severity, with LightGBM, a powerful Gradient Boosted Decision Tree model, emerging as the frontrunner in terms of performance Saito & Rehmsmeier, (2015). To assess the efficacy of the models, both ROC (Receiver Operating Characteristic) and PR (Precision-Recall) curves were utilized. These curves provided a clear demonstration of improved performance compared to a classifier based solely on acute respiratory distress syndrome prevalence within the cohort. To make it simple, LightGBM can translate meaningful patterns, resulting in more accurate predictions. Additionally, we subjected the model's performance to rigorous evaluation by comparing it with the clinical expertise of five healthcare professionals. Impressively, the models exhibited similar precision and recall values to these experts, demonstrating its clinical utility. The model's interpretability was enhanced by SHAP (Shapley Additive Explanations), which allowed to clarify, Age (0.5) and Gender (0.2) (mean |SHAP value|) as the most significant constants. These results support previously known information about COVID-19 severity risk factors, further supporting the validity. Also, the feature selection process was unbiased, utilizing all available ICD (International Classification of Diseases) codes from patients' medical histories.

While the study showcases the potential of machine learning models in forecasting COVID-19 severity, it has some limitations. The use of claims data may cause sample bias, as it might not represent patients with limited or no access to the healthcare system. Equally noteworthy is the fact that, potential data reporting delays by extending the timeframe utilized for identifying severity. However, the prevalence of ARDS in the cohort was in the same page with published literature Matthay et al. 2020.

Future research projects will evaluate models trained on EMR, broaden the COVID-19 patient selection window to cover an even larger cohort, and then compare the existing model with available predictive tools. This way we could be in position to investigate vaccine efficacy and explore potential long-term side effects. Ultimately, the aim is to optimize vaccination strategies by individual’s segments based on age or other important variables Saito & Rehmsmeier, (2015). To sum it up, employing machine learning with claims data can aid in predicting which COVID-19 patients are at higher risk. This way can eventually enhance the allocation of hospital resources and prioritize patient care.

4. CHAPTER

MACHINE LEARNING

4.1 Types of Machine Learning

This next chapter explores the various categories of machine learning, emphasizing their significance and how they work. By leveraging complex algorithms, machine learning provides sophisticated tools for uncovering patterns, making predictions, and driving decision-making processes. These advanced methodologies extend the capabilities of classical statistics, enabling more nuanced and accurate insights across diverse applications. As we explore the different facets of machine learning, we will highlight the important of statistics and its profound impact on various occasions.

Diverse approaches to problem-solving are offered by mainly three primary types of machine learning: supervised, unsupervised and reinforcement learning Sapoval et al.,2022. Supervised learning is used for tasks like picture classification and regression since it relies on labeled datasets. Conversely, unsupervised learning finds uses in dimensionality reduction and clustering by uncovering hidden patterns and structures within unlabeled data. Reinforcement learning, which has been successful in fields like game play and autonomous systems, presents the idea of an agent interacting with an environment and adopting the best approaches, by trial and error. The agent receives feedback in the form of rewards or punishments based on the actions it takes and aims to collect the most rewards possible each time.

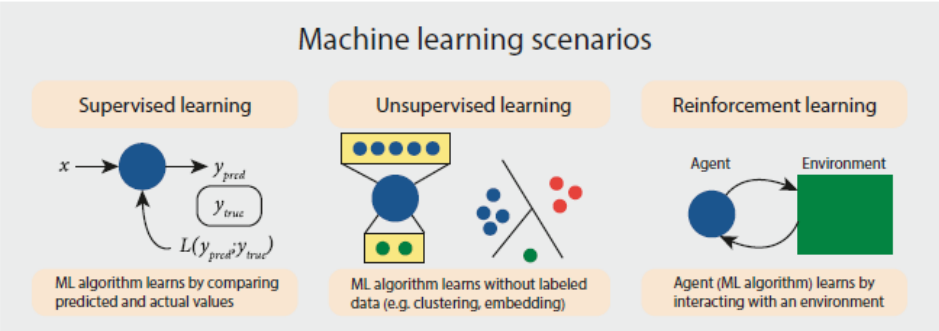

FIGURE 4.1: The main types of machine learning (source: https://www.researchgate.net/publication/359672214_Current_progress_and_open_challenges_for_applying_deep_learning_across_the_biosciences)

Apart from the types we just mentioned (Figure 4.1), new kinds are emerging such as: transfer, semi-supervised, and self-supervised learning that provide innovative solutions to problems in different fields. To be specific, semi-supervised learning leverages both labeled and unlabeled data to improve model accuracy. While self-supervised learning empowers models to generate their own labels, this way the model is trained to predict parts of the input data from other parts, creating a pseudo-supervised learning scenario, this mainly works efficiently for pre-training tasks. Transfer learning, on the other hand, can reuse the knowledge gained from one task to

improve the results on another, this may be crucial when dealing with limited labeled data and can save computational resources and time. The ongoing development of these methods emphasizes how dynamic machine learning is and how important a role it plays in forming the field of artificial intelligence.

Data preprocessing is a critical step in all types of machine learning that includes handling missing values, normalization, and outlier detection. Raw data is often inconsistent, incomplete, and unreliable, making it unsuitable for direct use in machine learning algorithms or statistical models. Through data preprocessing, we can enhance the quality of the data, making it more accurate and reliable, while reducing the likelihood of errors in the subsequent analysis Mehmed Kantardic (2020). Overall, data preprocessing is a crucial step in any data analysis task that can lead to better insights and results.

4.1.1 Preprocessing Steps

Preprocessing is a critical phase in data analysis that ensures the data is clean, well-formatted, and suitable for analysis. One of the first steps in preprocessing is data cleaning, which involves identifying and correcting errors, inconsistencies, and anomalies in the dataset. This step is crucial for maintaining the quality and integrity of the data, as even small errors can significantly skew the results. Data cleaning can include tasks such as correcting typos, standardizing formats, and resolving duplicates. Ensuring that the data is accurate and consistent lays a solid foundation for subsequent analysis.

Another important preprocessing step is data integration, where data from multiple sources is combined into a single, coherent dataset. This process often involves aligning different data formats, which is particularly important when dealing with large and diverse datasets, as it provides a comprehensive view of the information. Data reduction is also a key preprocessing step, aimed at simplifying the dataset without losing significant information. Data reduction helps in mitigating the dimensionality, improving the efficiency of the analysis, and enhancing the interpretability of the results. By focusing on the most informative features, data reduction enables more effective modeling and analysis, ensuring that the insights derived are meaningful.

Moreover, handling missing values is a significant step in data preprocessing because incomplete data can impact the accuracy and reliability of our analysis. First, an analyst with a professional, can modify samples with empty registries and enter an appropriate, or believed value based on the experience of each. The technique is relatively easy for small numbers but, the danger of adding noise into the data must be taken into consideration if there are a lot of missing values.

The next method is even simpler. It is based on a formal, often automatic replacement of missing values with constants, such as:

1. Replace all missing values with a single global constant (depends highly on the application).
2. Replace a missing value with its feature mean.
3. Replace a missing value with its feature mean for the given class (mainly for

classification problems where samples are classified in advance).

The last approach, typically chosen when dealing with a large amount of data, involves removing rows or columns that contain missing values from the dataset. However, this can result in data loss and may not be practical for datasets with many missing values. Although these solutions are tempting, the selection of the methodology for handling missing values depends on the specific characteristics of the data set and the goals of our research. The method we choose should be carefully considered, as it can have a significant impact on the results of our analysis.

4.1.2 Data Scaling and Normalization

Moving on to normalization there are certain methods, particularly those relying on distance computations in an n-dimensional space, often necessitate normalized data for optimal results. Normalization involves scaling measured values to a specific range, such as [-1, 1] or [0, 1]. Failure to normalize may result in distance measures overweighting features with, on average, larger values. Several effective techniques for normalizing data exist, with three notable methods outlined below:

(a) **Decimal Scaling:** Decimal scaling involves moving the decimal point while preserving most of the original digit value. The values are typically maintained within a range of -1 to 1. The scaling process is described by the equation: $v'(i) = \frac{v(i)}{10^k}$, where $v(i)$ is the value of the feature v for case i and $v(i)$ is the scaled value for the smallest k such that max($|v'(i)|$) < 1. First, the maximum $|v'(i)|$ is found in the data, and then, the decimal point is moved until the new, scaled maximum absolute value is less than 1. The divisor is then used to all other $v(i)$.For instance, if 455 is the biggest value and –834 the smallest, then the maximum absolute value of the feature becomes 0.834, and the divisor for all v(i) is 1000 (k = 3).

(b) **Min–Max Normalization:** Suppose now that the data for a feature v are in a range between 150 and 250. Then, the previous method of normalization will give all normalized data between 0.15 and 0.25 but it will accumulate the values on a small subinterval of the entire range. To obtain better distribution of values on a whole normalized interval, e.g., [0, 1], we can use the min–max formula

$$v'(i) = \frac{\big(v(i) - \min\big(v(i)\big)\big)}{\big(\max\big(v(i)\big) - \min\big(v(i)\big)\big)}$$

where the minimum and the maximum values for the feature v are computed on a set automatically or they are estimated by an expert in the domain. Similar transformation may be used for the normalized interval [−1,1]. The automatic computation of min and max values requires one additional search through the entire data set, but, computationally, the procedure is simple. On the other hand, expert estimations of min and max values may cause unintentional accumulation of normalized values.

(c) **Standard Deviation Normalization:** Normalization by standard deviation usually works well with distance measures but makes the data unrecognizable from the original form. For a feature v, the mean value mean (v) and the standard deviation sd (v) are

computed for the entire data set. Then, for a case i, the feature value is transformed using the equation

$$v*(i) = \frac{\big(v(i) - mean(v)\big)}{\text{sd}(v)}$$

For example, if the initial set of values of the attribute is $v = \{1, 2, 3\}$, then mean(v) = 2, sd(v) = 1, and the new set of normalized values is $v* = \{-1, 0, 1\}$. In machine learning, data scaling and normalization is an essential preprocessing step that has a big impact on the efficacy and accuracy of models and algorithms. It is crucial to give careful thought to which scaling strategy is best for the dataset and analysis being done.

4.1.3 Outlier Detection

In machine learning, outlier detection is the process of identifying data points or observations that significantly differ from others in a dataset Smiti (2020). Specifically, outliers can result from errors in data collection and entry, or they may represent genuinely unusual or rare events. Additionally, outliers can be removed from the dataset if they are deemed incorrect or transformed to mitigate their impact. Some simple ways to achieve this include taking the logarithm or square root of the value examined. In general, outliers can be of particular interest, and their impact on the analysis should be carefully examined each time.

Outliers are very different from noisy data, while noises are useless and must be removed like erroneous data values 999 instead of 99 for ‘‘age’’ attribute or incorrect data type (string type entered for a numeric attribute). On the contrary, outliers can provide both useless and interesting information. To make it clear we give a definition by the statistician, Hawkins 1980, «An observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism». Stated differently, outliers are data points that significantly vary from clearly defined norms within a data set or from predetermined notions of expected behavior. Sometimes they are helpful and should be kept, but other times we would like to eliminate them because they mislead our analysis. Outlier detection approaches can be categorized based on different criteria. For our thesis we briefly mention four principal methods: Statistical-based, Distance-based, Clustering based and Density-based.

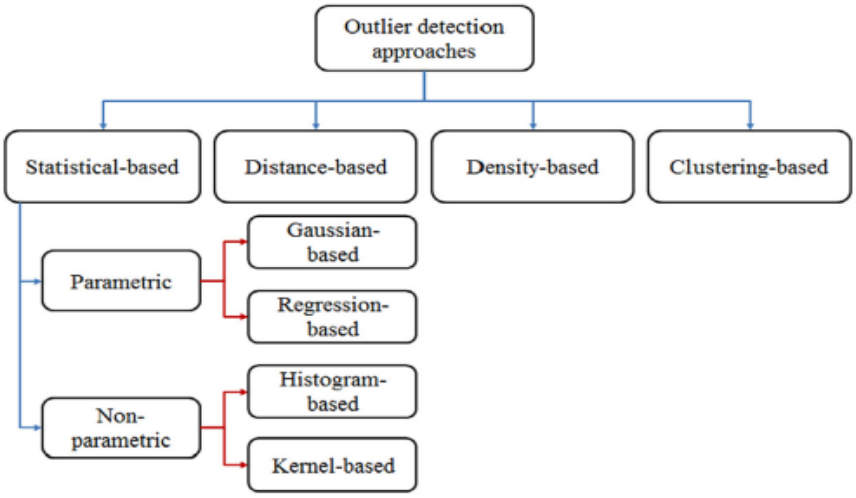

FIGURE 4.2: Outlier detection methods (source: https://www.sciencedirect.com/science/article/pii/S1574013720304068)

Outlier detection through statistical methods, has two main categories parametric and non-parametric and aim to identify outliers by assessing how significantly a data point deviates from a standard distribution. Parametric methods, such as Gaussian-based and regression-based techniques, rely on predefined distribution knowledge. Gaussian-based methods, exemplified by box plots and mean-variance calculations, offer a visual representation of data distribution characteristics, allowing for the identification of outliers. Regression-based methods, on the other hand, involve constructing models during the training phase and testing data points against these models during the test phase. Non-parametric methods, such as histogram-based and kernel-based approaches, provide alternatives for cases where the data distribution is unknown.

Although outlier detection is an interesting topic, we will not delve further to all the categories mentioned above. If someone wants to search more, he can start by reading the paper of Abir Smiti, 2020. We briefly mention only that Statistical-based methods

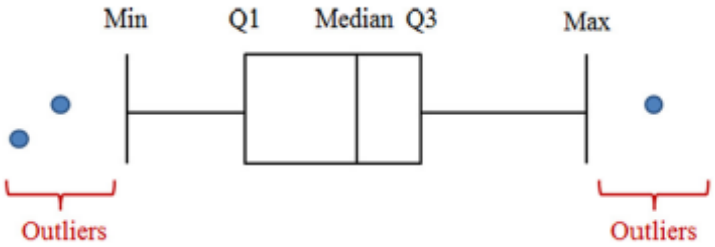

FIGURE 4.3: Boxplot that visualizes outliers (source: https://www.sciencedirect.com/science/article/pii/S1574013720304068)

may be effective for given distribution models, but they cannot be used when this distribution is not known. Distance-based techniques avoid this issue by being independent of data distribution yet, they may be very expensive when dealing with multivariate and high dimensional data. Density-based methods are more efficient, but they remain inappropriate for huge amounts of data and data streaming. While on the other hand, cluster-based techniques can manage data streams but require an excessive number of parameters.

Even though outlier detection has been the subject of numerous studies, there are still certain drawbacks with each approach. New outlier detection methods could be proposed, or existing methods could be improved.

4.2 Classification Methods

In machine learning, selection of the appropriate algorithm and the optimization of its performance depends on the type and characteristics of the available data, as well as on the nature of the problem faced Mehmed Kantardic (2020). Therefore, understanding the basic algorithms in both supervised and unsupervised learning is crucial. In the next units we will examine the two main categories of problems, each type is suitable, for supervised learning we have the classification (4.2) and regression (4.3) problems while for unsupervised learning clustering (4.4) and dimensionality reduction (4.5).

To begin with we start with classification, which is a fundamental problem in machine learning and involves predicting the class or category of data. The classification methods are algorithms that use labeled data to learn a boundary decision that can be used to classify new, unlabeled data into one or more predefined classes. An example of a real-world classification problem is when a doctor wants to decide, based on some specific characteristics, whether patients are suitable to have a surgery.

Overall, classification methods are powerful tools for addressing various real-world problems and can provide valuable insights for complex datasets. In this unit, we will try to analyze the methods mentioned below:

- Logistic Regression (4.2.1)
- Linear Discriminant Analysis (4.2.2)
- K-Nearest Neighbours (4.2.3)
- Support Vector Machines (4.2.4)

4.2.1 Logistic Regression

Continuous-value functions can be modeled using linear regression. Generalized regression models symbolize the theoretical basis on which the linear regression can be used to model categorical response parameters. Logistic regression is a popular form of generalized linear model. The chance of an event occurring as a linear function of a set of predictor variables can be calculated by logistic regression. It was developed to describe properties of population growth in ecology, rising quickly and maxing out, is presented in next figure Abdulhussein et al., (2021). Any real number can be mapped onto this S-shaped curve (Figure 4.4) to represent a value between 0 and 1.

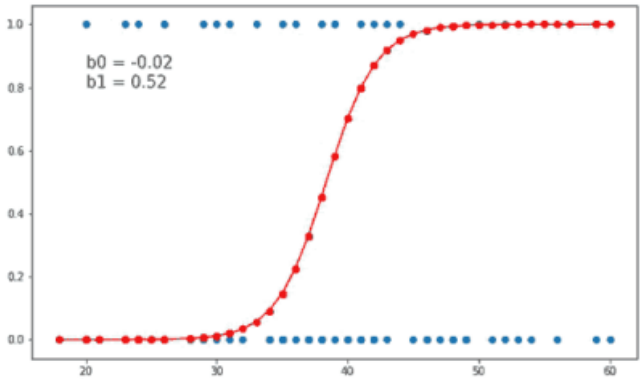

FIGURE 4.4: Logistic Regression S-curve (source: http://pen.ius.edu.ba/index.php/pen/article/viewFile/2507/1023)

Logistic regression is only applied when the output variable of the model can be categorized as a categorical binary variable, Kantardzic (2020). On the contrary, there is no reason why any of the inputs should not also be quantitative, and, thus, logistic regression can support a wide general input data set.

For instance, output Y has two possible categorical values 0 and 1. Based on the available data we can compute the probabilities for both values for the given input sample: $P(y_j = 0) = 1 - p_j$ and $P(p_j = 1) = p_j$. The model that we will fit these probabilities is accommodated linear regression:

$$log\left(\frac{p_j}{1-p_j}\right) = a + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \cdots + \beta_n X_{nj}$$

This equation is known as the linear logistic model. The function $log(\frac{p_j}{1-p_j})$ is usually symbolized also as logit(p). The main reason for using the logit form of output is to prevent the predicting probabilities from becoming values out of required range [0, 1]. Assume that a linear equation represents the estimated model that was created using the linear regression process and one training data set give us the result below:

$$A = logit(p) = 1.5 - 0.6\,x_1 + 0.4\,x_2 - 0.3\,x_3$$

and further assume that the input values for the new sample for classification are $\{x_1, x_2, x_3\} = \{1, 0, 1\}$. It is simple to calculate the probability of the output value 1, (P(Y = 1)) for this sample using the linear logistic model. First, finding the corresponding $logit(p) = 1.5 - 0.6\,1 + 0.4\,0 - 0.3\,1 = 0.6$ and then the probability of the output value 1 for the given inputs:

$$log\left(\frac{p}{1-p}\right) = 0.6 \Rightarrow p = log\left(\frac{e^{0.6}}{1+e^{0.6}}\right) = 0.65$$

We may determine that the output value Y = 1 is more likely than the other categorical value Y = 0 based on the final value for probability p usually, if the estimated probability is greater than 0.50 (threshold value), then the prediction is closer to YES. Logistic regression is a simple yet effective classification tool, as even this basic example demonstrates. One piece of data (the training set) can be used to create a logistic regression model, and another set of data (the testing set) can be used to assess how well the model predicts categorical values.

Logistic regression offers numerous advantages, including simplicity, interpretability, and the ability to handle numerical and categorical independent variables. However, it also comes with certain limitations, such as assuming linearity between independent variables and the logarithmic odds of the dependent variable, as well as sensitivity to extreme values and multicollinearity. Nevertheless, logistic regression remains a widely utilized and valuable tool in data analysis and machine learning.

4.2.2 Linear Discriminant Analysis

When there are categorical, nominal or ordinal, independent variables and metric dependent variables, linear discriminant analysis (LDA) is used to solve classification problems. The objective of LDA is to create a discriminant function that produces different scores when calculated with data from different output classes. The form of a linear discriminant function is $z = w_1 x_1 + w_2 x_2 + \ldots + w_k x_k$ where $x_1, x_2, \ldots x_k$ are

independent variables. Moreover, z is named discriminant score and $w_1 w_2, \ldots, w_k$ weights. The discriminant score (Figure 4.5) for a data sample represents its projection onto a line defined by the set of weight parameters as it demonstrated below.

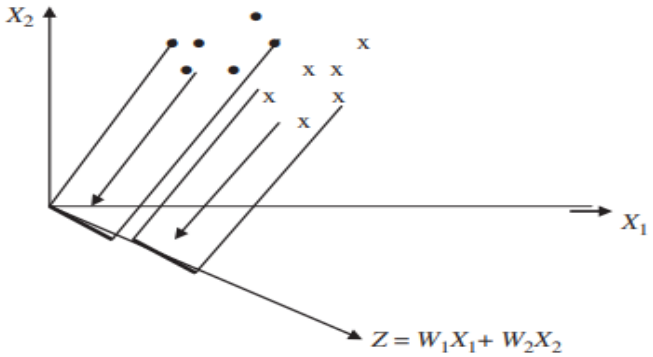

FIGURE 4.5: Geometric interpretation of the discriminant score, Kantardzic 2020

Finding a group of weight values wi that maximizes the ratio of the discriminant score between-class to within-class variance for a pre classified set of samples is the crucial step in creating the discriminant function z. Once constructed, the discriminant function z is used to forecast the class of a new non-classified sample. Cutting scores operate as the criteria against which each individual discriminant score is assessed the choice of cutting scores depends upon a distribution of samples in classes. Letting za and zb be the mean discriminant scores of pre classified samples from class A and B, respectively, the optimal choice for the cutting score zcut-ab is given as $Z_{cut\text{-}ab} = \frac{(Z\alpha - Zb)}{2}$ when the two classes of samples are of equal size and are distributed with uniform variance. A new sample will be classified to one or another class according to its score $z > Z_{cut\text{-}ab}$ or $z < Z_{cut\text{-}ab}$. While a weighted average of mean discriminant scores is implied when the group of samples for each of the classes is not the same size.

$$z_{cut-ab} = \frac{(n_a \cdot z_a + n_b \cdot z_b)}{(n_a + n_b)}$$

n_a and n_b indicate how many samples there are in each class. Although a single discriminant function z with a few discriminant cuts could distinctly sample into multiple classes, multiple discriminant analysis is preferred. Moreover, multiple discriminant analysis is employed in situations when separate discriminant functions are constructed for each class. In these circumstances, we choose the class whose discriminant score is the highest.

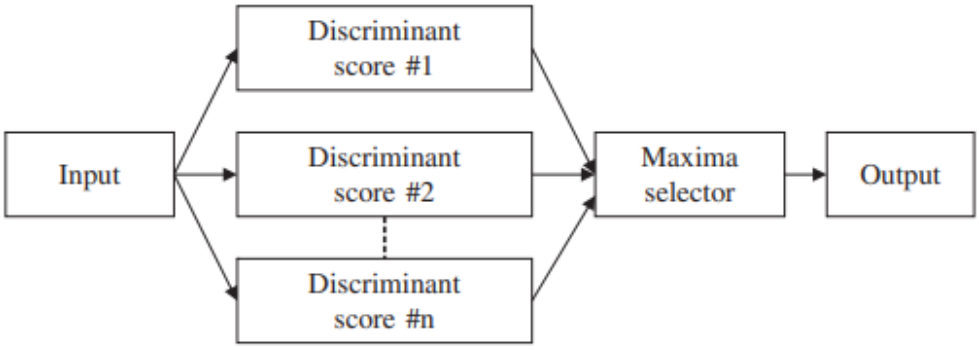

FIGURE 4.6: Classification process in multiple discriminant analysis, Kantardzic 2020

To conclude, LDA has several advantages, its ability to handle multiple classes and its tendency to produce more stable and interpretable results compared to other dimensionality reduction methods, are the worthiest mention. Also, the assumption that the data follow a normal distribution, which is a common one in many statistical models. However, LDA has certain limitations, such as its sensitivity to outliers and its assumption of equal covariance matrices across all classes. Additionally, we must highlight that it may be less effective when the number of classes is high or are not well separated.

4.2.3 K-Nearest Neighbours

The k-Nearest Neighbours (k-NN) technique provides a straightforward and understandable way for determining a molecule's class, property, or rank from its nearest training examples in the feature space by Lavecchia, (2015). K-NN is a form of lazy learning or instance-based learning in which all computations are postponed until classification and the function is only locally as it is shown in the proceeding table. Also, k-NN can be used for regression.

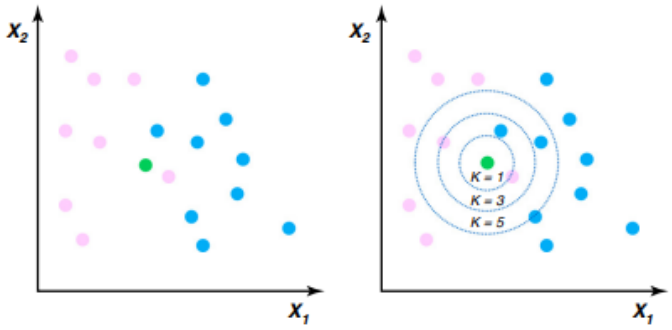

FIGURE 4.7: K-Nearest Neighbors on a 2D data set (source: https://www.sciencedirect.com/science/article/pii/S1359644614004176)

To clear things up, all that is needed for the k-NN classifier to determine distances in n-dimensional space is a metric measure, a set of labeled training samples, and a parameter k, Kantardzic (2020). The following steps are often the cornerstone of the k-NN classification process:

- Find the number of nearest neighbors, or parameter k.
- Determine the distance between every training sample and every testing sample.
- Using the kth threshold, sort the distance and identify the closest neighbors.
- Determine the category (class) for each of the nearest neighbors.
- Use simple majority of the category of nearest neighbors as the prediction value of the testing sample classification.

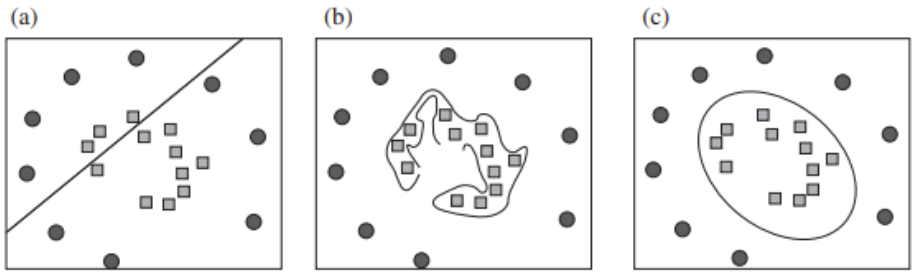

FIGURE 4.8: Trade-off between model complexity and the amount of data: (a) too simple (b) too complex (c) appropriate, Kantardzic 2020

These distance metrics aid in shaping decision boundaries that separate the studied points into different categories. Various distance measures have created, to optimize this process with the most frequent used listed below:

- **Euclidean Distance**: This distance metric is limited to vectors of real values. Using the formula below, it measures a straight line between the point under study and another reference point.

$$\boldsymbol{d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}}$$

- **Manhattan Distance**: Another popular distance metric, that calculates the absolute value between two points. It is also referred to as "taxicab distance" or "city block distance," as it is often represented on a grid, depicting how one can navigate from one address to another through city streets.

$$\boldsymbol{d(x,y) = (\sum_{i=1}^{m}|x_i - y_i|)}$$

- **Minkowski Distance**: This distance metric is the generalized form of Euclidean distance and Manhattan distance. In the formula below, the parameter *p* allows the creation of other distance metrics. When *p*=2, Minkowski Distance corresponds to the Euclidean distance, while *p*=1 corresponds to the Manhattan distance.

$$\boldsymbol{d(x,y) = \left(\sum_{i=1}^{n}|x_i - y_i|\right)^{1/p}}$$

There are numerous methods for enhancing k-NN functionality and speed. Selecting a subset of the training data for classification is one way. Analytically, the idea of the condensed nearest neighbor (CNN) is to choose the smallest subset Z of training data X such that when Z is used instead of X, error in classification of new testing samples does not increase. k-NN (with k=1) is used as the nonparametric estimator for classification. It approximates the classification function in an individually linear manner with only the samples that define the classifier need to be kept. Since they are members of the same class, samples inside regions do not require storage. An example

of CNN classifier in 2D space is given in the table here after (Figure 4.9). Greedy CNN algorithm is defined from the steps listed below:

- Begin with a blank set Z.
- Passing samples from X one at a time in a random order to see if they can be accurately sorted by Z instances.
- If a sample is misclassified, it is added to Z if it is correctly classified, Z is unchanged.
- Continue using the training data set several times until Z remains the same. The algorithm does not ensure that Z has a minimum subset.

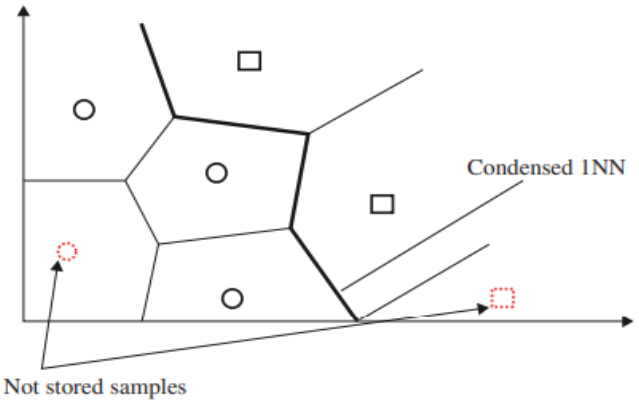

FIGURE 4.9: CNN classifier in 2D space, Kantardzic 2020

Finally, k-NN has many advantages for classification tasks, including simplicity and its ability to handle non-linear relationships between features and the target variable. However, k-NN also has its limitations, such as the sensitivity of k selection, which may affect the performance of the algorithm. It is also computationally expensive and memory intensive, especially when dealing with big data. Nevertheless, k-NN remains a widely used and useful tool in machine learning especially for classification tasks and it is highly useful for complex attributes and target variables.

4.2.4 Support Vector Machines

Supervised machine learning methods (SVMs) were created mainly by Vapnik, 2000 and enable compound classification, ranking, and regression-based property value prediction. SVMs are typically employed for binary property or activity predictions, such as differentiating between chemicals with or without a particular activity or between drugs and non-drugs.

SVM needs a relatively small number of samples for training, and experiments showed that it is insensitive to the number of sample's dimensions. The program first tackles the broad issue of learning to distinguish between individuals belonging to two classes that are represented as n-dimensional vectors. The function can be a classification function the output is binary, or the function can be a general regression function. The idea of decision planes, which specify the boundaries between decision

classes for samples, is the cornerstone of SVM function. A basic example is given below.

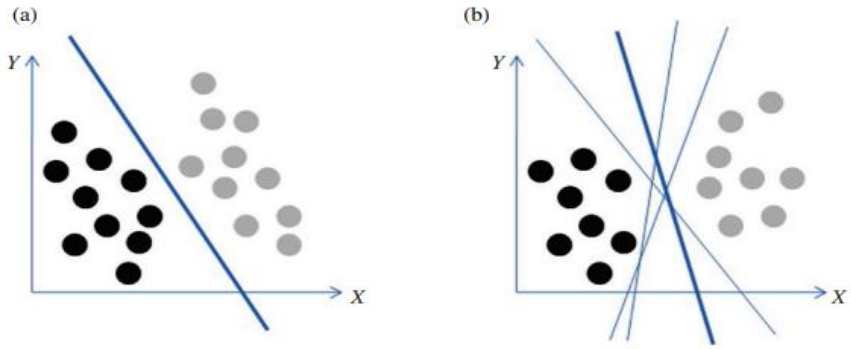

FIGURE 4.10: (a) A decision plane in 2D space is a line (b) How to select optimal separating line (hyperplane), Kantardzic 2020

The decision boundary should be as far away from the data points of both classes as possible, according to the fundamental notion. In other words, finding the ideal hyperplane to divide classes of n-dimensional vectors is the aim of SVM modeling in n-dimensional domains.

Next, we will examine the Kernel Trick which involves representing the data points in a higher-dimensional space compared to the original dimensions Kantardzic (2020). For instance, a 1D data point can be elevated to a 2D representation in space, and similarly, a 2D dataset can be projected into a 3D space, and so on. SVM adeptly handles non-linear data points by utilizing various kernel functions, making it seem as though the data has undergone a transformation. The SVM then identifies the optimal separating hyperplane in this higher-dimensional space. It is crucial to note that despite this apparent transformation, the underlying data points remain unchanged. The kernel trick allows SVM to intelligently navigate non-linearities without physically altering the data. Some of the most known, kernel functions are presented next (Figure 4.11)

Name of the Kernel	**Function**
Linear	$k(x,y)=x^T y$
Polynomial	$k(x,y) = (x^T, y)^P$ or $k(x,y) = (x^T y + 1)^p$ *where p is the polynomial degree*
RBF(Gaussian)	$\varphi(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right), \sigma > 0$

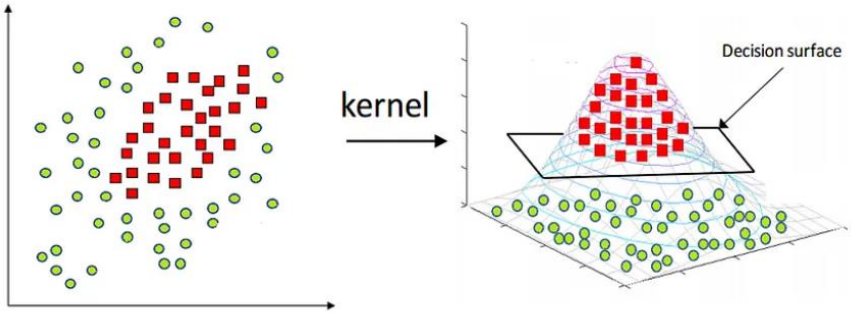

FIGURE4.11: a) Most known Kernel Functions b) Altering data perception (source: https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-important-98a98db0961d)

Numerous benefits are provided by the SVM for classification tasks. Because kernel functions are used, it can handle separable data that is both linear and nonlinear. Also, because SVM aims to optimize the margin between classes rather than tightly fitting the data, it is also less prone to overfitting than other techniques. SVM does, however, have several drawbacks, such as its sensitivity to the selection of the kernel function and algorithm hyperparameters, such as the kernel parameters and regularization parameter. Additionally, SVM can be computationally expensive, particularly for complex kernel functions or when dealing with big data. Lastly, because the hyperplane of the high-dimensional feature space might not exactly match a straightforward decision boundary in the original feature space, making the understanding of SVM findings a difficult task.

4.3 Regression Methods

Regression methods are a category of machine learning techniques used to model the relationship between a dependent variable and one or more independent variables. The goal of regression analysis is to find the parameters of the model that better predict the dependent variable for a set of input variables. Regression methods are commonly employed for prediction, forecasting, and understanding the relationship between variables in a dataset.

In supervised learning, regression methods involve training a model on a labeled dataset, where the algorithm learns from input-output pairs to make predictions on new, unseen data. The selection of regression method depends on the characteristics and the assumptions that align with the underlying relationships between variables.

Overall, regression methods are crucial in extracting meaningful insights and making accurate predictions in various fields. In the upcoming unit, we examine the following methods as it shows:

- Linear Regression (4.3.1)
- Ridge Regression (4.3.2)
- Lasso Regression (4.3.3)
- Partial Least Square Regression (4.3.4)

4.3.1 Linear Regression

Linear regression is a statistical technique used to model the linear relationship between a dependent and one or more independent variables. It is a simple and widely used regression method that assumes a linear relationship between variables. There are two types of linear regression: the simple and the multiple. In simple linear regression the goal is to locate the best-fitting line to describe the relationship between the independent and dependent variable. While in the multiple linear regression we search for the best-fitting hyperplane. In linear regression, the relationship between independent variables and the dependent variable is modeled using a linear equation of the form: $Y = a + \beta Xi$. The Y is the dependent variable, $x_1, x_2, \ldots, x_n$ are the independent variables, β_0 is the intercept, and $\beta_1, \beta_2, \ldots, \beta_n$ are the slope coefficients representing the change in y for a unit change in each of the independent variables.

The equation that needs to be minimized is: $\sum_{i=1}^{v} \varepsilon_i^2 = \sum_{i=1}^{v} (y_i - a \pm \beta x_i)^2$

The values of α and β that minimize the above equation are called unbiased estimators of least squares and are calculated using: $\bar{y} = \frac{1}{v}\left(\sum_{i=1}^{v} y_i\right)$, $\bar{x} = \frac{1}{v}\left(\sum_{i=1}^{v} x_i\right)$ the following relationships with: $\hat{a} = \bar{y} - \hat{\beta}\bar{x}$ and

$$\hat{\beta} = \frac{v\sum_{i=1}^{v} x_i y_i - (\sum_{i=1}^{v} x_i)(\sum_{i=1}^{v} y_i)}{v\sum_{i=1}^{v} x_i^2 - \left(\sum_{i=1}^{v} x_i\right)^2}$$

The goal of linear regression is to estimate the values of coefficients that minimize the sum of squared differences between the predicted values and the actual values of the dependent variable also known as the method of least squares. Linear regression is the best solution if you want simplicity, interpretability, and robustness to outliers. It is also a useful tool for making predictions and understanding the relationship between variables. However, linear regression has also its drawbacks, such as the assumption of linearity, which may not hold for certain datasets, and its inability to capture non-linear relationships between variables. Linear regression is employed mainly for prediction, forecasting, and hypothesis testing.

4.3.2 Ridge Regression

Next, we move to ridge regression, which shrinks the regression coefficients by imposing a penalty on their size. Ridge regression is usually used when the independent variables are highly correlated, thus we avoid overfitting, and we improve the accuracy of our model. Also, the ridge coefficients minimize a penalized residual sum of squares. Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage Hastie et al., 2009.

$$\hat{\beta}^{ridge} = arg_{\beta} min \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{P} {\beta_j}^2 \right\}$$

Therefore, ridge regression tries to balance between, fitting the data well and keeping the model's complexity in check. Furthermore, ridge regression can handle situations where the number of predictors is greater than the number of observations or when the predictors exhibit multicollinearity. In such cases, the ordinary least squares (OLS) method may lead to unstable and unreliable coefficient estimates. Ridge regression mitigates this problem by introducing regularization, which stabilizes the model and prevents the coefficients from becoming overly sensitive to small changes in the data. By minimizing the penalized residual sum of squares, ridge regression offers a more robust and reliable solution for linear regression problems with correlated predictors.

One benefit of ridge regression is that it helps mitigate multicollinearity, a common problem when dealing with independent variables in a regression model that are highly correlated. Moreover, ridge regression introduces a regularization term that prevents the coefficients from becoming too large, thus stabilizing the model, and improving its overall performance. On the other hand, one limitation of ridge regression is that it assumes all independent variables that are significant for the model. In practice, some variables may be irrelevant or have a negligible impact on the outcome. For these problems, a more advanced method is preferred called Lasso regression, which we will discuss shortly.

4.3.3 Lasso Regression

The Least Absolute Shrinkage and Selection Operator (LASSO) is a shrinkage method like Ridge, with subtle but important differences. This method is proposed by statistician Robert Tibshirani in 1996 and is defined by:

$$\hat{\beta}^{lasso} = arg_{\beta} min \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \right\}$$

Notice the similarity to the ridge regression equation: the L_2 ridge penalty $\sum_{j=1}^{P} {\beta_j}^2$ is replaced by the L_1 lasso penalty $\sum_{j=1}^{P} |\beta_j|$.

In the context of an orthonormal input matrix X, both methods (Lasso & Ridge) have explicit solutions involving transformations to the least squares estimate β_j. Ridge regression achieves proportional shrinkage, while the lasso employs soft thresholding by translating each coefficient with the λ and truncating at zero. In the visualizations that follows, illustrate the differences between lasso and ridge regression. For two parameters, ridge regression's constraint region is depicted as a disk, while lasso as a diamond. Both methods identify the first point where elliptical contours of the residual sum of squares intersect the constraint region. The diamond's corners, unique to lasso, signify instances where one parameter (β_j) equals zero. As the number of parameters (p) increases, the diamond becomes a rhomboid, providing more opportunities for

estimated parameters to be precisely zero. We can generalize Lasso & Ridge and view them as Bayes estimates with q ≥ 0 consider the criterion:

$$\tilde{\beta} = argmin \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{P} |\beta_j|^q \right\}$$

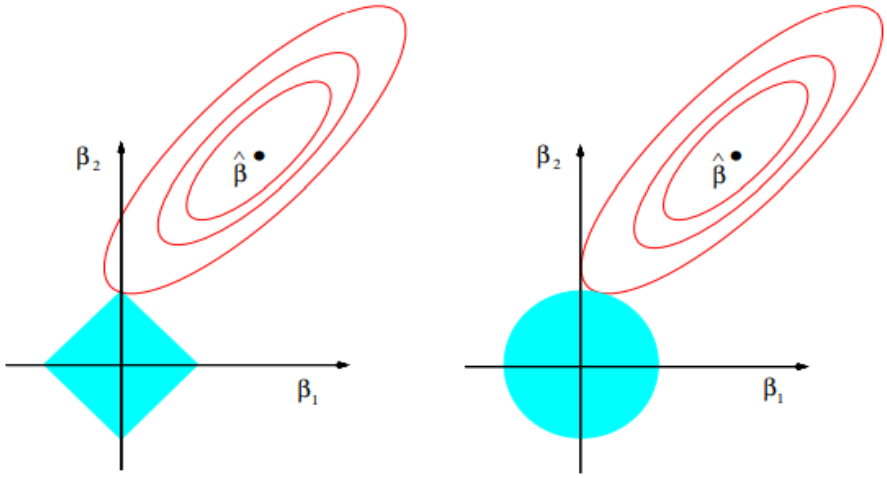

FIGURE 4.12: Estimation picture for the lasso (left) and ridge regression (right) (source:https://hastie.su.domains/Papers/ESLII.pdf)

4.3.4 Partial Least Squares Regression

Partial Least Squares (PLS) regression is a designed for situations where the number of predictors is large relative to the number of observations, a scenario often referred to as the "small N large P problem." Its primary goal is to analyze or predict a set of dependent variables from a set of independent variables or predictors by Abdi, (2010). In other words, the primary objective of PLS regression is to predict Y from X and describe their common structure usually is preferred when traditional multiple regression becomes impractical cause multicollinearity is prevalent.

A briefly overview of how the algorithm works is shown right below:

- **Simultaneous Decomposition:** PLS simultaneously decomposes both X and Y into latent vectors and specific loadings.
- **Latent Vectors:** PLS identifies a set of latent vectors, often called latent variables or components, that capture the maximum covariance between the predictor and response variables.
- **Score and Loading Matrices:** X is decomposed into a score matrix (T) and a loading matrix (P), while Y is estimated as a product of a score matrix (U), regression weights (B), and a weight matrix (C).
- **Regression Step:** The latent vectors obtained from X are then used to predict Y through a regression step, involving regression weights and the weight matrix.

- **Optimal Covariance:** PLS aims to find latent vectors that explain as much as possible of the covariance between X and Y. This makes PLS effective in situations where there are more predictors than observations, and it provides a means of predicting the dependent variables.

Therefore, PLS regression is particularly effective in situations where multicollinearity is present, a common challenge when dealing with many predictors. Also, it excels when predicting a set of dependent variables from a large set of independent variables, making it suitable for modern data analysis domains like bioinformatics and data mining. PLS regression has found applications in various fields, including economics, chemometrics and social sciences. Tus it has become a versatile tool in both experimental and nonexperimental data analysis. Despite its strengths, PLS regression also has some limitations. One drawback is the potential for overfitting, especially when the number of predictors is much larger than the number of observations. Overfitting may lead to a model that performs well on the training data but generalizes poorly to new, unseen data. Another limitation lies in the interpretability of the latent vectors. While PLS regression excels in predictive modeling, the interpretability of the latent vectors may be challenging, making it less suitable for scenarios where a clear understanding of the underlying relationships is essential.

4.4 Clustering Methods

Grouping data sensibly is a fundamental aspect of learning. Cluster analysis, is the formal examination of methods and algorithms for the natural grouping of objects based on intrinsic characteristics or similarity, is integral to this process. In other words, clustering, which is the most prevalent unsupervised learning task, involves identifying a finite set of categories or clusters to describe data. Overall, the final clusters are defined by general characteristics, and the solutions can be different according to which clustering technique we choose. Following the clustering process, new samples can be assigned to previously identified clusters based on their similarity to the cluster characteristics. Clustering poses a significant challenge as data can unveil clusters of different shapes and sizes within an n-dimensional data space.

Each type of clustering algorithm has its own strengths and weaknesses, and the choice often depends on the specific problem and the characteristics of the data. There are several types of clustering algorithms and in this section, we will briefly examine the following:

- Hierarchical Clustering (4.4.1)
- K-means Algorithm (4.4.2)
- DBSCAN Algorithm (4.4.3)

4.4.1 Hierarchical Clustering

In hierarchical cluster analysis, we do not specify the number of clusters as a part of the input. In particular, the input to a system is (X, s), in which X is a set of samples

and s is an index of similarity. Therefore, most hierarchical clustering processes are not based on the idea of optimization but instead, they aim to discover a roughly suboptimal solution by repeatedly improving the divisions until convergence. Overall hierarchical cluster analysis algorithms fall into two main categories: agglomerative and divisible algorithms.

To begin with, a divisible algorithm starts from the entire set of samples X and divides it into a partition of subsets, then divides each subset into smaller sets, and so on. As a result, a divisible algorithm produces a series of partitions ordered from coarser one to a finer one. On the other hand, in an agglomerative algorithm, every object serves as an initial cluster at the beginning. The clusters are combined into a coarser partition, and this merging process is continued until the result is the trivial partition, which consists of one big cluster containing all the objects. This clustering procedure separates data from a finer to a coarser level in a bottom-up method. Since agglomerative algorithms are more commonly employed in practical settings than divisible algorithms, we will dive into more detail on the agglomerative methods.

Most agglomerative hierarchical clustering algorithms are variants of the single-link or complete-link algorithms. These two algorithms vary only in the way they define the similarity between a pair of clusters. For instance, the single-link approach calculates the distance between two clusters as the minimum of all the pairs of samples (one from each of the two clusters, one from each element) that are selected from them. While in the complete-link algorithm, the distance between two clusters is the maximum of all distances between all pairs drawn from the two clusters. To further understand we present a relevant graphic.

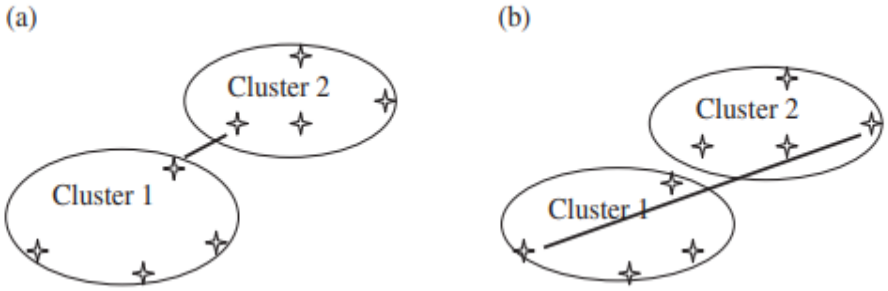

FIGURE 4.14: (a) Single-link distance & (b) Complete-link distance, Kantardzic 2020

Now that we discuss their differentiation, we can examine the agglomerative-clustering fundamental steps that are the same for both.

- Arrange the samples into separate clusters. Construct the list of inter-cluster distances for all distinct unordered pairs of samples and sort this list in ascending order.
- Proceed through the sorted list of distances, creating a graph of the samples where pairs of samples closer than dk are joined to form a new cluster by a graph edge for each unique threshold value dk. If all the samples are members of a connected graph, stop. If not, go back and do this again.
- The output of the algorithm is a nested hierarchy of graphs, which can be cut at the desired dissimilarity level forming a partition (clusters) identified by simple connected components in the corresponding subgraph.

Agglomerative clustering offers several benefits, making it a versatile method for exploring data relationships. One notable advantage is its intuitive hierarchy, represented through a dendrogram, facilitating a clear understanding of the hierarchical structure of clusters. Another strength lies in its flexibility regarding the number of clusters, as users can determine the appropriate quantity based on the dendrogram. The method's adaptability to different linkage methods and its insensitivity to the shape and size of clusters further enhance its applicability to diverse datasets.

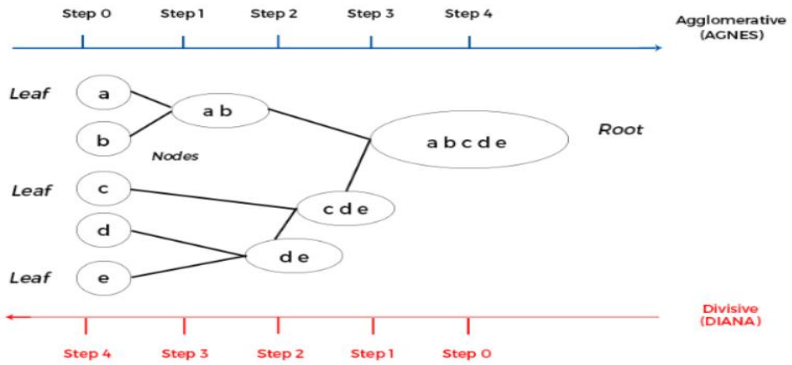

FIGURE 4.15: Agglomerative and Divisible algorithms steps (source:https://www.researchgate.net/figure/Hierarchical-clustering-structure-1-Divisive-Hierarchical-Clustering-Algorithm-Division_fig5_315747115)

However, agglomerative clustering comes with its set of disadvantages as it is computationally expensive, limiting its scalability for large datasets. Additionally, the method may struggle with noisy data and outliers, as their influence can significantly impact the resulting clusters. Moreover, the non-reversibility of mergers poses a challenge for adjusting or refining clustering later in the analysis. These factors mainly highlight the trade-offs involved in choosing agglomerative clustering for data analysis.

4.4.2 K-means Algorithm

One of the most often used clustering techniques is the K-means algorithm. The aim of the K-means algorithm is to partition the data into K clusters so that the within-group sum of squares is minimized. The simplest form of the K-means algorithm is based on alternating two steps. The assignment of objects to groups is the first. In general, an object is placed in the group whose mean is the closest in Euclidean sense. The second action is the calculation of new group means based on the assignments. For each cluster we calculate the mean with the following type:

$$K_j = \frac{1}{n_j} \sum_{x_i \to K_j} x_i$$

When moving an item to a different group does not result in a decrease in the within-group sum of squares, the process comes to an end.

There are many variants of the K-means algorithm that improve its efficiency in terms of reducing the computing time and achieving a smaller error. Some algorithms allow new clusters to be created and existing ones to be deleted during the iterations. Others may move an object to another cluster based on the best improvement in the objective function, by Larrañaga et al., 2006. Alternatively, the first encountered improvement while passing by the dataset could be used.

The algorithm is easy to apply and can handle large datasets. Moreover, the number of K-clusters can be adjusted depending on the problem at hand, allowing flexible clustering based on the desired level of detail. However, K-means clustering also has some limitations such as the sensitivity to the initial selection of means, which can lead to the formation of different clusters depending on the starting point. Therefore, multiple runs to ensure stability is a necessity. In addition, the K-means clustering assumes that the clusters are spherical and have similar deviations, which may not always be the case.

4.4.3 DBSCAN Algorithm

The algorithm Density-Based Spatial Clustering of Applications with Noise (DBSCAN) targeting low-dimensional data is the major representative in the category of density-based clustering algorithms, Kantardzic (2020). DBSCAN can recognize the clusters mainly because within each cluster, we have a typical density of points that is considerably higher than outside of the cluster.

DBSCAN, is grounded in the principles of density reachability and density connectivity, guided by two crucial parameters: the neighborhood size (ε) and the minimum points in a cluster (m). The algorithm operates on the premise that for each point in a cluster, its ε neighborhood must encompass at least m points to surpass a predefined density threshold. This ensures that the density of points within the neighborhood is sufficiently high. Furthermore, if a point like p has only two neighbors within its ε radius, while another point q has eight, the density around q is deemed higher than around p, forming the basis for cluster identification.

The density reachability concept dictates that two points, p_1 and p_2, are considered density reachable if they are close (distance $(p_1, p_2) < \varepsilon$) and if there are enough points in the ε neighborhood of p_2 (distance $(r,p_2) > m$), where r denotes other database points. Building upon this, density connectivity establishes relationships between points, declaring points p_0 and p_n as density connected if there exists a sequence of density reachable points $(p_0,p_1, p_2, \dots)$ from p_0 to p_n, where each subsequent point $(p_i + 1)$ is density reachable from the previous one. This way forms the fundamental steps for DBSCAN clusters, representing sets of all density-connected points and enabling the algorithm to uncover clusters of varying shapes while handling noise.

If a point has more than a certain number of points (m) within neighborhood ε, it is considered a core point (points in a cluster's interior). A border point is located close to a core point but has fewer than m points in its neighborhood (ε). Any point that is neither a border nor a core point is considered a noise point. To illustrate this, we attach the next diagram.

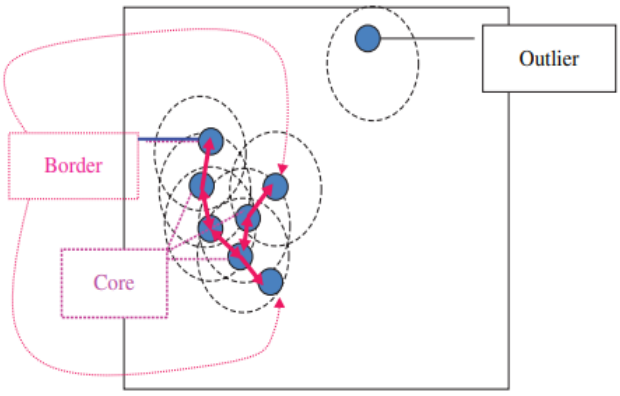

FIGURE 4.16: Examples of core, border, and noise points, Kantardzic 2020

In an ideal world, we would have to know each cluster's proper parameters, ε and m, but there is no simple way to find out in advance. Thus, DBSCAN uses the same values for all the clusters. Beyond that, studies show that, although the approach requires significantly more calculations, DBSCAN clusters for m > 4 do not differ significantly from m = 4. For low-dimensional databases, we can therefore set the parameter to 4 to eliminate it in practice. The DBSCAN algorithm's key phases are:

- Random pick a point, p.
- Get every point density that may be reached from p in connection with ε and m.
- A new cluster is created, or an existing cluster is expanded if p is a core point.
- If p is a border point, and no points can be densely reached from p, then move on the next point.
- Once every point in the database has been processed, continue the procedure with the remaining points.
- Since DBSCAN uses global values for ε and m, two clusters may be merged into one cluster if the distance between is less than ε.

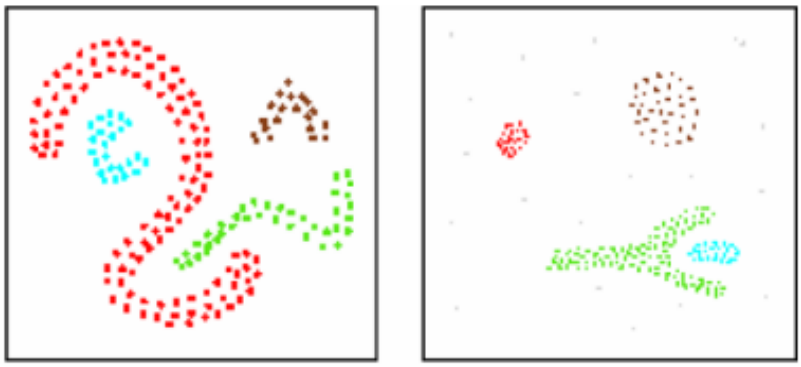

FIGURE 4.17: DBSCAN builds clusters of different shapes, Kantardzic 2020

The DBSCAN clustering algorithm presents several distinct advantages firstly, it eliminates the need to determine the number of clusters beforehand, setting it apart from clustering methods like K-means, Kantardzic (2020). Additionally, DBSCAN excels in identifying clusters with arbitrary shapes, a feature not shared by many popular clustering algorithms. Moreover, it incorporates a concept of noise, effectively removing outliers from clusters. Furthermore, the algorithm's simplicity is evident in

its requirement for just two parameters, displaying robustness against variations in the order of points within the database.

On the contrary, DBSCAN is not without its drawbacks. The algorithm's complexity remains notably high, posing a challenge for certain applications. Specifically, the operation of finding neighbors relies on distance calculations, often using Euclidean distance, and this may exacerbate the curse of dimensionality problem, particularly in high-dimensional datasets. Consequently, while DBSCAN proves advantageous in handling low-dimensional real-world data, its application to high-dimensional datasets may be limited.

4.5 Dimensionality Reduction

Dimensionality reduction techniques are employed when dealing with high-dimensional datasets to make them more manageable while preserving the integrity of the data. One commonly used method in data preprocessing is Principal Component Analysis (PCA). In unsupervised learning, reducing dimensionality involves decreasing the number of features or variables in a dataset while retaining crucial information. This process is valuable for visualizing high-dimensional data in a lower-dimensional space. Although there are various other methods, PCA is the most famous technique for dimensionality reduction. The choice of technique depends on the specific dataset and problem at hand, but PCA is usually preferred at most cases and in the next paragraph we will examine why.

4.5.1 PCA

PCA is considered to be the best linear dimension reduction method as we said, mainly because it uses the covariance matrix of the features. In simple terms, PCA seeks to decrease the dimension of the data by locating a few orthogonal linear combinations of the original features with the largest variance. It is preferred to first normalize each variable to have a mean equal to zero and a standard deviation one, because the variance depends on the scale of the variables. After the standardization, the original variables with potentially different units of assessing are all in similar units.

The fundamental theory is as follows, a set of n-dimensional vector samples X = $\{x_1, x_2, \dots, x_m\}$ should be transformed into another set Y= $\{y_1, y_2, \dots, y_m\}$ of the same dimensionality, but Y have the property that most of their information content is stored in the first few dimensions. This will minimize the information loss, while we are reducing the data dimensions. Thus, if we want to reduce a set of input dimensions X to a single dimension Y, we should transform X into Y as a matrix computation (Y = A × X) choosing A such that Y has the largest variance possible for a given data set. The single dimension Y is known as the first principal component. An axis pointing in the direction of maximum variance makes up the first main component. As seen in Figure 3, which shows the transformation of a two-dimensional space into a one-dimensional space where the data set has the maximum variance, it minimizes the distance of the sum of squares between data points and their projections on the component axis.

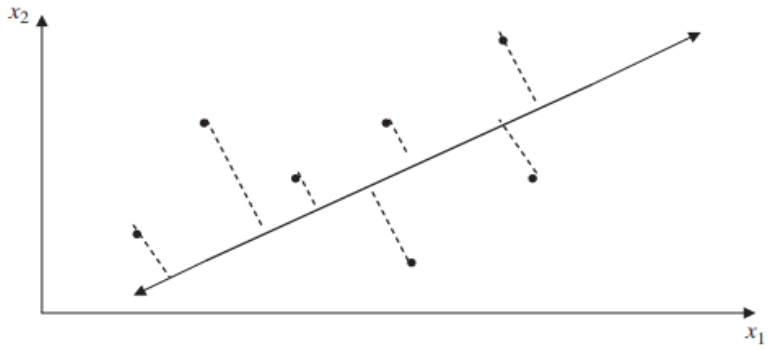

FIGURE 4.18: The first principal component is an axis in the direction of maximum variance, Kantardzic 2020

In practice, it is not possible to determine matrix A directly, and therefore we compute the covariance matrix S as a first step in feature transformation. Matrix S is defined as:

$$\underset{\text{nxn}}{\text{S}} = \frac{1}{(\text{n}-1)}\left[\sum_{\text{j=1}}^{\text{n}}(\text{x}_\text{j}-\text{x}')^{\text{T}}(\text{x}_\text{j}-\text{x}')\right]$$

To continue, we calculate the eigenvalues of the covariance matrix S for the given data. Finally, the m eigenvectors corresponding to the m largest eigenvalues of S define a linear transformation from the n-dimensional space to an m dimensional space in which the features are uncorrelated.

Furthermore, determining the optimal number of principal components is a necessity for a meaningful representation of the data. To address this, analyzing the proportion of variance becomes crucial. By dividing the sum of the first m eigenvalues by the total sum of variances (all eigenvalues), we obtain a measure of the representation quality based on the first m principal components. This result, expressed as a percentage, can be satisfactory if it is more than 90% of the total variance.

$$R = \frac{(\sum_{i=1}^{m}\lambda_\text{i})}{(\sum_{i=1}^{n}\lambda_\text{i})}$$

Although PCA offers advantages such as simplifying complex datasets, aiding in visualization, and facilitating dimensionality reduction, which can enhance computational efficiency and mitigate the risk of overfitting. It is important to note that PCA assumes that the data is related linearly and may not work well with non-linear data. Furthermore, interpretation of principal components can be difficult, especially when dealing with large data sets with many variables.

4.6 Neural Networks

Neural networks (NN) are designed as computational models to replicate the functioning of the brain. The brain consists of small functional units known as neurons, each comprising a cell body, multiple short dendrites, and a single long axon. Neurons connect to one another through dendrites and axons. Dendrites serve as inputs to the neuron, receiving signals from other neurons. These inputs either increase or decrease

the electrical potentials of the cell body, and if a threshold is reached, an electrical pulse is transmitted down the axon. This output then becomes the input for several other neurons.

Likewise, an artificial neural network (ANN) is constructed from computational units, often referred to as neurons. There are linkages connecting these units, and each link has a weight. Long-term memory is analogous to weights. Like a real neuron, each unit gets information from input links. Next, each unit determines the weighted total a final value that serves as the unit's output is transformed using a transfer function and the input values. Next, figure depicts a basic neural network model.

The block diagram, which is a model of an artificial neuron, consists mainly of three basic elements:

1. A set of connecting links from different inputs x_i (or synapses), each of which is characterized by a weight or strength w_{ki}. The first index refers to the neuron in question, and the second index refers to the input of the synapse to which the weight refers. In general, the weights of an ANN may lie in a range that includes negative as well as positive values.
2. An adder for summing the input signals x_i weighted by the respective synaptic strengths w_{ki}. The operation described here constitutes a linear combiner.
3. An activation function f for limiting the amplitude of the output y_k of a neuron.

The model of the neuron given in Figure 4.19 also has an externally applied bias, symbolized bk. The bias has the effect of increasing or lowering the net input of the activation function, depending on whether it is positive or negative. To make it simple, an artificial neuron is an abstract model of a natural neuron, and its processing capabilities are formalized using the following notation. First, there are several inputs x_i $i = 1, \ldots m$. Each input xi is multiplied by the corresponding weight wki where k is the index of a given neuron in an ANN. The weights represent the biological synaptic strengths in a natural neuron.

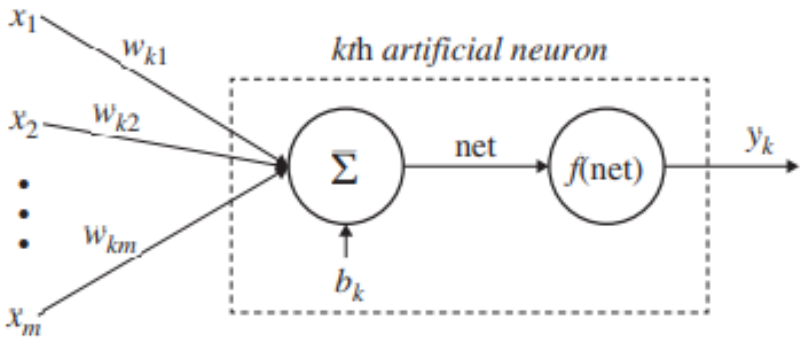

FIGURE 4.19: Representation of a basic neural network model , Kantardzic 2020

4.6.1 Architecture of Neural Networks

In general, network architecture is defined by the total amount of inputs to the network, the quantity of outputs, the whole number of elementary nodes that are usually equal processing elements for the entire network, and their organization and interconnections. Based on the kind of connections, neural networks are typically divided into two groups: feedforward and recurrent.

The network is feedforward if the processing propagates from the input side to the output side unanimously, lacking any cycles or feedback. In a layered representation of the feedforward neural network, there are no links among nodes in the same layer; outputs of nodes in a specific layer are always linked as inputs to nodes in succeeding layers. On the other hand, A network is considered recurrent if a feedback connection creates a cyclical path within it, sometimes with a delay element acting as a synchronization component. The most popular model in terms of real-world applications is the multilayer feedforward network with a backpropagation-learning mechanism, even though many neural-network models have been published in both categories.

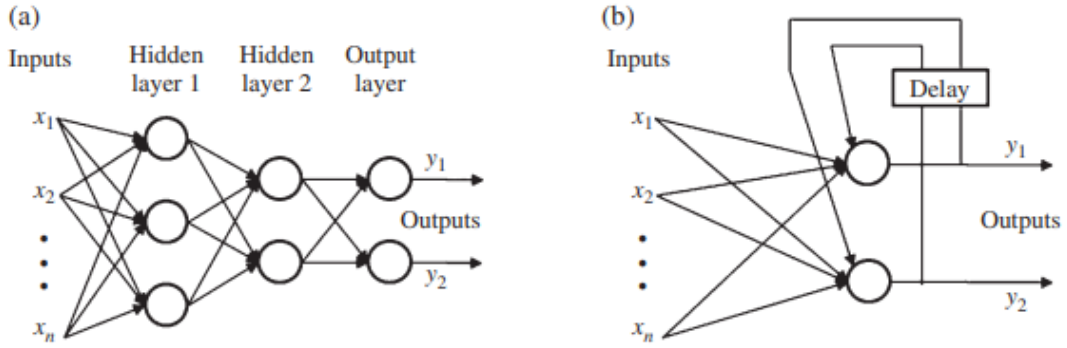

FIGURE 4.20 (a) Feedforward network (b) Recurrent network, Kantardzic 2020

There are several common network architectures used in bioinformatics which have unique application. Perceptron is for example is the simplest form of neural network which has only two layers, the input layer, and the output layer. Because they can only be used to categorize patterns into one of two classes, perceptron has very limited applications, Kantardzic (2020). Moving to multi-layer perceptron (MLP) perceptron with more than two layers of neurons. MLP has an input layer, one or more hidden layers and an output layer. Normally, the hidden and output layer's transfer function is either a logistic or sigmoid function. A fully linked network is typically made up of all the neurons in the layer above it, however there are occasionally exceptions. By dividing the data into separate sections using hyper-planes, MLP can classify the given data set.

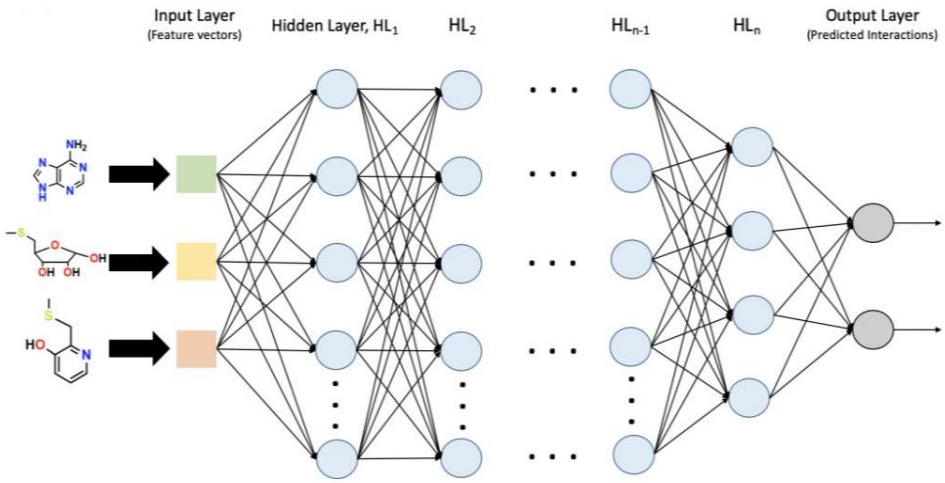

FIGURE 4.21: A multilayer-perceptron architecture with n hidden layers, (source: https://pubmed.ncbi.nlm.nih.gov/33198233/)

Next there is the architecture of the radial basis function, which is quite alike with the MLP, but the principle of action and training is different from MLP. The data can be clustered into a limited number of ellipsoid areas using the radial basis function. Normally, a transfer function is one of Splice, Gaussian, or other quadratic functions. Each hidden unit of the network acts as the center of the region. Inputs to these units are not a weighted sum but a distance measure with the Euclidean function being the most often used. After that, the output is calculated by the hidden unit as a function of the input vector and its center.

Compared to earlier networks, Kohonen self-organizing maps (SOM) are significantly different. While Kohonen self-organizing maps feature an input layer, they lack an output or hidden layer. A grid of discrete units is connected to input layer units. They have complete connections and links connected with weights. By calculating the distance matrix for each grid point and selecting the point that most closely matches the input, the input vectors are transferred to one of the grid points.

In other words, SOMs are construed as unsupervised neural networks designed to address clustering problems through cluster visualization. Through a learning process, SOM serves as a crucial tool for visualization and data reduction, providing a comprehensive overview of the data. This involves transforming similar data items into a lower dimension, automatically grouping them together for enhanced understanding and analysis. Therefore, SOMs suits best if we want to reduce dimensions and display similarities.

The primary benefits of SOM technology include results that are straightforward to comprehend and interpret, simplicity in implementation, and effectiveness in addressing numerous practical issues. However, there are drawbacks as well. SOMs involve significant computational costs, exhibit high sensitivity to similarity measures, and are not suitable for real-world datasets containing missing values.

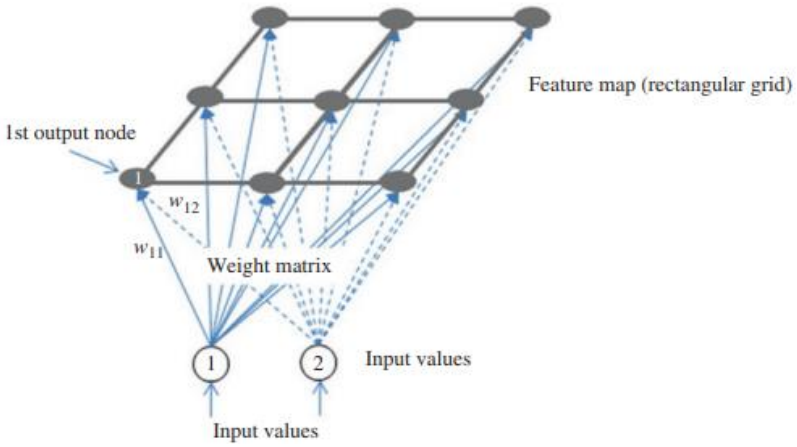

FIGURE 4.22. SOM with 2D input and 3 × 3 output, Kantardzic 2020

5 CHAPTER

Application of Machine Learning Techniques

5.1 Purpose of the Analysis

In contemporary society, diabetes looms as a pervasive health concern, with its prevalence reaching alarming heights. The modern lifestyle, characterized by sedentary habits and diets rich in processed foods, has fueled an epidemic of both type 1 and type 2 diabetes. Type 2 diabetes has seen a surge in cases, often attributed to factors such as obesity, poor dietary choices, and blood pressure. This rise not only places a significant strain on healthcare systems but also underscores the urgent need for widespread awareness and proactive measures to mitigate its effects.

Diabetes, characterized by elevated levels of blood sugar resulting from either insufficient insulin production or ineffective insulin utilization, poses a significant challenge to global health. Its impact reverberates through communities, affecting individuals of all ages and backgrounds. Symptoms of this conditions manifest in various forms, from increased thirst and frequent urination to persistent fatigue, often disrupting daily life and potentially leading to severe complications if left unmanaged. Despite the absence of a definitive cure, the management of diabetes encompasses an array of treatment options aimed at mitigating symptoms and regulating blood sugar levels, offering hope and improved quality of life for those affected by this pervasive condition.

Moreover, the impact of diabetes extends far beyond individual health, permeating societal and economic spheres. The condition contributes substantially to healthcare expenditures, with costs associated with diabetes management, complications, and related conditions skyrocketing. Furthermore, diabetes poses formidable challenges in terms of productivity losses due to disability, premature mortality, and diminished quality of life for affected individuals and their families. As such, addressing diabetes comprehensively is not merely a matter of health but also an imperative for sustainable healthcare systems and economic prosperity. In this context, efforts to combat diabetes must be multifaceted and proactive, spanning prevention, early detection, and effective management strategies

The purpose of analyzing this dataset is to develop predictive models for identifying individuals at risk of diabetes based on various health-related attributes. Utilizing machine learning algorithms and statistical techniques, we aim to create robust models that can accurately classify individuals into those with or without diabetes. This will facilitate early diagnosis and personalized treatment strategies, ultimately contributing to improved healthcare outcomes for individuals at risk of diabetes. This dataset, sourced from the National Institute of Diabetes and Digestive and Kidney Diseases, serves a critical objective to predicting the likelihood of diabetes. All variables are numerical and are presented below in detail.

5.2 Presentation of the Dataset and Descriptives Statistics

Before starting the presentation of the dataset (Figure 5.1) which consists of 2.728 records, it's essential to understand the key attributes and their significance. Each entry in the dataset is uniquely identified by an 'Id'. The features include 'Pregnancies', indicating the number of times pregnant, 'Glucose', representing plasma glucose concentration over a 2-hour oral glucose tolerance test, 'BloodPressure', which denotes diastolic blood pressure in mm Hg, 'Skin Thickness', measuring triceps skinfold thickness in mm, 'Insulin', indicating 2-Hour serum insulin in mu U/ml, 'BMI' (Body Mass Index), 'Diabetes Pedigree Function', a genetic score of diabetes, and 'Age' in years. Finally, the 'Outcome' column provides a binary classification indicating the presence (1) or absence (0) of diabetes, serving as the target variable for predictive modeling. Understanding these attributes lays the foundation for meaningful analysis and interpretation of the dataset.

Variable Name	Description	Role
Id	Unique identifier for each data entry.	Feature
Pregnancies	Number of times pregnant.	Feature
Glucose	Plasma glucose concentration over 2 hours in an oral glucose tolerance test.	Feature
BloodPressure	Diastolic blood pressure (mm Hg).	Feature
Skin Thickness	Triceps skinfold thickness (mm).	Feature
Insulin	2-Hour serum insulin (mu U/ml).	Feature
BMI	Body mass index (weight in kg / height in m^2).	Feature
Diabetes Pedigree Function	Diabetes pedigree function, a genetic score of diabetes.	Feature
Age	Age in years.	Feature
Outcome	Binary classification indicating the presence (1) or absence (0) of diabetes.	Target

FIGURE 5.1: Structure of our Dataset

5.2.1 The Feature Variables

In the following figures, the density plots of all variables are presented (Figure 5.2) along with their respective box plots (Figure 5.3), providing this way a comprehensive insight into their distributions and variability.

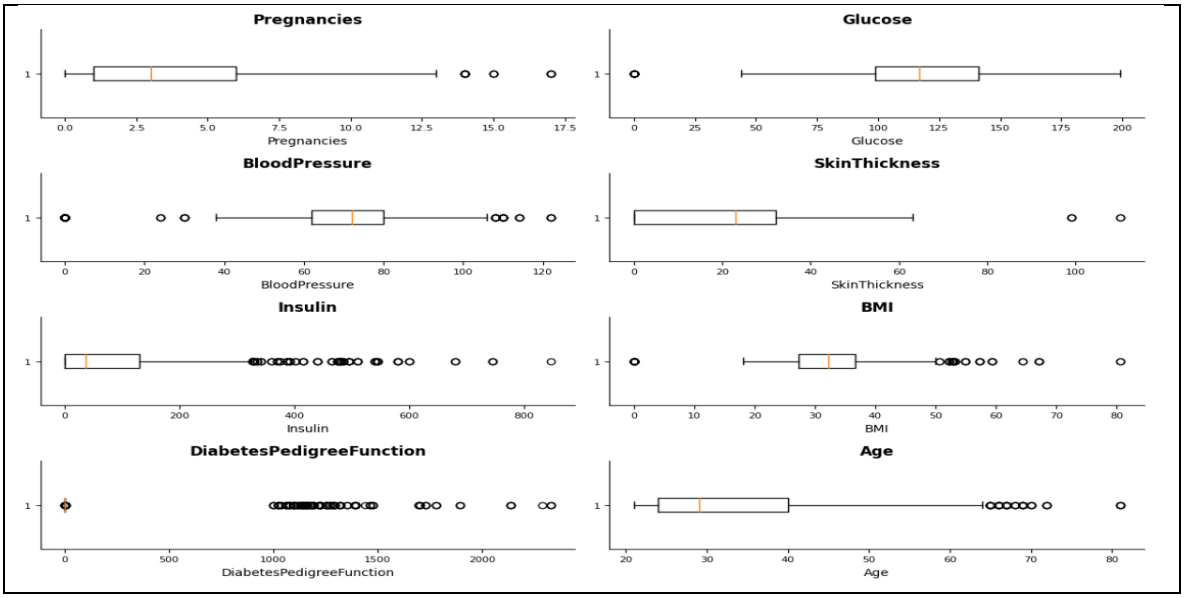

FIGURE 5.2: Boxplots of the Variables

It's evident from our data that most of our participants exhibit blood pressure readings within the range of 60 to 80. BMI values are predominantly clustered around 30, indicating a significant proportion of individuals with this body mass index. Regarding the age distribution of our participants, the majority fall within the range of 25 to 40 years old, with the mean age being 29. These insights offer valuable glimpses into the characteristics of our dataset, paving the way for more nuanced analysis and interpretation. A more detailed picture of what we examine is presented with the density plots below for every variable.

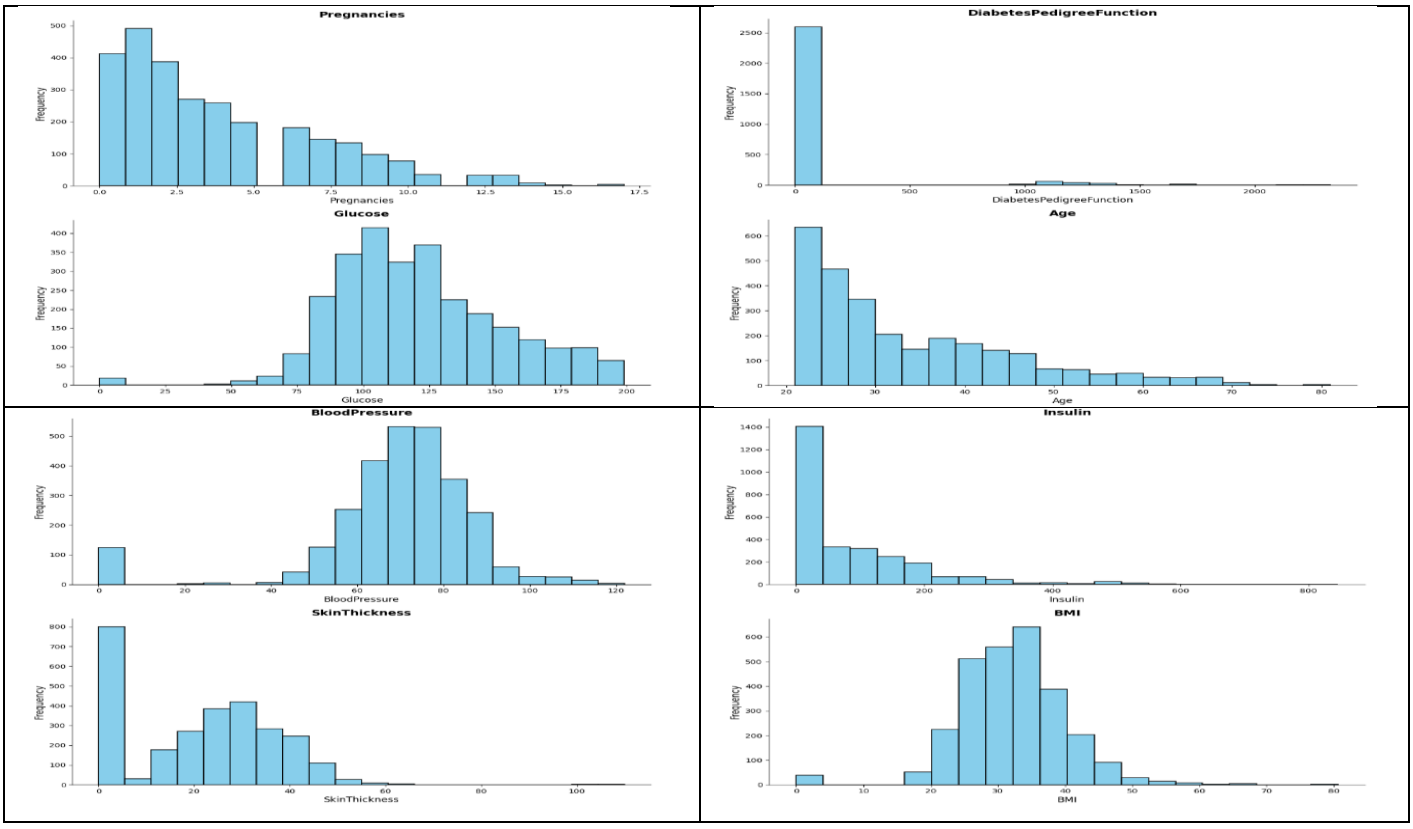

FIGURE 5.3: Density plots of the Variables

Upon inspecting the dataset for missing values, we found none (Figure 5.4). Following this, we employed a technique known as Synthetic Minority Over-sampling Technique (SMOTE) due to the observed class imbalance in the target variable. Subsequently, we conducted normalization, an essential preprocessing step for machine learning algorithms that we will conduct later.

```
#no missing values
print(df.isnull().mean())
```

```
Id                          0.0
Pregnancies                 0.0
Glucose                     0.0
BloodPressure               0.0
SkinThickness               0.0
Insulin                     0.0
BMI                         0.0
DiabetesPedigreeFunction    0.0
Age                         0.0
Outcome                     0.0
dtype: float64
```

FIGURE 5.4: Missing Values

5.2.2 The Target Variable

Continuing our presentation, with the target variable 'Outcome', which serves as a binary classification indicating the presence (1) or absence (0) of diabetes, as described in the dataset. We observe that we have 1816 instances labeled as (0) and 952 instances labeled as (1). In graphical representation (Figure 5.5), this corresponds to a pie chart where we depict proportions. Specifically, we see that 65.6% of the dataset corresponds to the absence of diabetes, while 34.4% corresponds to the presence of diabetes. This distribution provides insight into the prevalence of the outcome categories within our dataset, highlighting the imbalance between the two classes.

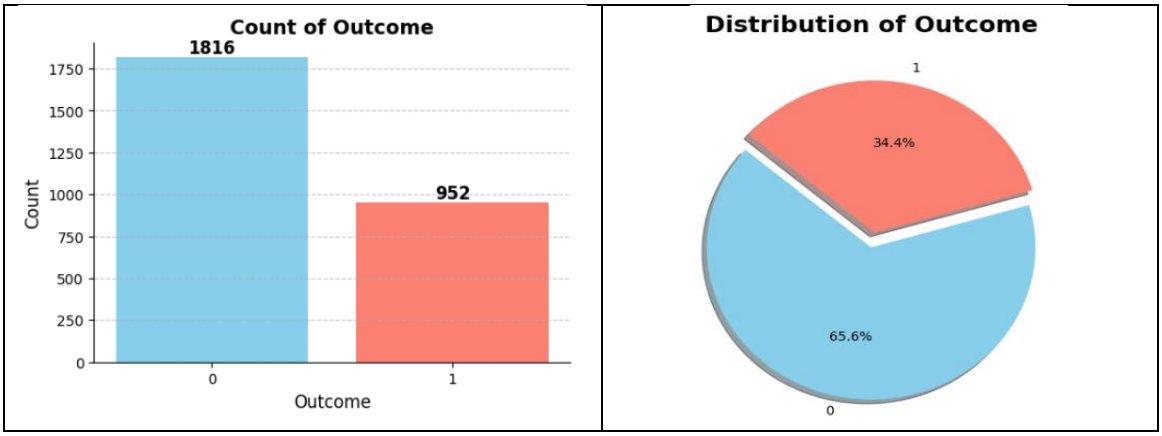

FIGURE 5.5: Describe of the Target Variable: Outcome

Next, we will present (Figure 5.6) the target variable alongside with the density plots of the variables according to the values (0 or 1 of the Outcome) to better understand the structure of our dataset.

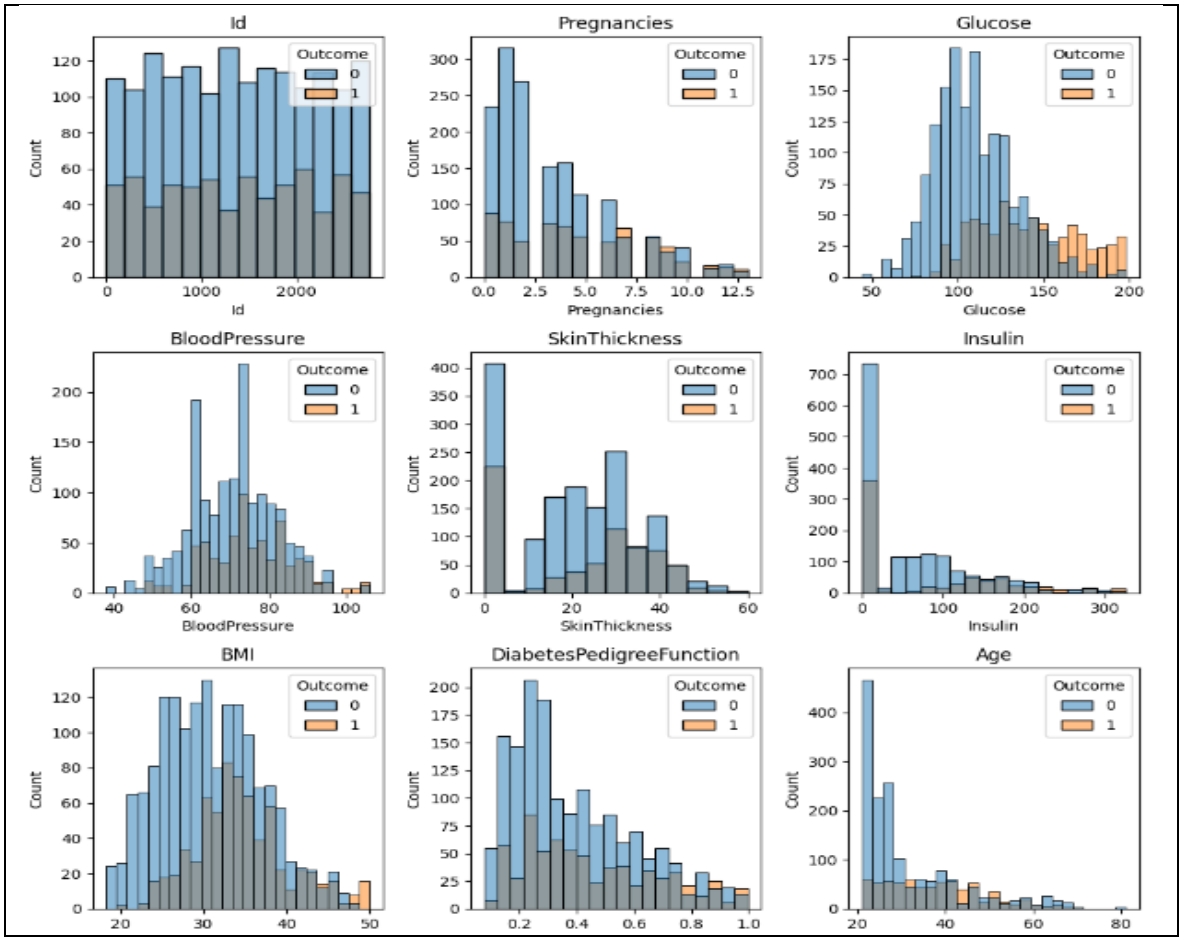

FIGURE 5.6: Density plots of each Variable according to the Target Variable

5.2.3 Correlation-Matrix

The correlation matrix (Figure 5.7) reveals the relationships between different variables within our dataset. Each value represents the correlation coefficient between the corresponding variables. Positive values indicate a positive correlation, while negative values indicate a negative correlation. The closer the value is to 1 or -1, the stronger the correlation, whereas values close to 0 indicate weak or no correlation.

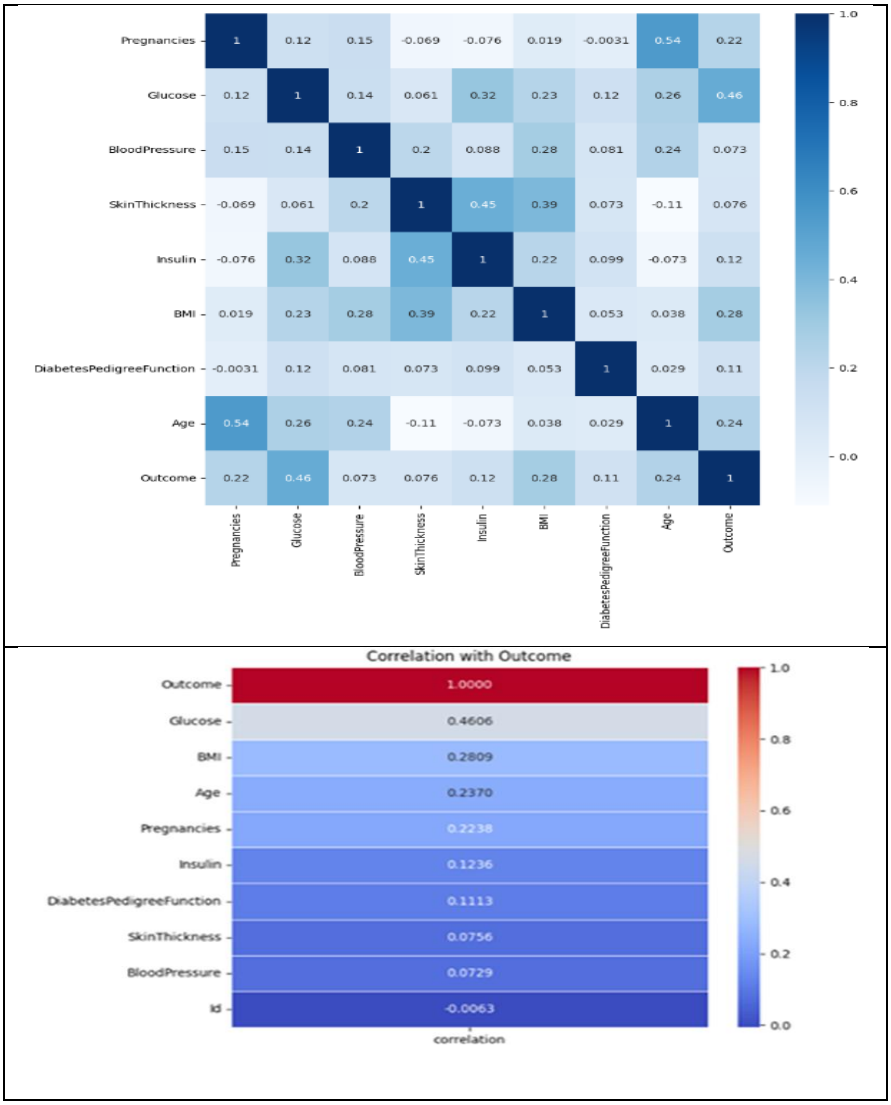

FIGURE 5.7: Correlation Matrix

Specifically, we observe a moderate positive correlation of 0.46 between "Glucose" levels and the "Outcome" variable. This suggests that higher levels of glucose in the blood may be associated with a higher likelihood of the Outcome. Additionally, we note a correlation coefficient of 0.28 between BMI and the Outcome, indicating a moderate positive relationship. However, the correlations of the other variables with the Outcome are relatively low, ranging from 0.23 to 0.07, with the lowest correlation observed for BloodPressure at 0.07. These insights into the correlation structure provide valuable information for understanding the interplay between different variables.

5.2.4 Scatterplots

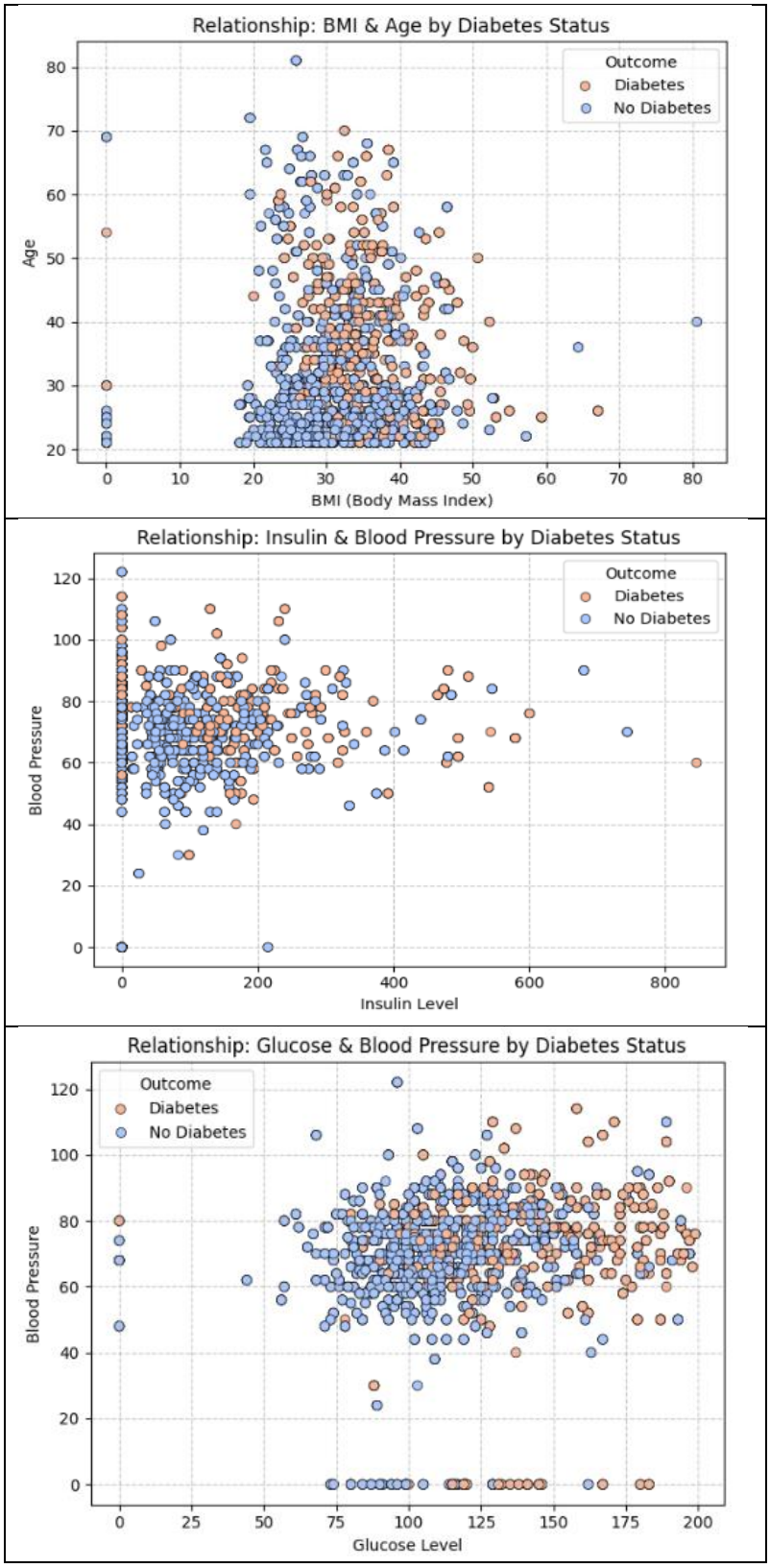

FIGURE 5.8: Scatterplots:
(a) Age – BMI
(b) Blood Pressure – Insulin Level
(c) Blood Pressure – Glucose Level

Next, we examine some of the most important scatterplots (Figure 5.8) that we find in our research. Scatterplots, serve as visual tools for uncovering relationships between variables within our dataset. These plots provide a graphical representation of how two variables are distributed relative to each other, offering valuable insights into potential patterns, trends, and correlations. By plotting each data point on a graph with one variable on the x-axis and the other on the y-axis, we can visually assess the nature of their relationship. These visualizations play a crucial role in identifying any discernible associations or dependencies between variables, thereby enhancing our understanding of the underlying data structure.

In the initial scatterplot featuring Age and BMI variables, we discern a notable trend: participants above the age of 30 tend to exhibit higher incidences of diabetes, particularly those with a BMI exceeding 30. Moving to the second scatterplot, we observe a distinct pattern indicating that elevated levels of Insulin and Blood Pressure are strongly associated with a higher likelihood of diabetes.

Lastly, in the third scatterplot, we have a clear separation between individuals with and without diabetes, arguably the most discernible thus far. Those with diabetes predominantly cluster on the right side, while those without are concentrated on the left. Specifically, for glucose levels, the distinguishing threshold appears to hover around 125, offering a clear demarcation between the two groups.

5.3 Statistics Methods and Techniques

When dealing with imbalanced data, relying solely on accuracy for evaluation can be misleading. Instead, the F1 score proves to be a more reliable metric. Unlike accuracy, the F1 score considers both precision and recall, providing a balanced assessment, especially in situations where class distribution is uneven. In binary classification, results fall into four categories: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Accuracy and F1 metrics take into account these categories to gauge the performance of the model. Below, we provide a more detailed exposition of the types of calculations involved in the metrics:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

and

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Precision and Recall can be found according to,

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Moreover, we use K-fold cross-validation which is a widely used technique for model selection and hyperparameter tuning. In this method, the dataset is divided into k equal-sized parts, or folds. During each iteration, one of these folds is held out as the test set while the remaining k-1 folds are used for training. This process is repeated until each fold has been used as a test set exactly once. The final performance metric is then computed by averaging the results of all k iterations. However, when dealing with imbalanced data, traditional k-fold cross-validation may lead to biased results as it doesn't consider the class distribution. To address this issue, stratified k-fold cross-validation is employed. Like k-fold, it partitions the data into k folds, but it ensures that the ratio of each class within each fold closely matches the original dataset, thus providing a more accurate evaluation of model performance.

In this study, an initial analysis of all variables within the dataset is conducted, followed by a thorough examination of the obtained results. Subsequently, a comparative analysis is performed where extreme values are extracted, and only the interquartile range (25th to 75th percentiles) is retained to assess the performance of our models under altered conditions. Lastly, we engage in a rigorous Principal Component Analysis (PCA), an analytical technique employed to distill the dataset's multitude of variables into a more manageable subset, thereby elucidating the nuanced impact of variable reduction on our analytical framework.

5.3.1 The Statistical Analysis without PCA

After ensuring that the data is properly formatted and arriving at our final dataset, we perform the technique SMOTE which is a statistical technique for increasing the number of cases in our dataset in a balanced way to address the imbalance observed in

the target variable (Figure 5.5). Subsequently, we proceed with normalization and splitting the dataset into training and testing sets, with proportions of 80% and 20% respectively, which is essential for machine learning algorithms.

The supervised classification algorithms that we will examine next are Logistic Regression, Support Vector Machines, k-Nearest Neighbors, Linear Discriminant Analysis, Random Forest, and Extreme Gradient Boosting. Evaluation of the models was performed using metrics such as Accuracy, Precision, Recall, F-score, and AUC. Notably, a 10-fold cross-validation technique was applied, followed by the computation of the mean and standard deviation across the 10 repetitions.

Model	Accuracy(%)	Precision(%)	Recall(%)	F-score(%)	AUC(%)
Logistic Regression	0.752(0.025)	0.759 (0.036)	0.729(0.035)	0.743 (0.026)	0.752 (0.025)
LDA	0.754 (0.025)	0.762 (0.038)	0.728 (0.030)	0.744 (0.024)	0.753 (0.025)
SVM	0.754 (0.026)	0.759 (0.037)	0.735 (0.037)	0.746 (0.026)	0.754 (0.026)
KNN	0.832 (0.022)	0.809 (0.024)	0.862 (0.026)	0.835 (0.021)	0.833 (0.022)
RF	0.864 (0.017)	0.837 (0.025)	0.898 (0.024)	0.866 (0.016)	0.864 (0.017)
XGB	0.911 (0.016)	0.898 (0.014)	0.923 (0.021)	0.910 (0.016)	0.911 (0.016)

TABLE 5.1: Results of Statistical Analysis without PCA

At first, we examine Logistic Regression which proves reliable, offering consistent performance across metrics. Next, we have LDA that shows competitiveness, particularly excelling in precision and F-score. SVM impresses with its commendable results, especially in precision and recall. K-NN stands out for its high accuracy and recall, making it suitable for healthcare classification tasks. Moreover, RF exhibits robust performance, notably excelling in precision and recall. However, Extreme Gradient Boosting (XGB) emerges as the standout performer, consistently surpassing others across all metrics, making it an excellent choice for high-stakes healthcare applications.

Overall, XGB outperforms other models, as we just saw, across all metrics, exhibiting the highest accuracy, precision, recall, F-score, and AUC. More analytical, for each meter we had:

1. **Accuracy**: All models achieve relatively high accuracy, ranging from 75.2% to 91.1%. XGB (Extreme Gradient Boosting) demonstrates the highest accuracy among all models, achieving 91.1%.

2. **Precision**: Precision measures the percentage of true positive predictions out of all positive predictions made by the model. The precision values range from 75.9% to 89.8%, with XGB achieving the highest precision of 89.8%.

3. **Recall**: Recall, also known as sensitivity, measures the percentage of true positive predictions out of all actual positive instances in the dataset. The recall values range from 72.9% to 92.3%, with XGB achieving the highest recall of 92.3%.

4. **F-score**: The F-score is the harmonic mean of precision and recall and provides a balance between the two metrics. F-score values range from 74.3% to 91.0%, with XGB again demonstrating the highest F-score of 91.0%.

5. **AUC**: AUC measures the model's ability to discriminate between positive and negative instances. All models achieve AUC values above 0.75, indicating good discriminative performance. XGB has the highest AUC of 91.1%.

Considering the nature of health-related application, where identifying an individual at risk is crucial, it's worth noting that recall, which indicates the model's ability to correctly identify positive cases, is particularly important. In this regard, XGB's superior performance also in recall suggests its suitability for this diabetes application where accurately identifying individuals with diabetes is paramount.

5.3.2 The Statistical Analysis after the removal of extreme values

After removing extreme values, the dataset initially comprising 2,728 records was reduced to 2,264 records. The 'DiabetesPedigreeFunction' is a function that scores the probability of diabetes based on family history, with a realistic range of 0.08 to 1 most of the times. In our case we come across some values reaching extremes such as 1000 or even 2000. Recognizing these outliers, we opted to exclude values below the 25th percentile and above the 75th percentile to obtain a more accurate representation of the data. Through this adjustment, reflected in the presented boxplot, we ensure that all values of the variable fall within the expected range of 0.08 to 1, providing a clearer and more meaningful depiction of the dataset.

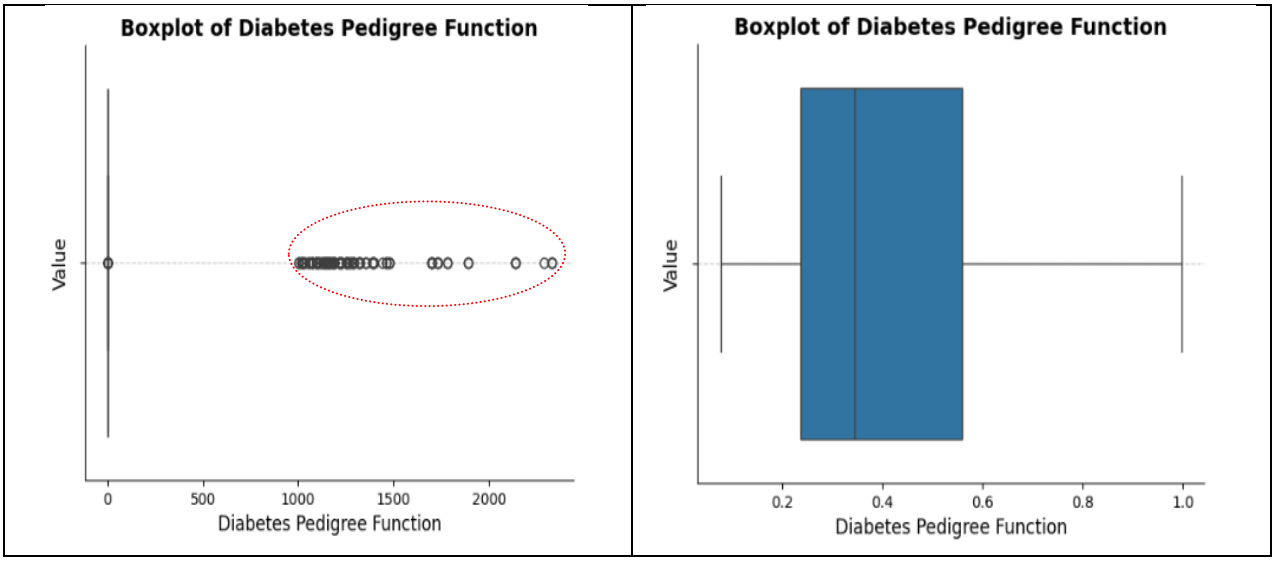

FIGURE 5.9: The change of the variable, Diabetes Pedigree Function after the extreme values has been removed

Subsequently, we repeated the procedure, yielding improved results, as demonstrated below.

Model	Accuracy(%)	Precision(%)	Recall(%)	F-score(%)	AUC(%)
Logistic Regression	0.745 (0.014)	0.758 (0.024)	0.722 (0.031)	0.739 (0.015)	0.745 (0.014)
LDA	0.748 (0.013)	0.766 (0.024)	0.720 (0.038)	0.741 (0.016)	0.748 (0.013)
SVM	0.747 (0.015)	0.769 (0.030)	0.710 (0.032)	0.737 (0.015)	0.747 (0.015)
KNN	0.892 (0.022)	0.850 (0.027)	0.953 (0.024)	0.898 (0.020)	0.891 (0.022)
RF	0.974 (0.010)	0.965 (0.020)	0.984 (0.012)	0.974 (0.010)	0.974 (0.010)
XGB	0.987 (0.007)	0.985 (0.015)	0.988 (0.010)	0.987 (0.006)	0.987 (0.007)

TABLE 5.2: Results of Statistical Analysis after the removal of extreme values

Logistic Regression, LDA, and SVM models experienced slight decreases in accuracy compared to their pre-processed counterparts, although these changes were within a small margin. Also, the KNN model showcased a significant enhancement in accuracy, precision, recall, F-score, and AUC, indicating a substantial improvement in its predictive capabilities.

Remarkably, RF and XGB models exhibited remarkable improvements in all performance metrics, with both achieving high accuracy, precision, recall, F-score, and AUC values. These outcomes underscore the effectiveness of removing extreme values in enhancing the predictive performance of the models, particularly evident in the notable enhancements observed in the RF and XGB models.

5.3.3 The Statistical Analysis with PCA

In our next steps, we'll follow a similar approach to our previous analysis, but with a notable difference: this time, we'll meticulously select features to account for 95% of the variance using Principal Component Analysis PCA. By opting for PCA, we aim to streamline the dataset's dimensionality, particularly crucial as we're dealing with a sizable dataset of 2800 patients and all the features. Additionally, PCA offers the potential to transform correlated variables into a set of linearly uncorrelated ones, enhancing the interpretability of our data. Moreover, given that we haven't yet verified whether each algorithm's assumptions are met (particularly regarding linear independence for linear discriminant analysis) incorporating PCA might further refine our analysis results.

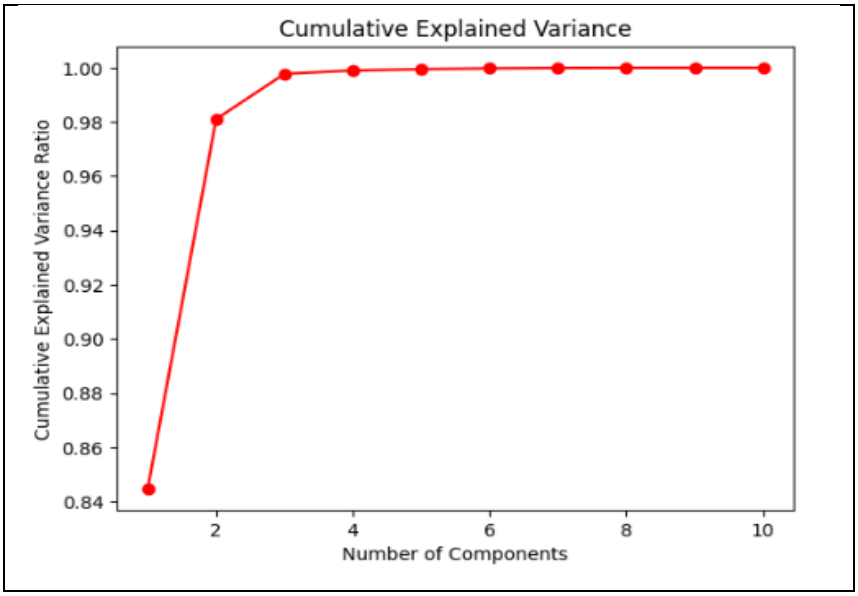

FIGURE 5.10: Plot for the PCA

The PCA plot we see above offers a concise representation of the multidimensional nature of our dataset. Through dimensionality reduction, it projects the original variables onto a lower-dimensional space while preserving the maximum variance within the data. Notably, the first principal component explains approximately 84% of the variance, indicating a strong explanatory power in capturing the underlying structure of the dataset. As we progress along subsequent principal components, the explained variance increases significantly, with the second component reaching 98% and the third component attaining 100%.

This suggests a gradual refinement in capturing the remaining variance within the data, ultimately resulting in a comprehensive representation of the dataset's variability. Due to the discernible structure revealed by the PCA analysis, we opt to select the top two principal components for further analysis, as they offer a balanced trade-off between dimensionality reduction and explanatory power.

Furthermore, employing fewer variables not only reduces the resource-intensive nature of data preparation but also cuts down on costs and time commitments. Therefore, achieving comparable scores with only 2 variables compared to the initial would signify a significant success in our analysis journey.

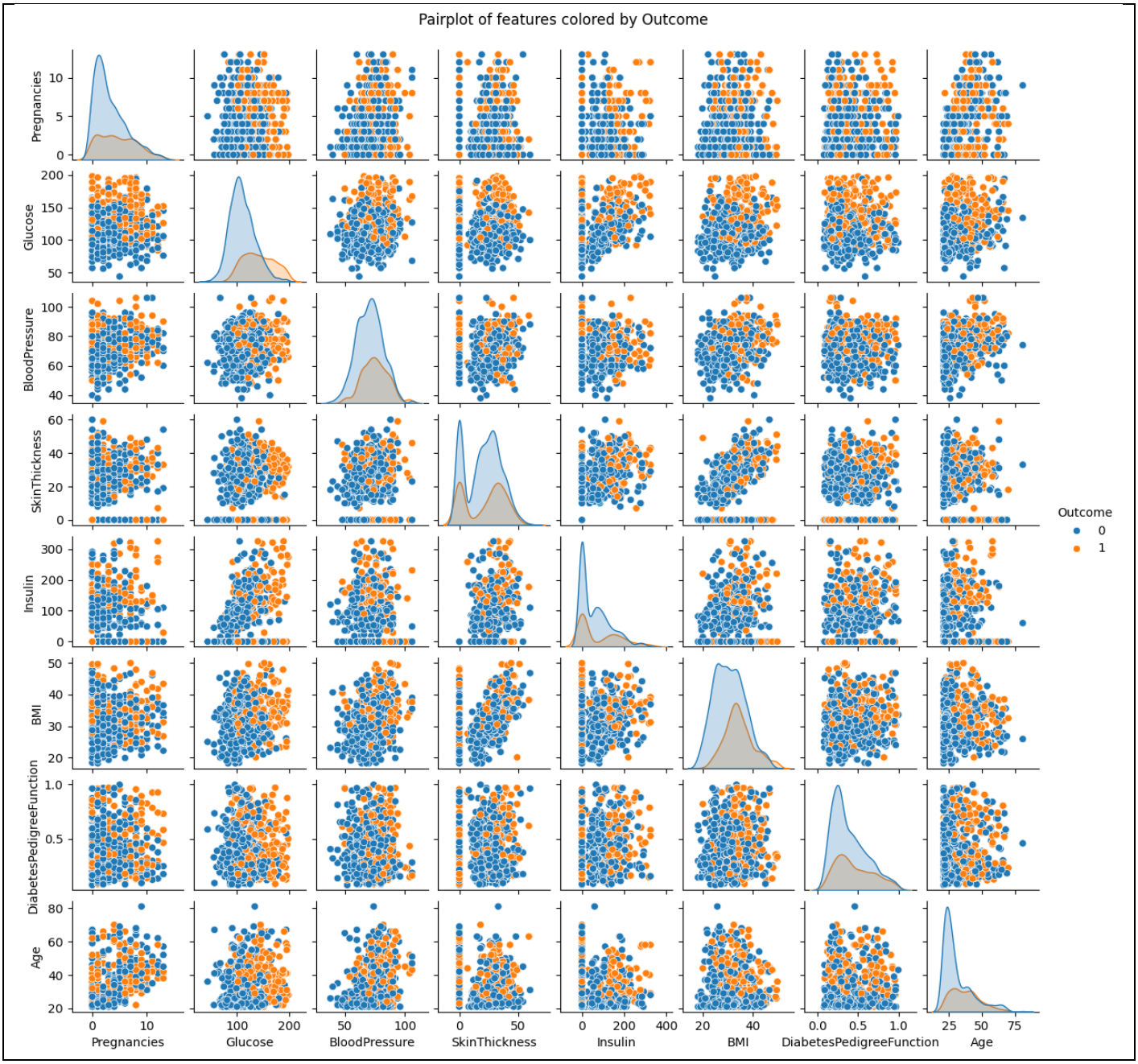

FIGURE 5.11: Pairplot Analysis Colored by Outcome

The pairplot above presents an extensive view of the interrelationships among all the features in our dataset. Each scatter plot illustrates the correlation between two variables, while the diagonal plots depict the distributions of individual features. The incorporation of color coding based on the 'Outcome' variable enables us to discern the distribution of classes across the feature space. This visualization approach facilitates the detection of potential patterns or correlations between features and the target variable, thereby enriching our analytical understanding.

Up next, we presented the results of statistical analysis for each model separately, this time with the PCA. Despite the reduction to only two variables, the models perform remarkably similarly to those analyzed in our initial assessment without PCA. This observation underscores the robustness of the models and suggests that the essential information captured by the original feature set is effectively retained in the reduced

dimensional space. By leveraging the insights gained from PCA, we are able to streamline our analysis without sacrificing predictive performance, thereby enhancing the efficiency and interpretability of our modeling approach.

Model	Accuracy(%)	Precision(%)	Recall(%)	F-score(%)	AUC(%)
Logistic Regression	0.700 (0.029)	0.697 (0.028)	0.694 (0.039)	0.695 (0.031)	0.700 (0.029)
LDA	0.698 (0.027)	0.694 (0.027)	0.695 (0.039)	0.694 (0.030)	0.698 (0.028)
SVM	0.698 (0.028)	0.694 (0.028)	0.693 (0.031)	0.693 (0.031)	0.698 (0.028)
KNN	0.846 (0.016)	0.835 (0.021)	0.857 (0.033)	0.846 (0.018)	0.846 (0.016)
RF	0.923 (0.014)	0.914 (0.023)	0.932 (0.018)	0.923 (0.013)	0.923 (0.014)
XGB	0.805 (0.029)	0.800 (0.037)	0.809 (0.033)	0.804 (0.027)	0.805 (0.029)

TABLE 5.3: Results of Statistical Analysis with PCA

We also add below two graphs to demonstrate how the models are learning and how the Cross-Validation procedure works with CV=10.

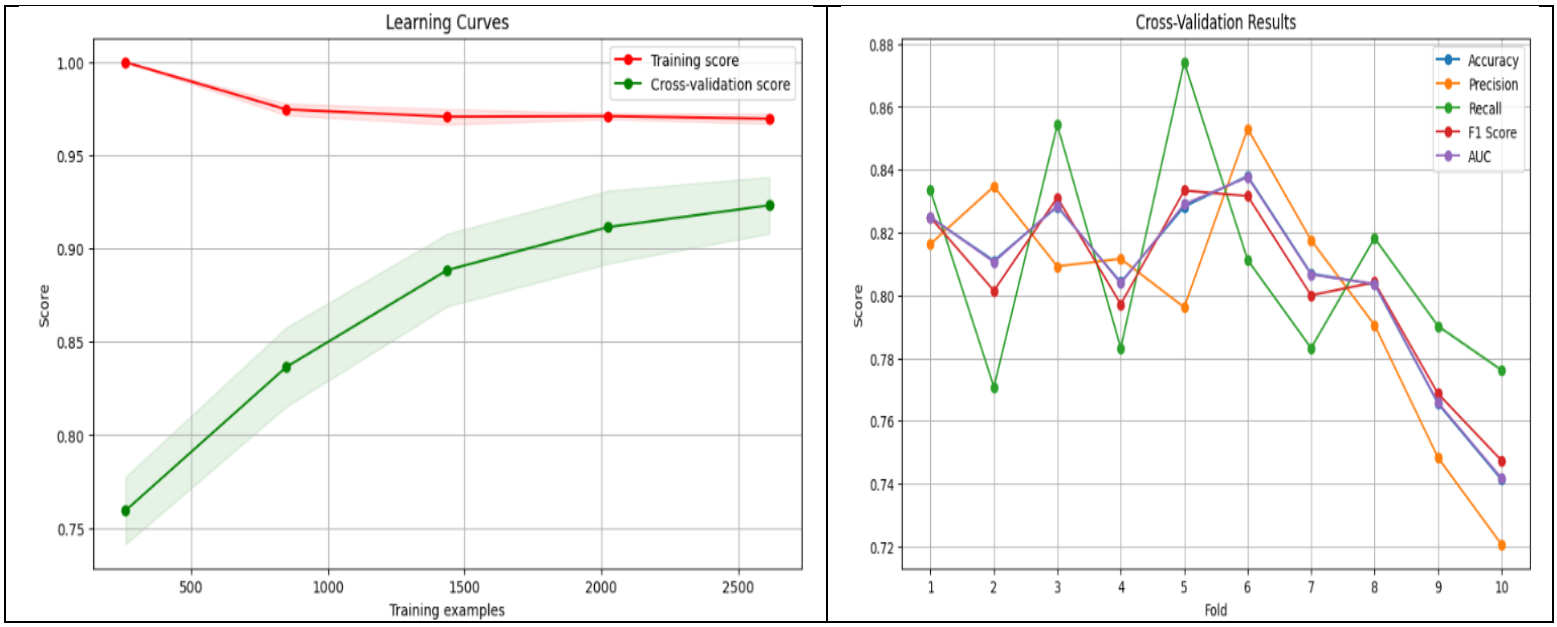

FIGURE 5.12: Plots of how the models work a) Learning Curves b) Cross-Validation results for cv-score = 10

To be precise, we further enhance our understanding of the models' performance by incorporating two insightful graphical representations. First, the learning curves provide a visual depiction of how the models are learning over time, offering valuable insights into their convergence behavior and potential for overfitting or underfitting. Secondly, the cross-validation results, conducted with a 10-fold cross-validation procedure, offer a comprehensive evaluation of model performance across multiple subsets of the data. These plots serve as indispensable tools for assessing the robustness and generalization capabilities, providing another view of their performance under varying conditions. By integrating these graphical analyses into our analysis, we gain

deeper insights about how the CV works and the learning process that has been applied in general.

Generally, we observe that classification models without PCA in this classification problem performed better, although there are no significant differences between the metrics. After applying PCA, we observe a decrease in accuracy for each model examined except, RF that it showed an increase. However, the accuracy remains marginally in good levels, so the predictions can still be considered acceptable. In fact, the best-performing model appears to be the Random Forest with an accuracy of 0.923 this time. On the other hand, the model with the biggest accuracy reduction is the Extreme Gradient Boosting with accuracy from 0.911without PCA went to 0.805 with PCA.

5.4 Artificial Neural Network

The dataset has been partitioned into training and testing subsets, with proportions of 80% and 20% respectively. The neural network undergoes two iterations: one employing the conventional approach of dividing the data into training and testing sets, and another utilizing the 10-fold cross-validation technique.

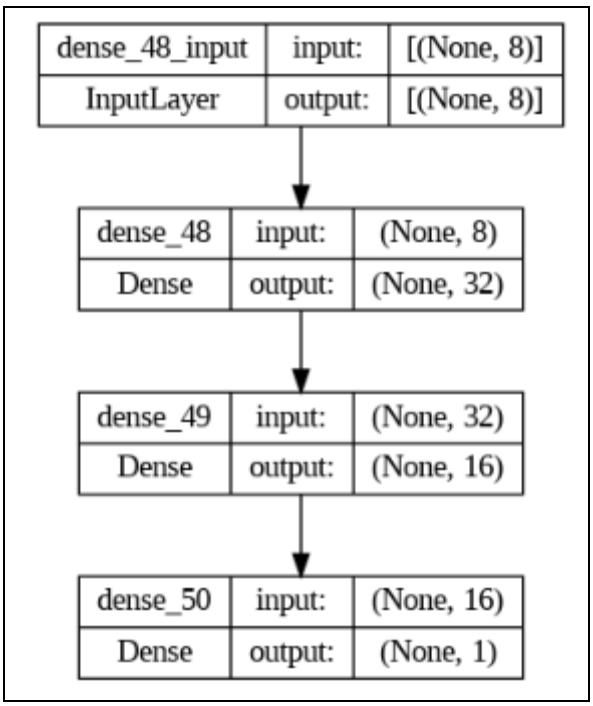

FIGURE 5.13: Architecture of the ANN

The architectural configuration of this neural network model is as follows: an input layer comprising 8 neurons, followed by the first hidden layer with 32 neurons, the second hidden layer with 16 neurons, and finally, the output layer with a single neuron. The rectified linear unit (ReLU) activation function is applied in both hidden layers to introduce non-linearity, enabling the model to capture intricate patterns within the data. The sigmoid activation function is employed in the output layer, facilitating the

mapping of the network's output to a probability value between 0 and 1, given that our neural network is designed to address a binary classification task.

The results, displayed in the table below, demonstrate considerable efficacy. Furthermore, the comprehensive performance evaluation is illustrated in the graph presented subsequently.

Accuracy	**0.910**
Precision	**0.874**
Recall	**0.922**
F1-score	**0.914**
AUC-ROC	**0.910**

TABLE 5.4: Results of ANN

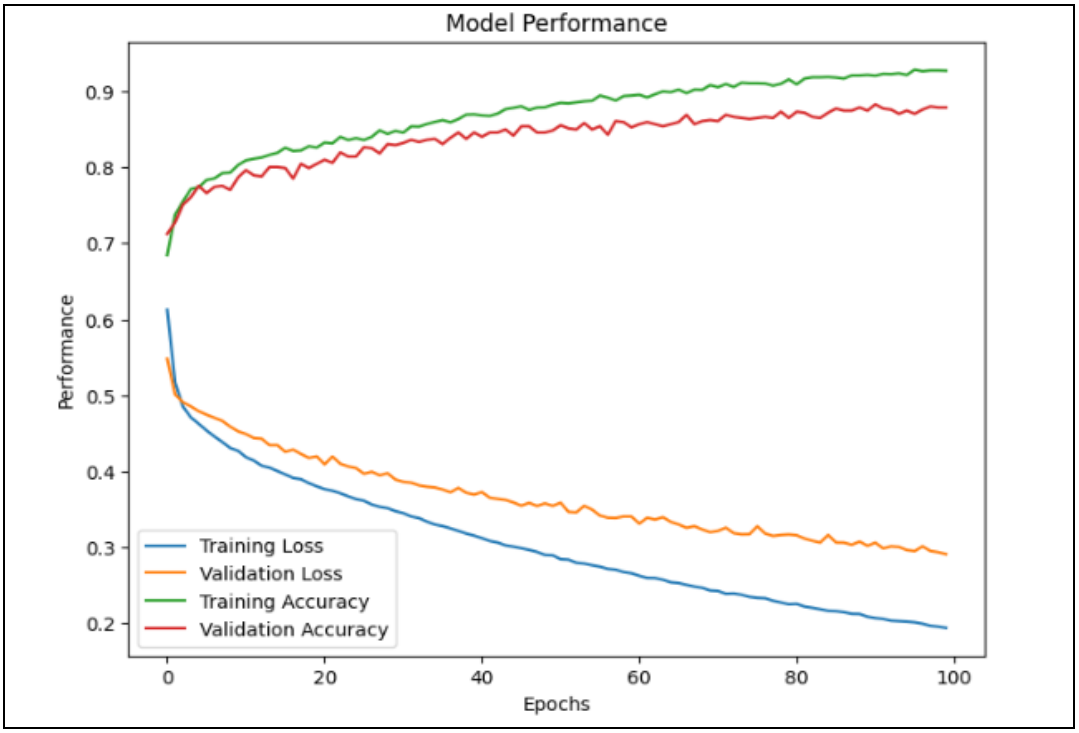

FIGURE 5.14: Performance of ANN and results

The metrics obtained using the 10-fold cross-validation technique is presented in the final table. It includes accuracy and loss metrics for the neural network model.

The cross-validation accuracy of 99.62% indicates a high level of model performance across multiple validation folds. This highlights, the neural network's ability to generalize to unseen data. Additionally, the cross-validation loss of 0.02 further corroborates the effectiveness of the model, with minimal error incurred during training and validation.

Cross-validation accuracy	**99.62% (+/- 0.00%)**
Cross-validation loss	**0.02 (+/- 0.00)**

TABLE 5.5: Comparison of the results of ANNs

5.5 Conclusion of the Analysis

To sum it up, we present a comprehensive overview of the statistical analysis conducted throughout our study and summarize the key findings. Our analysis encompasses various aspects, including descriptive statistics, model evaluation metrics, and validation techniques.

The initial exploration of the dataset revealed important insights into the distribution and variability of the feature variables. Through visualizations such as boxplots, density plots, and scatterplots, we gained valuable understanding of the data characteristics and how they interact with each other. Preprocessing steps such as handling missing values, normalization, and outlier removal were essential for ensuring data quality and enhancing model performance. Techniques like SMOTE addressed class imbalance, while dimensionality reduction via PCA streamlined the dataset for modeling.

The evaluation of predictive models was conducted using a range of metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve AUC-ROC. These metrics provided comprehensive look into model performance across different aspects of classification tasks, guiding model selection and examining it from different views.

Overall, our analysis highlights the effectiveness of machine learning algorithms in predicting diabetes risk based on health-related attributes. The comparison of models highlights the superior performance of XGB and RF algorithms, particularly after preprocessing steps like outlier removal and dimensionality reduction.

In addition to our analysis of traditional machine learning algorithms, we also explored the efficacy of ANN in predicting diabetes risk. The ANN, with its intricate architecture and ability to capture complex patterns within data, yielded promising results in our study. Through two iterations—one employing conventional training and testing sets, and another utilizing a 10-fold cross-validation technique—we observed consistently high levels of accuracy, precision, recall, and F1-score. The architectural configuration of the ANN, comprising multiple layers with varying numbers of neurons and activation functions, enabled the model to effectively learn and generalize from the data. Notably, the ANN exhibited remarkable performance in both training and validation phases, with cross-validation accuracy reaching an impressive 99.62% and minimal loss incurred during training. These results underscore the potential of neural network models in healthcare analytics, offering a complementary approach to traditional machine learning algorithms and further enriching our understanding of diabetes risk prediction.

Through the above statistical analysis and evaluation, we have gained valuable insights into the predictive modeling of diabetes risk. Our findings underscore the importance of RWD and how to process them to make decisions that previously would take more time and money to make. Also, crucial role played each time, the model selection in developing accurate and reliable models.

As our understanding of diabetes continues to evolve, so too does our approach to its management and treatment. The complex interplay of genetic predisposition, lifestyle factors, and environmental influences underscores the multifaceted nature of diabetes. Research endeavors tirelessly seek to unravel its mysteries, for innovative

therapies and preventative strategies to confront this global health challenge head-on. Amidst the ongoing quest for breakthroughs, the imperative remains to empower individuals with diabetes with knowledge, support, and access to comprehensive care. By fostering awareness, advocating for equitable healthcare policies, and fostering collaborative efforts across disciplines, we can strive towards a future where the burden of diabetes is alleviated, and the promise of optimal health and well-being is within reach for all.

6 CHAPTER

6.1 Discussion

Overall, this study stands as a thorough examination of the vast potentials residing within RWD and its profound impact on healthcare. While embracing the opportunities it presents, the study also acknowledges the complexities and challenges that must be addressed with diligence and strategic foresight. Through our investigation, we've discerned that RWD holds immense potential in reshaping healthcare practices, but its integration requires meticulous navigation and strategic planning. One of the fundamental distinctions we've established is between RWD and RWE. While RWD serves as the raw material, RWE emerges from the analysis and interpretation of this data, offering valuable insights into real-world patient outcomes, treatment patterns, and the overall effectiveness and safety of medical interventions. Regulatory bodies such as the EMA and the FDA play pivotal roles in shaping the landscape of healthcare approvals. Understanding their functions and regulations is essential for harnessing the potential of RWD in informing decision-making processes.

Moreover, we've identified various strategies for confronting the inherent challenges associated with RWD, from harmonizing health data through initiatives like the OMOP to embracing HTA as a tool for evaluating the value of healthcare interventions. Also, as it comes to the next steps in drug development, RWD offers a paradigm shift through innovative approaches like PK–PD–PE modeling, enabling early prediction and differentiation in clinical trials. Additionally, the integration of Social Listening and other innovative strategies facilitates a more personalized approach to medicine, bridging the gap between biology and pharmacology.

Looking towards the future, the leveraging of RWD holds promise in maximizing clinical development results and driving advancements in pharmaceutical research and development. From improving healthcare systems to elevating specific fields like arthroplasty and acoustics, the potential of RWD is vast and far-reaching. Furthermore, the synergy between RWD and machine learning presents exciting opportunities for enhancing predictive biomarkers, designing clinical trials, and optimizing pharmaceutical research. By embracing machine learning methodologies and leveraging the power of advanced analytics, we can unlock new insights and accelerate the pace of discovery in healthcare, ushering in a new era of data-driven decision-making and personalized medicine.

As established in our Literature Review, the diverse applications and transformative potential of RWD in the health sector are both profound and far-reaching. Our review underscored how RWD, encompassing everything from genetic sequences and protein structures to clinical records and epidemiological datasets, functions as a continuously evolving record of biological processes. This extensive scope of data enables a more comprehensive understanding of the intricate relationships between genetics, diseases, and therapeutic interventions. By integrating insights from various studies, we observed that RWD significantly enhances pharmaceutical research and development through the application of AI, DL, and ML. These technologies leverage vast datasets to uncover hidden patterns, optimize treatment protocols, and predict disease outcomes with remarkable accuracy. Moreover, the real-world applicability of these data-driven approaches in diverse medical fields, such as orthopedic surgery and stroke treatment,

demonstrates their critical role in improving patient outcomes and advancing personalized medicine. Ultimately, the Literature Review elucidated how RWD not only drives innovation in bioinformatics but also empowers healthcare professionals to make more informed decisions, thereby fostering a more effective and responsive health system.

In summary, our exploration of RWD and its profound impact on healthcare signifies not just a shift in methodology, but a revolution in the very fabric of medical practice. As we stand at the nexus of data-driven decision-making and precision medicine, the possibilities are limitless. By embracing innovation, ethical principles, and collaborative efforts, we have the opportunity to harness the full potential of RWD to revolutionize patient care, drive advancements in medical research, and ultimately, improve the lives of individuals worldwide.

This journey is not without its challenges, but with perseverance and dedication, we can overcome obstacles and chart a course towards a future where healthcare is not just reactive, but proactive and personalized. As we continue to refine our understanding and utilization of RWD, let us remain steadfast in our commitment to ethical stewardship, ensuring that patient privacy and welfare are paramount in all endeavors. The RWD is not a vision for the future of healthcare it has already arrive and is here to stay, evolve and expand before our eyes. This ever-changing landscape, marked by rapid advancements in technology and data analytics, must keep us vigilant and attuned to the ongoing developments reshaping the healthcare sector. From the innovative applications of machine learning to the intricacies of data-driven decision-making, every stride forward propels us closer to a future where healthcare is not just reactive but anticipatory and personalized.

Code in Python

```
#Import Data
df = pd.read_excel('Healthcare-Diabetes(new).xlsx')
df=pd.DataFrame(df)
# print the shape of data
df.shape
# print the columns of the data
df.columns.to_list()
# describe the columns in the data set
df.describe()
variables = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome']
num_variables = len(variables)

#Scatterplot 1
sns.scatterplot(data=df, x='Glucose', y='BloodPressure', hue='Outcome',
palette='coolwarm', marker='o', edgecolor='k')
plt.title('Relationship: Glucose & Blood Pressure by Diabetes Status')
plt.xlabel('Glucose Level')
plt.ylabel('Blood Pressure')
plt.legend(title='Outcome', loc='best', labels=['Diabetes', 'No Diabetes'])
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
#Scatterplot 2
sns.scatterplot(data=df, x='Insulin', y='BloodPressure', hue='Outcome',
palette='coolwarm', marker='o', edgecolor='k')
plt.title('Relationship: Insulin & Blood Pressure by Diabetes Status')
plt.xlabel('Insulin Level')
plt.ylabel('Blood Pressure')
plt.legend(title='Outcome', loc='best', labels=['Diabetes', 'No Diabetes'])
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
#Scatterplot 3
sns.scatterplot(data=df, x='BMI', y='Age', hue='Outcome', palette='coolwarm',
marker='o', edgecolor='k')
plt.title('Relationship: BMI & Age by Diabetes Status')
plt.xlabel('BMI (Body Mass Index)')
plt.ylabel('Age')
plt.legend(title='Outcome', loc='best', labels=['Diabetes', 'No Diabetes'])
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()  # Adjust layout to prevent overlap
plt.show()
```

```
# Calculate the number of rows and columns needed
num_rows = (num_variables + 1) // 2
num_cols = 2

# Create subplots
fig, axs = plt.subplots(num_rows, num_cols, figsize=(15, 20))
    # Flatten the axs array
axs = axs.ravel()

# Histograms for each variable
for i, var in enumerate(variables):
axs[i].hist(df[var], bins=20, color='skyblue', edgecolor='black', linewidth=1.2)
axs[i].set_title(var, fontsize=14, fontweight='bold')
axs[i].set_xlabel(var, fontsize=12)
axs[i].set_ylabel('Frequency', fontsize=12)
axs[i].tick_params(axis='both', which='major', labelsize=10)
axs[i].spines[['top', 'right']].set_visible(False)
# Hide any empty subplots
for j in range(num_variables, len(axs)):
axs[j].axis('off')
plt.tight_layout()
plt.show()
# Define the number of variables
num_variables = len(variables)
# Calculate the number of rows needed
num_rows = (num_variables + 1) // 2
# Create subplots with smaller size and nicer layout
fig, axs = plt.subplots(num_rows, 2, figsize=(12, 10))
# Flatten the axs array
axs = axs.ravel()

# Boxplots for each variable
for i, var in enumerate(variables):
axs[i].boxplot(df[var], vert=False)
axs[i].set_title(var, fontsize=12, fontweight='bold')
axs[i].set_xlabel(var, fontsize=10)
axs[i].tick_params(axis='both', which='major', labelsize=8)
axs[i].spines[['top', 'right']].set_visible(False)
# Hide any empty subplots
for j in range(num_variables, len(axs)):
axs[j].axis('off')
plt.tight_layout()
plt.show()
outcome_counts = df['Outcome'].value_counts()
plt.figure(figsize=(6, 4))
plt.bar(outcome_counts.index, outcome_counts.values, color=['skyblue', 'salmon'])
plt.title('Count of Outcome', fontsize=14, fontweight='bold')
plt.xlabel('Outcome', fontsize=12)
plt.ylabel('Count', fontsize=12)
plt.xticks(outcome_counts.index, ['0', '1'])
```

```
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.gca().spines[['top', 'right']].set_visible(False)
for i, count in enumerate(outcome_counts.values):
plt.text(i, count, str(count), ha='center', va='bottom', fontsize=12, fontweight='bold')
plt.show()
outcome_counts = df['Outcome'].value_counts()
labels = outcome_counts.index
sizes = outcome_counts.values
colors = ['skyblue', 'salmon']
explode = (0.1, 0)  # explode the first slice

# Plot for Target Variable: Outcome
plt.figure(figsize=(4, 5))
plt.pie(sizes, explode=explode, labels=labels, colors=colors, autopct='%1.1f%%',
startangle=140, shadow=True)
plt.title('Distribution of Outcome', fontsize=18, fontweight='bold')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
# Display
plt.tight_layout()
plt.show()


num_list=['Id','Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
'BMI', 'DiabetesPedigreeFunction', 'Age']
fig = plt.figure(figsize=(15,25))
for i in range(len(num_list)):
plt.subplot(6,3,i+1)
plt.title(num_list[i])
sns.histplot(data=df,x=df[num_list[i]],hue='Outcome')
plt.tight_layout()

# Corellation Matrix
cor = df.drop(["Id"],axis=1).select_dtypes(include='number').copy()
corr = cor.corr()
fig , ax = plt.subplots(figsize=(10 , 10))
sns.heatmap(corr ,annot= True , ax=ax , cmap= 'Blues');
corr = df.corrwith(df['Outcome']).sort_values(ascending=False)
plt.figure(figsize=(8, 6))
sns.heatmap(pd.DataFrame(corr , columns=['correlation']), annot=True,
cmap='coolwarm', fmt=".4f", linewidths=.5)
plt.title('Correlation with Outcome')
plt.show()

X=df.drop(["Id", "Outcome"],axis=1)
y=df["Outcome"]

# Check for missing values
print(df.isnull().mean())
# Choose the best number of components based on the explained variance ratio
cumulative_variance_ratio = np.cumsum(variance_ratio_all)
```

```
n_components = np.argmax(cumulative_variance_ratio >= 0.95) + 1
print(f"Best number of components: {n_components}")
# Perform PCA with the chosen number of components
pca_best = PCA(n_components=n_components).fit(df)
df = pca_best.transform(df)

# Balance dataset
sm = SMOTE(random_state = 1234)
X, y = sm.fit_resample(X, y)
# Print the class distribution of the resampled dataset
print('Resampled class distribution: ', Counter(y))
# Scale the features
scaler = StandardScaler()
X = scaler.fit_transform(X)
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20)

# Fit the logistic regression model on the training set
model = LogisticRegression(penalty="l2", C=1, solver="lbfgs", max_iter=500)
model.fit(X_train, y_train)

# Fit the LDA model on the training set
model = LinearDiscriminantAnalysis()
model.fit(X_train, y_train)

# Fit the SVM model on the training set
model = SVC(kernel='linear', C=1, random_state=1234)
model.fit(X_train, y_train)

# Fit the KNN model on the training set
model = KNeighborsClassifier()
model.fit(X_train, y_train)

# Define XGBoost classifier model
params = {'objective': 'binary:logistic', 'n_estimators': 300, 'learning_rate': 0.05,
'max_depth': 10, 'min_child_weight': 1, 'gamma': 0.1, 'subsample': 0.8,
'colsample_bytree': 0.8, 'scale_pos_weight': 1, 'seed':42}
model = xgb.XGBClassifier(**params)
model.fit(X_train, y_train)

# Define Random Forest classifier model
model = RandomForestClassifier(n_estimators=500, max_depth=10,
random_state=1234)
model.fit(X_train, y_train)

# Models performance
# Predict on the test data
y_pred = model.predict(X_test)
# Print the classification report and AUC score
print(classification_report(y_test, y_pred, digits=4))
print('AUC score:', roc_auc_score(y_test, y_pred))
```

```
scorers = {'accuracy': make_scorer(accuracy_score),
'precision': make_scorer(precision_score, pos_label=1),
'recall': make_scorer(recall_score, pos_label=1),
'f1': make_scorer(f1_score, pos_label=1),
'auc': make_scorer(roc_auc_score)}
cv_results = cross_validate(model,X_train, y_train,cv=10,  scoring=scorers)
print('Accuracy: {:.3f} ({:.3f})'.format(np.mean(cv_results['test_accuracy']),
np.std(cv_results['test_accuracy'])))
print('Precision: {:.3f} ({:.3f})'.format(np.mean(cv_results['test_precision']),
np.std(cv_results['test_precision'])))
print('Recall: {:.3f}
({:.3f})'.format(np.mean(cv_results['test_recall']),np.std(cv_results['test_recall'])))
print('F1 Score: {:.3f} ({:.3f})'.format(np.mean(cv_results['test_f1']),
np.std(cv_results['test_f1'])))
print('AUC: {:.3f} ({:.3f})'.format(np.mean(cv_results['test_auc']),
np.std(cv_results['test_auc'])))
# Assuming y_true and y_pred are the true and predicted labels respectively
cm = confusion_matrix(y_test, y_pred)
print("Confusion matrix:")
print(cm)
print("Total samples:", np.sum(cm))
print("True positives:", cm[1, 1])
print("False positives:", cm[0, 1])
print("True negatives:", cm[0, 0])
print("False negatives:", cm[1, 0])
```

#Handling Extreme Values

```
def outliers(df,ft):
q1 = df[ft].quantile(0.25)
q3 = df[ft].quantile(0.75)
iqr = q3 - q1
lower_limit = q1 - iqr *1.5
upper_limit = q3 + iqr *1.5
ls = df.index[(df[ft]<lower_limit) | (df[ft]>upper_limit)]
return ls
index_list = []
num = ['Id', 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
'BMI', 'DiabetesPedigreeFunction']
for feature in num:
index_list.extend(outliers(df,feature))
def remove(df,ls):
ls = sorted(set(ls))
df = df.drop(ls)
return df
df = remove(df,index_list)
df.shape
sns.boxplot(data=df, x="DiabetesPedigreeFunction")
plt.title('Boxplot of Diabetes Pedigree Function', fontsize=14, fontweight='bold')
plt.xlabel('Diabetes Pedigree Function', fontsize=12)
plt.ylabel('Value', fontsize=12)
```

```
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.gca().spines[['top', 'right']].set_visible(False)
plt.show()
```

#Perform PCA

```
pca = PCA()
pca.fit(df)
plt.figure()
plt.plot(range(1, pca.n_components_ + 1),
np.cumsum(pca.explained_variance_ratio_), marker='o', linestyle='-', color='r')
plt.xlabel('Number of Components')
plt.ylabel('Cumulative Explained Variance Ratio')
plt.title('Cumulative Explained Variance')
plt.show()
#Plot for the PCA
pca = PCA(n_components=2)
```

 # Pairplot for all the Variables

```
sns.pairplot(df.drop(columns=['Id']), hue='Outcome', diag_kind='kde', height=1.5)
plt.suptitle("Pairplot of features colored by Outcome", y=1.02)
plt.show()
```

 # Plotting Learning Curves

```
train_sizes, train_scores, test_scores = learning_curve(model, X_train_pca, y_train,
cv=10, scoring='accuracy')
train_scores_mean = np.mean(train_scores, axis=1)
train_scores_std = np.std(train_scores, axis=1)
test_scores_mean = np.mean(test_scores, axis=1)
test_scores_std = np.std(test_scores, axis=1)
plt.figure(figsize=(10, 6))
plt.fill_between(train_sizes, train_scores_mean - train_scores_std, train_scores_mean
+ train_scores_std, alpha=0.1, color="r")
plt.fill_between(train_sizes, test_scores_mean - test_scores_std, test_scores_mean +
test_scores_std, alpha=0.1, color="g")
plt.plot(train_sizes, train_scores_mean, 'o-', color="r", label="Training score")
plt.plot(train_sizes, test_scores_mean, 'o-', color="g", label="Cross-validation score")
plt.title("Learning Curves")
plt.xlabel("Training examples")
plt.ylabel("Score")
plt.legend(loc="best")
plt.grid(True)
plt.show()
```

 # Plotting cross-validation results

```
plt.figure(figsize=(10, 6))
plt.plot(range(1, cv_results['test_accuracy'].shape[0] + 1), cv_results['test_accuracy'],
marker='o', label='Accuracy')
plt.plot(range(1, cv_results['test_precision'].shape[0] + 1), cv_results['test_precision'],
marker='o', label='Precision')
```

```
plt.plot(range(1, cv_results['test_recall'].shape[0] + 1), cv_results['test_recall'],
marker='o', label='Recall')
plt.plot(range(1, cv_results['test_f1'].shape[0] + 1), cv_results['test_f1'], marker='o',
label='F1 Score')
plt.plot(range(1, cv_results['test_auc'].shape[0] + 1), cv_results['test_auc'], marker='o',
label='AUC')
plt.title('Cross-Validation Results')
plt.xlabel('Fold')
plt.ylabel('Score')
plt.xticks(range(1, cv_results['test_accuracy'].shape[0] + 1))
plt.legend()
plt.grid(True)
plt.show()
```

 # Code for the ANN

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20)
# Define the model
model = Sequential()
model.add(Dense(32, input_dim=8, activation='relu'))
model.add(Dense(16, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
# Compile the model
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
# Train the model
history = model.fit(X_train, y_train, epochs=100, batch_size=32,
verbose=1,validation_data = (X_test, y_test))
# Evaluate the model on test data
y_pred = model.predict(X_test)
y_pred = (y_pred > 0.5).astype(int)
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
auc_roc = roc_auc_score(y_test, y_pred)
print('Test accuracy:', accuracy)
print('Test precision:', precision)
print('Test recall:', recall)
print('Test F1-score:', f1)
print('Test AUC-ROC:', auc_roc)
plot_model(model, show_shapes=True)
```

 # Plot loss and accuracy over epochs

```
plt.figure(figsize=(8, 6))
plt.plot(history.history['loss'], label='Training Loss')
plt.plot(history.history['val_loss'], label='Validation Loss')
plt.plot(history.history['accuracy'], label='Training Accuracy')
plt.plot(history.history['val_accuracy'], label='Validation Accuracy')
plt.title('Model Performance')
plt.xlabel('Epochs')
```

```
plt.ylabel('Performance')
plt.legend()
plt.show()

# 10-fold Cross-Validation
kfold = StratifiedKFold(n_splits=10, shuffle=True, random_state=1234)
# Initialize lists to store cross-validation results
acc_scores = []
loss_scores = []
# Perform 10-fold cross-validation
for train_index, test_index in kfold.split(X, y):
 X_train, X_test = X[train_index], X[test_index]
 y_train, y_test = y[train_index], y[test_index]
 # Train the model on the current fold
 history = model.fit(X_train, y_train, epochs=100 , batch_size=32, verbose=1,
validation_data=(X_test, y_test))
# Evaluate the model on the current fold
loss, acc = model.evaluate(X_test, y_test, verbose=0)
acc_scores.append(acc)
loss_scores.append(loss)

# Print the cross-validation results
print("Crossvalidation accuracy: {:.2f}% (+/-
{:.2f}%)".format(np.mean(acc_scores)*100,
np.std(acc_scores)*100))
print("Crossvalidation loss: {:.2f} (+/- {:.2f})".format(np.mean(loss_scores),
np.std(loss_scores)))
```

Abbreviations and Acronyms

RWE	*Real-World Evidence*
RWD	*Real-World Data*
RCTs	*Randomized Clinical Trials*
EHRs	*Electronic Health Records*
HIS	*Healthcare Information Systems*
EMRs	*Electronic Medical Records*
FDA	*Food and Drug Administration*
EMA	*European Medicines Agency*
WHO	*World Health Organization*
RP	*Reference Product*
SBPs	*Similar Biotherapeutic Products*
AI	*Artificial Intelligence*
HTA	*Health Technology Assessment*
OTA	*Office of Technology Assessment*
EUnetHTA	*European network for Health Technology Assessment*
PHM	*Population Health Management*
PK	*Pharmacokinetic*
PD	*Pharmacodynamic*
PE	*Pharmacoeconomic*
QALY	*Quality-Adjusted Life Year*
IMI	*Innovative Medicines Initiative*
EH4CR	*Electronic Health Records for Clinical Research*
NLP	*Natural Language Processing*
FOrMAT	*FOCUS4 Molecularly Stratified Trial*
RADAR-CNS	*Remote Assessment of Disease and Relapse - Central Nervous System*
ML	*Machine Learning*
DL	*Deep Learning*
TREVO	*Thrombectomy Revascularization of Large Vessel Occlusions*
SARS-CoV-2	*Severe Acute Respiratory Syndrome Coronavirus 2*
MT-DTI	*Molecule Transformer-Drug Target Interaction*
IC50	*Half-maximal Inhibitory Concentration*
BATTLE	*Biomarker Approaches of Targeted Therapy for Lung Cancer Elimination*
PFS	*Progression-Free Survival*
BHMM	*Bayesian Hierarchical Mixture Model*
EXNEX	*Exchangeability-Non-exchangeability Model*
COVID-19	*Coronavirus Disease 2019*
POS	*Probability of Success*
MLAL	*Machine Learning Arthroplasty Laboratory*
THA	*Total Hip Arthroplasty*
BS	*Biosimilars*
TKA	*Total Knee Arthroplasty*
LOS	*Length of Stay*
ANN	*Artificial Neural Network*
SPL	*Sound Pressure Level*
SML	*Sound Modulation Level*
SNR	*Signal to Noise Ratio*

YNH-U	*Younger listeners with normal hearing from an urban area*
YNH-R	*Younger listeners with normal hearing from a rural area*
OHL-U	*Older listeners with hearing loss from an urban area*
OHL-R	*Older listeners with hearing loss from a rural area*
ANOVA	*Analysis of Variance*
LME	*Linear Mixed-Effects*
HR	*Heart Rate*
NIH	*National Institutes of Health*
AHA	*American Heart Association*
CI	*Confidence Interval*
mRS	*Modified Rankin Scale*
ARDS	*Acute Respiratory Distress Syndrome*
ICD-10	*International Classification of Diseases, 10th Edition*
AUCPR	*Area Under Precision-Recall Curve*
ROC	*Receiver Operating Characteristic*
PR	*Precision-Recall*
SHAP	*Shapley Additive Explanations*
LDA	*Linear Discriminant Analysis*
k-NN	*k-Nearest Neighbours*
SVM	*Support Vector Machines*
OLS	*Ordinary Least Squares*
CNN	*Condensed Nearest Neighbor*
PLS	*Partial Least Squares*
PCA	*Principal Component Analysis*
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
MLP	*Multi-layer Perceptron*
SOM	*Self-Organizing Map*
TP	*True Positives*
TN	*True Negatives*
FP	*False Positives*
FN	*False Negatives*
SMOTE	*Synthetic Minority Oversampling Technique*
ReLU	*Rectified Linear Unit*

Bibliography

1. Breckenridge, A. M., Breckenridge, R. A., & Peck, C. C. (2019). Report on the current status of the use of real-world data (RWD) and real-world evidence (RWE) in drug development and regulation. In *British Journal of Clinical Pharmacology* (Vol. 85, Issue 9, pp. 1874–1877). Blackwell Publishing Ltd. https://doi.org/10.1111/bcp.14026

2. Elia Stupka. (2019). Extended Real-World Data: The Life Science Industry's Number One Asset. Health Catalyst. https://www.healthcatalyst.com/insights/real-world-data-chief-driver-drug-development

3. Liu, F., & Demosthenes, P. (2022). Real-world data: a brief review of the methods, applications, challenges and opportunities. In *BMC Medical Research Methodology* (Vol. 22, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12874-022-01768-6

4. Eduardo Valencia. (2017). *What is Real World Evidence and why does it matter?*. Meaning.Cloud. https://www.meaningcloud.com/blog/real-world-evidence

5. Garrison, L. P., Neumann, P. J., Erickson, P., Marshall, D., & Mullins, C. D. (2007). Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force report. *Value in Health : The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *10*(5), 326–335. https://doi.org/10.1111/J.1524-4733.2007.00186.X

6. Grimberg, F., Asprion, P. M., Schneider, B., Miho, E., Babrak, L., & Habbabeh, A. (2021). The Real-World Data Challenges Radar: A Review on the Challenges and Risks regarding the Use of Real-World Data. In *Digital Biomarkers* (Vol. 5, Issue 2, pp. 148–157). S. Karger AG. https://doi.org/10.1159/000516178

7. Izmailova, E. S., Wagner, J. A., & Perakslis, E. D. (2018). Wearable Devices in Clinical Trials: Hype and Hypothesis. *Clinical Pharmacology and Therapeutics*, *104*(1), 42. https://doi.org/10.1002/CPT.966

8. Karamehic, J., Ridic, O., Ridic, G., Jukic, T., Coric, J., Subasic, D., Panjeta, M., Saban, A., Zunic, L., & Masic, I. (2013). Financial Aspects and the Future of the Pharmaceutical Industry in the United States of America. *Materia Socio-Medica*, *25*(4), 286. https://doi.org/10.5455/MSM.2013.25.286-290

9. Mahajan, R. (2015). Real world data: Additional source for making clinical decisions. *International Journal of Applied and Basic Medical Research*, *5*(2), 82. https://doi.org/10.4103/2229-516x.157148

10. Miksad, R. A., & Abernethy, A. P. (2018). Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality. *Clinical Pharmacology and Therapeutics*, *103*(2), 202. https://doi.org/10.1002/CPT.946

11. Pink, J., Lane, S., & Hughes, D. A. (2012). Mechanism-based approach to the economic evaluation of pharmaceuticals: Pharmacokinetic/pharmacodynamic/pharmacoeconomic analysis of rituximab for follicular lymphoma. *PharmacoEconomics*, *30*(5), 413–429. https://doi.org/10.2165/11591540-000000000-00000/METRICS

12. Risson, V., Saini, D., Bonzani, I., Huisman, A., & Olson, M. (2016). Patterns of Treatment Switching in Multiple Sclerosis Therapies in US Patients Active on Social Media: Application of Social Media Content Analysis to Health Outcomes Research. *Journal of Medical Internet Research*, *18*(3). https://doi.org/10.2196/JMIR.5409

13. Schneider, B., Asprion, P. M., & Grimberg, F. (2019). Human-centered artificial intelligence: A multidimensional approach towards real world evidence. *ICEIS 2019 - Proceedings of the 21st International Conference on Enterprise Information Systems*, *1*, 369–378. https://doi.org/10.5220/0007715503810390

14. Swift, B., Jain, L., White, C., Chandrasekaran, V., Bhandari, A., Hughes, D. A., & Jadhav, P. R. (2018). Innovation at the Intersection of Clinical Trials and Real-World Data Science to Advance Patient Care. In *Clinical and Translational Science* (Vol. 11, Issue 5, pp. 450–460). Blackwell Publishing Ltd. https://doi.org/10.1111/cts.12559

15. Togo, K., & Yonemoto, N. (2022). Real world data and data science in medical research: present and future. *Japanese Journal of Statistics and Data Science*, *5*(2), 769–781. https://doi.org/10.1007/s42081-022-00156-0

16. U.S. Food & Drug Administration. (2018). *Real-World Evidence*. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

17. Wise, J., Möller, A., Christie, D., Kalra, D., Brodsky, E., Georgieva, E., Jones, G., Smith, I., Greiffenberg, L., McCarthy, M., Arend, M., Luttringer, O., Kloss, S., & Arlington, S. (2018). The positive impacts of Real-World Data on the challenges facing the evolution of biopharma. *Drug Discovery Today*, *23*(4), 788–801. https://doi.org/10.1016/J.DRUDIS.2018.01.034

18. Anshari, M. (2019). Redefining Electronic Health Records (EHR) and Electronic Medical Records (EMR) to Promote Patient Empowerment. In *IJID International Journal on Informatics for Development* (Vol. 8, Issue 1). https://www.researchgate.net/publication/337443049_Redefining_Electronic_Health_Records_EHR_and_Electronic_Medical_Records_EMR_to_Promote_Patient_Empowerment

19. Banta, D. (2002). *The development of health technology assessment*. ELSEVIER. Health Policy.www.elsevier.com/locate/healthpol

20. Gherghescu, I., & Delgado-Charro, M. B. (2021). The biosimilar landscape: An overview of regulatory approvals by the EMA and FDA. *Pharmaceutics*, *13*(1), 1–16. https://doi.org/10.3390/pharmaceutics13010048

21. Hogervorst, M. A., Pontén, J., Vreman, R. A., Mantel-Teeuwisse, A. K., & Goettsch, W. G. (2022). Real World Data in Health Technology Assessment of Complex Health Technologies. *Frontiers in Pharmacology*, *13*. https://doi.org/10.3389/fphar.2022.837302

22. JOÃO ALMEIDA, A. T. N. H. P. R. J. L. O. (2018). *The European Health Data & Evidence Network Portal*. https://www.offidocs.com/public/?v=ext&pdfurl=https://www.ohdsi-europe.org/images/symposium-2019/posters/30_Alina_Trifan.pdf

23. Quiroz, J. C., Chard, T., Sa, Z., Ritchie, A., Jorm, L., & Gallego, B. (2022). Extract, transform, load framework for the conversion of health databases to OMOP. *PLoS ONE*, *17*(4 April). https://doi.org/10.1371/journal.pone.0266911

24. Beck, B. R., Shin, B., Choi, Y., Park, S., & Kang, K. (2020). Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Computational and Structural Biotechnology Journal*, *18*, 784–790. https://doi.org/10.1016/j.csbj.2020.03.025

25. Binning, M. J., Bartolini, B., Baxter, B., Budzik, R., English, J., Gupta, R., Hedayat, H., Krajina, A., Liebeskind, D., Nogueira, R. G., Shields, R., & Veznedaroglu, E. (2018). Trevo 2000: Results of a large real-world registry for stent retriever for acute ischemic stroke. *Journal of the American Heart Association*, *7*(24). https://doi.org/10.1161/JAHA.118.010867

26. Christensen, J. H., Saunders, G. H., Porsbo, M., & Pontoppidan, N. H. (2021). The everyday acoustic environment and its association with human heart rate: Evidence from real-world data logging with hearing aids and wearables. *Royal Society Open Science*, *8*(2). https://doi.org/10.1098/rsos.201345

27. Hwang, T. J., Carpenter, D., Lauffenburger, J. C., Wang, B., Franklin, J. M., & Kesselheim, A. S. (2016). Failure of Investigational Drugs in Late-Stage Clinical Development and Publication of Trial Results. *JAMA Internal Medicine*, *176*(12), 1826–1833. https://doi.org/10.1001/JAMAINTERNMED.2016.6008

28. Jorgensen, E., Xu, J., Chipara, O., & Wu, Y. H. (2023). Auditory environment diversity quantified using entropy from real-world hearing aid data. *Frontiers in Digital Health*, *5*. https://doi.org/10.3389/fdgth.2023.1141917

29. Kolluri, S., Lin, J., Liu, R., Zhang, Y., & Zhang, W. (2022). Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: a Review. In *AAPS Journal* (Vol. 24, Issue 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1208/s12248-021-00644-3

30. Lazzarini, N., Filippoupolitis, A., Manzione, P., & Eleftherohorinou, H. (2022). A machine learning model on Real World Data for predicting progression to Acute Respiratory Distress Syndrome (ARDS) among COVID-19 patients. *PLoS ONE*, *17*(7 July). https://doi.org/10.1371/journal.pone.0271227

31. Li, B., Shin, H., Gulbekyan, G., Pustovalova, O., Nikolsky, Y., Hope, A., Bessarabova, M., Schu, M., Kolpakova-Hart, E., Merberg, D., Dorner, A., & Trepicchio, W. L. (2015). Development of a drug-response modeling framework to identify cell line derived translational biomarkers that can predict treatment outcome to Erlotinib or Sorafenib. *PLoS ONE*, *10*(6). https://doi.org/10.1371/journal.pone.0130700

32. Li, M., Liu, R., Lin, J., Bunn, V., & Zhao, H. (2020). Bayesian Semi-parametric Design (BSD) for adaptive dose-finding with multiple strata. *Journal of Biopharmaceutical Statistics*, *30*(5), 806–820. https://doi.org/10.1080/10543406.2020.1730870

33. Matthay, M. A., Leligdowicz, A., & Liu, K. D. (2020). Biological mechanisms of COVID-19 acute respiratory distress syndrome. In *American Journal of Respiratory and Critical Care Medicine* (Vol. 202, Issue 11, pp. 1489–1491). American Thoracic Society. https://doi.org/10.1164/rccm.202009-3629ED

34. Ramkumar, P. N., Haeberle, H. S., Bloomfield, M. R., Schaffer, J. L., Kamath, A. F., Patterson, B. M., & Krebs, V. E. (2019). Artificial Intelligence and Arthroplasty at a Single Institution: Real-World Applications of Machine Learning to Big Data, Value-Based Care, Mobile Health, and Remote Patient Monitoring. In *Journal of Arthroplasty* Churchill Livingstone Inc. https://doi.org/10.1016/j.arth.2019.06.018

35. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, *10*(3). https://doi.org/10.1371/journal.pone.0118432

36. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature 2020 577:7792*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

37. Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(1), 97–106. https://doi.org/10.1002/wics.51

38. Abdulhussein, A., Al-Magsoosi, D., Mohammed, G. N., & Ramadhan, Z. A. (2021). Comparison and analysis of supervised machine learning algorithms. *Original Research*, *9*(4), 1102–1109. http://pen.ius.edu.ba/index.php/pen/article/view/2507

39. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. https://hastie.su.domains/Papers/ESLII.pdf

40. Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., & Robles, V. (2006). Machine learning in bioinformatics. In *Briefings in Bioinformatics* (Vol. 7, Issue 1, pp. 86–112). https://doi.org/10.1093/bib/bbk007

41. Lavecchia, A. (2015). Machine-learning approaches in drug discovery: Methods and applications. In *Drug Discovery Today* (Vol. 20, Issue 3, pp. 318–331). Elsevier Ltd. https://doi.org/10.1016/j.drudis.2014.10.012

42. Mehmed Kantardic. (2020). *Data Mining Concepts, Models, Methods and Algorithms*. https://ieeexplore.ieee.org/servlet/opac?bknumber=5265979

43. Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C. J., Dannenfelser, R., Dun, C., Edrisi, M., Elworth, R. A. L., Kille, B., Kyrillidis, A., Nakhleh, L., Wolfe, C. R., Yan, Z., Yao, V., & Treangen, T. J. (2022). Current progress and open challenges for applying deep learning across the biosciences. In Nature Research. https://doi.org/10.1038/s41467-022-29268-7

44. Smiti, A. (2020). A critical overview of outlier detection methods. In *Computer Science Review* (Vol. 38). Elsevier Ireland Ltd. https://doi.org/10.1016/j.cosrev.2020.100306

45. Douglas M. Hawkins (1980) , Identification of Outliers, vol. 11, Springer, https://link.springer.com/book/10.1007/978-94-015-3994-4

46. Subash Sigdel. (2022). *Classification in Machine Learning*. https://subashsigdel.com.np/Re/researchimg/Classification%20in%20Machine%20Learning.pdf

47. Vladimir N. Vapnik (2000) The Nature of Statistical Learning Theory, Springer https://books.google.gr/books?hl=el&lr=&id=sna9BaxVbj8C&oi=fnd&pg=PR7&ots=orG7PSmk98&sig=58-9rn60LjctoMSoqw8LF9FMQ5E&redir_esc=y#v=onepage&q&f=false

48. Grace Zhang, (2023) What is the kernel trick? Why is it important? Medium. https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-important-98a98db0961d

49. Koller D, Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen, van de Vijver MJ, West RB, van de Rijn M, (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. https://pubmed.ncbi.nlm.nih.gov/22072638/

50. Dr. Muditha M. Hapudeniya (2010) Artificial Neural Networks in Bioinformatics. "Sri Lanka Journal of Bio-Medical Informatics" https://www.researchgate.net/publication/228347923_Artificial_Neural_Networks_in_Bioinformatics

51. Lauv Patel , Tripti Shukla , Xiuzhen Huang , David W. Ussery and Shanzhi Wang (2020) Machine Learning Methods in Drug Discovery https://pubmed.ncbi.nlm.nih.gov/33198233/