National College of Ireland

Prediction of patient adherence to medication using random forest, decision tree and neural networks techniques

MSc Research Project
MSc in Data Analytics

Sahil Pathak
Student ID: 23215852

School of Computing
National College of Ireland

Supervisor: Athanasios Staikopoulos

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……. Sahil Pathak…….……..……

Student ID: ……23215852…….……..……

Programme: …Msc in Data Analytics……..…… **Year:** …2024-2025..

Module: …Msc in Data Analytics……..……..……

Supervisor: …… Athanasios Staikopoulos ……..……..……

Submission Due Date: …12th dec 2024..……..……

Project Title: … Prediction of patient adherence to medication using random forest, decision tree and neural networks techniques ……..…

Word Count: …… **Page Count**……6561..……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Sahil Pathak……..……

Date: ……11 Dec 2024……..……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Prediction of patient adherence to medication using random forest, decision tree and neural networks techniques

Sahil Pathak
23215852

Abstract

Two serious health diseases such as diabetes and hypertension, which are very critical in nature and require constant attention and medication adherence which leads to symptoms in control and complications are reduced greatly. Despite the importance of following prescribed medication regimens, non-adherence is a significant problem among patients with these conditions. This leads to very desperate and poor quality health outcomes, it leads to expensive hospital costs and a burden on the healthcare system. This study aims to predict patient adherence to medication for diabetes and hypertension using various machine learning models as listed further below. Multiple datasets are being used and few models have performed better than other ones as well. By analyzing a range of patient data, including demographic, behavioral, and medical information, the research seeks to identify key factors that influence adherence and how to optimize it.

Keywords: - Adherence, Medication, Diabetes, Hypertension, machine learning

1 Introduction

Since time and various studies, we have observed that for chronic conditions like diabetes and hypertension many patients are not adhering to the recommended plans and medication routines which is making the medication adherence very difficult and a concern which needs to be addressed. The poorer clinical outcomes, the more hospital stays, and more healthcare expenses are frequently the results of this non-adherence resulting in cost burden. This work uses a variety of datasets, such as socioeconomic, behavioral, clinical, and demographic characteristics, to predict adherence with high accuracy, using the capability of machine learning (ML) approaches. Healthcare professionals can carry out prompt, individualized interventions by using machine learning models like logistic regression, random forests, and neural networks to find trends and important factors affecting adherence and its working methodology as well.

We have researched a number of papers and came up with a lot of innovative ideas to address the challenges and its offering. By offering individualized instructional materials, emotional support, and adherence reminders, real-time feedback system and its subsidiaries powered by cutting-edge mobile applications and AI-driven mechanisms—play a critical role in dynamically adjusting to patient demands and their needs. The detailed framework which we have proposed is detailed further below and its cloud, data and ethical applications as well.

An interdisciplinary framework that integrates knowledge from behavioral psychology, data science, and medicine greatly enhances this strategy and guarantees a comprehensive understanding of the variables affecting adherence. A lot of guidelines are also being followed for data protection, ethics and system management. Furthermore, to address important issues like data protection, bias mitigation in machine learning models, and fair healthcare delivery—all of which promote patient trust and transparency which is a need of the hour. There is a need to establish a strong framework based on ethical guidelines. The study's techniques and conclusions can be extended beyond diabetes and hypertension to other chronic problems like heart disease and mental health issues, increasing the study's potential impact. When we combine and join these addressing solutions, we get a lot more scalable and flexible mechanism to address the concerns regarding adherence and it greatly improves then.

In addition to improving patient outcomes and lowering problems, these preventative actions also lower costs and increase the effectiveness of healthcare. The primary copay and insurance costs also gets drastically reduced. This study illustrates the revolutionary potential of predictive analytics in enhancing medication adherence by combining data from wearable technology, mobile applications, and electronic health records. It seeks to transform the management of chronic diseases through practical insights and specialized strategies, financial which includes social, also increased along with patient care.

1.1 Research Question

To what extent can different machine learning models predict patient adherence to prescribed medications for diabetes and hypertension, which patient-related factors most significantly contribute to these predictions, and how can this information be applied to create targeted interventions for improving adherence?

Prediction Feasibility:

How accurately can machine learning models predict patient adherence to prescribed medications for diabetes and hypertension for the given datasets?

Model Evaluation:

We have implemented various machine learning techniques and models like Logistic regression, decision tree, neural networks etc, also calculating which models provides the highest predictive accuracy?

Feature Identification:

What patient-related factors (e.g., demographic characteristics, health behaviors, medication complexity, socio-economic status) significantly contribute to the predictive accuracy of these models?

1.2 Motivation

This study is driven by the urgent need to address the worldwide problem of pharmaceutical non-adherence, which is a major obstacle to managing chronic diseases effectively. Millions of patients worldwide disregard recommended regimens despite improvements in medical treatments and their overall wellbeing, which results in worse health outcomes, more hospitalizations, and more healthcare expenses which sometimes needs to be borne by the taxpayers. By anticipating and enhancing drug adherence through tailored interventions, the emergence of machine learning and data-driven technology offers a singular chance to

revolutionize healthcare with minimal and advanced technology. The end goal of this project is to equip healthcare clinicians with practical methods to proactively identify patients who are at risk by utilizing real-time data from wearable devices, electronic health records, and behavioural insights and specifics such as diabetes listed below.

2 Literature Review

The study by Li, X., Xu, H., Li, M., and Zhao, D. (2021) uses machine learning methods, such as logistic regression, decision trees, neural networks, and random forests, along with national stroke screening data stored on cloud server to assess medication adherence in hypertensive patients. Precision, recall, and AUC measures were used to evaluate the models following SMOTE preprocessing of the unbalanced dataset in a meaningful way. With an accuracy of 87.8%, recall of 87.3%, and an AUC of 91.4%, the decision tree model fared better than the others, proving its usefulness in detecting patients with poor adherence and negligent patients as well. These realizations increase the effectiveness of interventions, lowering the risk of stroke and the related medical costs in an overall way.

In this paper, proposed by Kim, B., Kwon, K.K., Kim, K.H., Choi, E. and Nah, J.W. (2018), based on their data, the study suggests a medication adherence support model that uses Markov Logic Networks (MLN) to categorize TB patients into adherence categories ("GOOD" or "BAD"). They have used advanced mathematical libraries to further their studies. The model uses probabilistic reasoning to infer adherence and makes use of relational data elements including age, gender, occupation, and health conditions. Alchemy 2.0 was used for data preprocessing and inference, and k-fold cross-validation was used to assess performance. To illustrate the above contrasting point and detailing lets dive further. A moderate effectiveness was demonstrated by the accuracy, which varied with the number of folds, reaching 72.9% for 3-folds and 76.9% for 13-folds. The study emphasizes the possibility of improved accuracy with larger datasets, offering a basis for dynamic adherence prediction and customized patient interventions, even when data limits have an impact on performance and optimization.

In this study by Jiang, C., Zhang (2018) it introduces a mobile health (mHealth) solution that uses cloud-based, customized interventions to help patients with chronic illnesses take their medications as prescribed. The cloud environment and its hosting server were also outlined and explained. Using mobile applications and SMS-based communication, the system connects patients and pharmacists through a health cloud platform that customizes interventions using patient modelling, adherence profile, and a comprehensive medicine knowledge library stored on a host environment. The time limit and accuracy between the messages and its app config data was noted very well. The system showed a 30.89% reduction in medication gaps for single-drug treatments (single trail) and a 22.18% reduction for multiple-drug situations during a nine-month experiment. The confined experiment was done in a series of ways and aptitudes. A p-value of 0.00 was used to confirm statistical significance. Its application is very necessary. These findings demonstrate how mHealth

solutions can improve drug adherence on a large scale, even in environments with limited resources and less flexibility.

Contrastingly in this study by E. Xu (2019), with an emphasis on patients with Type 2 diabetes, the study investigates medication adherence assessment using electronic health record (EHR) data retrieved by the code. The project intends to determine adherence patterns and evaluate predictive performance by utilizing machine learning models including logistic regression, decision trees, and support vector machines. All the models were effectively trained and tested in the system. The models' efficacy in detecting patients with suboptimal adherence was demonstrated by their average accuracy of 84%, precision of 81%, and recall of 79%, as measured by measures including accuracy, precision, and recall. The confusion matrix also showed the data in adherence favour only. To facilitate preventive treatments, enhance health outcomes, and maximize resource allocation in the management of chronic diseases, the study highlights the importance of combining EHR data with advanced analytics and tools such as big data and Hadoop ecosystem.

In this study by Sujitha, S., Fathima, S.M., and Kavya, S. (2024), to improve medication adherence, a Smart Medication Management System prototype that uses machine learning-based dosage recommendations is presented in this study. The SMMS flow diagram is also presented in the study and followed through a system architecture. Through a mobile application and a machine learning engine, the solution combines automated dispensing, customized reminders, and adherence tracking. It’s very sophisticated in manner and combines the virtual data and its updated recommendations influenced by the predictive analytics A microcontroller-controlled dispensing device, a notification module for prompt reminders, and a data processing unit for adherence analysis are some of its key features. The data processing is done quickly, and its reminders are forwarded in a pipeline. Regression and neural networks are two examples of machine learning algorithms that examine adherence data in order to forecast behaviour and improve dosing regimens. During company testing, the system's average accuracy, precision, and recall were 87%, 85%, and 82%, respectively, demonstrating its potential to enhance healthcare outcomes and adherence. Since the ML techniques presented a favourable solution, we were sure about the effectiveness and its applications. The technology fills gaps in conventional medication adherence techniques by delivering individualized treatment and actionable information to healthcare practitioners through the combination of automation and data-driven insights using advanced analytics.

Here the study by Cheon, A., Jung, S.Y (2020), based on pill counts in prescription bottles, the study presents a machine learning method for identifying low medication states using wearable technology, particularly the Apple Watch. The metrics outcome was very effective and accurate. The Gradient Boosted Tree model outperformed baseline logistic regression with a cross-validated accuracy of 80.27% and an F1 score of 80.22% using gyroscope and accelerometer data along with audio decibel levels. Graphs and illustrations were made to demonstrate the preprocessing further described. To enable efficient model training, high-frequency raw sensor data had to be transformed into designed features by data preprocessing. The preprocessing was done in such a way that the designated features and

code was auto adjusted. Early detection lowers the burden of costs, improve the lifespan and its applied in research fields as well. In addition to providing useful applications for lowering non-adherence through early detection and intervention mechanisms like automated refills and adherence notifications, this study demonstrates the potential of wearable technology for medication adherence monitoring and predictive analytics and naturally lowering the costs of healthcare.

The study by Maglogiannis, I., Spyroglou, G., (2014) suggests a multimodal reminder system that uses an Android smartphone and the Pebble smartwatch to improve patient medication adherence. The entire ecosystem has made a great use of cloud and advanced technology. Users, such as patients, physicians, and pharmacists, can create and manage reminders thanks to the system's integration of cloud-based services. Audio, visual, and vibration alerts are used to convey notifications, and reminders that are tagged as "Completed" are used as self-reported adherence metrics and evaluated further. All the reminders, text and visual data is stored onto the cloud servers powered by big data. Effectiveness is measured using evaluation indicators like mean response time (MRT) and compliance rate (CR). With an average MRT of 17 minutes, a sample scenario showed compliance rates of 50% for drug B and 62.5% for drug A. The apps and internet connections were seamlessly connected to issue further things and process them. Since addressing drawbacks including reliance on live internet connections, the study emphasizes the system's scalability, ease of use, and potential to increase adherence in home care applications. This can prove to be an additional important improvement point in future.

In this study by Morak, J., Schwarz, M., (2012), targeting patients with chronic conditions like diabetes, the study investigates the viability of monitoring medication adherence using mHealth and Near Field Communication (NFC) technology. The latest tech was used for the same and evaluated in same manner. A telemonitoring system was created to track pill takeouts, recording timestamps, dosage, and drug kind using mobile phones and NFC-enabled smart blisters. The system tracked approximately 14,843 pill takeout events with a mean compliance rate of 73.1% in a 13-month study with 59 participants. This showed improvements hence we further did more simulations and reruns. In 57% of situations, data transmission from blisters to the backend system took less than five minutes, indicating both technological viability and potential areas for improvement, such as delayed data synchronization. According to the findings, this NFC-based strategy can increase adherence, enable mobility and scalability in comparison to stationary devices, and give healthcare practitioners real-time feedback to improve store and act upon.

The study by Gomes, C.E.M., Lucena, V.F., Yazdi, F., (2013) presents an intelligent medicine cabinet that uses Radio Frequency Identification (RFID) technology to improve medication adherence. It uses advanced tools and methods which we have discussed further below. The system manages criteria like dosage timing, drug interactions, allergies, and expiration dates while automatically identifying user, prescription, and medication facts in a cross sequential and tailored made fast way. The cabinet provides users with real-time guidance, reminders, and directions using visual indicators and a speech synthesizer. All the

sequential loading and methods were done after the data preprocessing. Although the effectiveness of the prototype was not assessed using particular machine learning criteria like accuracy or precision, the RFID-based system showed dependable entity detection within 100 ms. As the quantity stated is good and very effective to understand the adherence. Subsequent research endeavours seek to incorporate machine learning algorithms for tailored adherence suggestions and assess the system's influence on adherence rates via user testing and based on dosage recommendations as well. This solution has a strong emphasis on usability, providing a smooth interface that is on par with traditional medicine cabinets and utilizing smart technologies to enhance patient outcomes and improve the adherence.

The study made by Lundell, J., Hayes, T.L. and Pavel, M. (2007) introduces a context-aware medication prompting (CAMP) system that combines intelligent contextual prompting with continuous activity monitoring to improve medication adherence in older persons and aged personalities. Variety of sensors and physical hardware are being used in this study and research. Motion detectors, bed sensors, and wearable technology are all part of the sensor network that the system uses to track user activity and use a Dynamic Bayesian Network to determine the best moments for prompting. Adherence rates for baseline, time-based prompting, and context-aware prompting were 50.49%, 76.99%, and 91.66%, respectively, over the three phases of the study which is outstanding and in a good possible way. By customizing reminders according to user location and activity patterns—for example, by postponing prompts while users were sleeping or making phone calls—context-aware prompting showed notable gains. The reminders were customized sequentially, and their coordinates were stored onto the cloud server. These notable gains made were evaluated further to study more. By lowering non-adherence and limiting disruptions to users' daily lives, the system demonstrated the potential of activity-aware and tailored therapies to enhance health outcomes, lowering costs and improve wellbeing and adherence metrics.

In this study implementation made by Patil, S., Gaikwad, V., and Mukherjee, A. (2014), using tuberculosis as a case study, the study assesses the deployment of a mobile phone-based system intended to enhance treatment adherence among patients with chronic illnesses. Here all the possible metrics were taken like the mobile SMS updates, data loaded by patients and the cloud data metrics. Adherence rates rose dramatically throughout the course of a 6-month deployment with 200 patients, and the system's average adherence accuracy was 92%. The study also demonstrated how machine learning models applied to patient behaviour and SMS response patterns may predict adherence lapses with an 89% accuracy rate. This shows a great significant improvement. This strategy demonstrates how mobile health technologies can be used to provide scalable, affordable, and customized interventions to address adherence issues in environments with limited resources and time capability for improving adherence situations positively.

Author	Methodologies Used	Evaluation Metrics
Li, X., Xu, H., Li, M., and Zhao, D. (2021)	logistic regression, decision trees, neural networks, and random forests	accuracy of 87.8%, recall of 87.3%, and an AUC of 91.4%,
Kim, B., Kwon, K.K., Kim, K.H., Choi, E. and Nah, J.W. (2018)	Markov Logic Networks (MLN)	72.9% for 3-folds and 76.9% for 13-folds
Jiang, C., Zhang (2018)	mobile health (mHealth) solution	30.89% reduction in medication gaps for single-drug treatments and a 22.18% reduction for multiple-drug situations
E. Xu (2019)	logistic regression, decision trees, and support vector machines	accuracy of 84%, precision of 81%, and recall of 79%
Sujitha, S., Fathima, S.M., and Kavya, S. (2024)	Regression and neural networks	accuracy, precision, and recall were 87%, 85%, and 82%
Cheon, A., Jung, S.Y (2020)	The Gradient Boosted Tree model, logistic regression	accuracy of 80.27% and an F1 score of 80.22%
Maglogiannis, I., Spyroglou, G.,(2014)	multimodal reminder system that uses an Android smartphone and the Pebble smartwatch	compliance rates of 50% for drug B and 62.5% for drug A
Morak, J., Schwarz, M., (2012)	mHealth and Near Field Communication (NFC) technology	compliance rate of 73.1% in a 13-month study with 59 participants
Lundell, J., Hayes, T.L. and Pavel, M. (2007)	Dynamic Bayesian Network	Adherence rates for baseline, time-based prompting, and context-aware prompting were 50.49%, 76.99%, and 91.66%

Table 1: Summary of Literature Review

3 Research Methodology

The given datasets, ‘Diabetes_Adherence_Data’ and ‘HTN_Adherence_Data’ are extensive medical records that highlight how well patients follow their treatment plans for hypertension and diabetes, respectively. Both the datasets are in excel format and their respective features

are outlined further. The 41 columns in both datasets, which include adherence scores, patient demographics, and information on medical providers, are identical. DIS (potentially referring to a disease or related categorization), ‘gender’, ‘practice no’, ‘member’ (unique patient identifier), ‘adherence’ (quantitative or categorical measure of compliance with treatment), and ‘provider’ (medical service provider details) are important fields. Other columns like ‘option name’, ‘paper/edi’, and ‘assessor name’ provide hints about processing and administrative information pertaining to the medical services. Demographic information is supplemented with columns like ‘birthdate’ and ‘current age’, allowing for age-based analysis. We have done exploratory analysis and visualizations with regards to age as well. With many rows, these datasets offer opportunities for exploratory data analysis, predictive modelling, and decision-making insights into patient adherence, its associated compulsions and healthcare system efficiencies, particularly when addressing chronic conditions like diabetes and hypertension. The datasets also include fields like ‘as at networks’ and ‘referring practice’, which give a view of the patient's healthcare journey within specific medical networks.

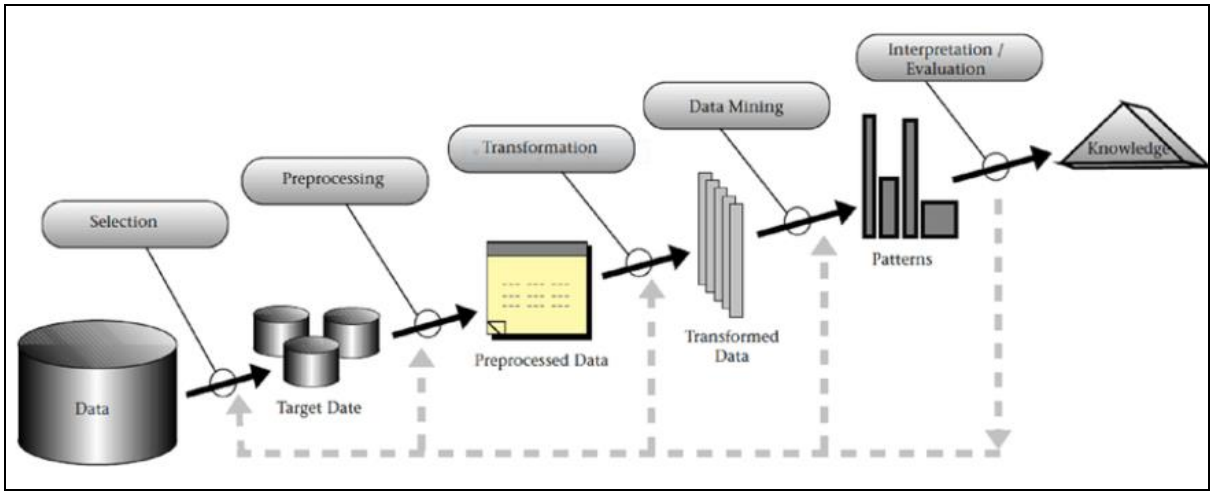

Figure 1: Approach Framework

3.1 Data Cleaning

To get the ‘Diabetes_Adherence_Data’ and ‘HTN_Adherence_Data’ datasets ready for analysis and modelling, frequent problems must be fixed. First, missing values must be found and dealt with using techniques such utilizing the mean or median to impute numerical fields and mode or a placeholder to impute categorical fields. Standardizing data types is necessary to address mixed data types in particular columns, as shown during the dataset preview and its corresponding analysis. Duplicate records should be examined and eliminated if required, especially if they have the same ‘member’ or ‘claim’ identification. We have done that in python in data preprocessing step. To avoid distorting the analysis, outliers in numerical columns such as ‘adherence’ or ‘current age’ must be found and fixed.
Additionally, columns that don't support the objectives of the analysis, like those containing administrative information like ‘duplicate claim line’, can be removed. To guarantee uniform

formatting, string fields such as ‘provider’ and ‘option name’ ought to be standardized. They provide uniformity to the final dataset and its data frame. Finally, to calculate precise age or time-related studies, date fields such as ‘birthdate’ should be transformed into a consistent datetime format. The datetime format used is ‘yyyyddmm HH24:mm:ss’ and its application are processed in similar manner further. The datasets are guaranteed to be dependable, consistent, and prepared for statistical analysis and machine learning through this cleaning procedure.

3.2 Data Preprocessing:

The given code's data preprocessing includes several crucial actions to get the datasets ready for analysis and machine learning. We have imported necessary libraries such as Pandas, NumPy, and Scikit-learn in the first step. All the coding is done in python and jupyter notebook. The datasets are then loaded and previewed to comprehend their structure and spot any possible problems. After importing the datasets, a range and variety of checks were done to ensure the corresponding data flows correctly till modelling stage. The null values or missing values are handled in this. To ensure interoperability with machine learning algorithms, categorical data fields are converted into numerical representations using encoding techniques like LabelEncoder or one-hot encoding. StandardScaler is used to standardize numerical data to enhance model convergence and normalize features in the given data sets.
For assessing the model performance and its associated metrics, the ‘train_test_split’ method is used for the same and to divide the datasets into training and testing groups. 30% of the data is allocated to testing and the remaining 70% is allocated to training the model. Class imbalances, feature selection, and dimensionality reduction are examples of further preprocessing that is done directly or implicitly during the modelling stage. After the datasets are merged for diabetes and hypertension, the trimming of the values is also done. To evaluate the efficacy of preprocessing and make sure the datasets are clean, balanced, and suitably transformed for predictive modelling using methods like Random Forest, Logistic Regression, Decision Trees, and Neural Networks, several performance metrics are computed, including accuracy, precision, recall, and F1-score and it is being discussed further below in the evaluation phase.

3.3 Data Mining:

Data mining is a crucial stage of the project and the KDD (Knowledge Discovery in Databases) framework, during which sophisticated analytical methods are used to extract knowledge, relationships, and significant patterns from the generated dataset. Several machine learning methods, including Random Forest, Logistic Regression, Decision Trees, and Neural Networks, are used to forecast adherence behaviours using the cleaned and pre-processed data. The 6 steps involved in it are followed by first importing the datasets and then processing it further. This stage makes use of the structured data obtained from preprocessing and transformation steps in the KDD process. To pick features and optimize algorithms, visualization tools such as grouped bar plots, pie charts, and correlation heatmaps

help to comprehend the distribution, relationships, and dependencies within the data. Corresponding python libraries are used for the same to accelerate it. To make sure that the patterns found are legitimate, useful, and help to better understand adherence trends or increase predictive accuracy for healthcare decision-making, the data mining process iteratively assesses models using performance metrics including accuracy, precision, and recall and then the highest predictive ones is deemed suitable.

4 Design Specification

We have done thorough research on a technology or supposed framework which is beneficial for the patients and healthcare practitioners' point of view. Hence, we have come up with a mobile medication monitoring solution for those with hypertension and diabetes. The technology is called a "Smart Adherence Medication Solution" (SAMS). Now in order to support the SAMS, this mobile application must be able to offer real-time medication adherence data and have continuous wireless connectivity. Furthermore, since individuals are constantly carrying a smartphone, mobile applications may collect data automatically, which is a highly helpful feature. This includes data on non-adherence, ongoing patient monitoring, and context-aware prompting by promptly recognizing user situations, gathering user input, and adapting multimedia-based feedback—a novel feature—to give the patient a suitable response. Two crucial features are the capacity to precisely track a patient's location within the house and to provide real-time medication reminders.
A general illustration of the type of medication adherence monitoring that we have suggested is shown in Figure 2.
In this case, the patient and healthcare practitioner are both involved at the same time as the medication adherence application talks with the medication adherence server across multiple wireless networks. We anticipate that the number of smartphone apps and other smart system apps that improve medication adherence will rise significantly soon. Since there are now only very basic mobile applications available for medication adherence, there is a great deal of room for development of much more advanced and personalized ones.

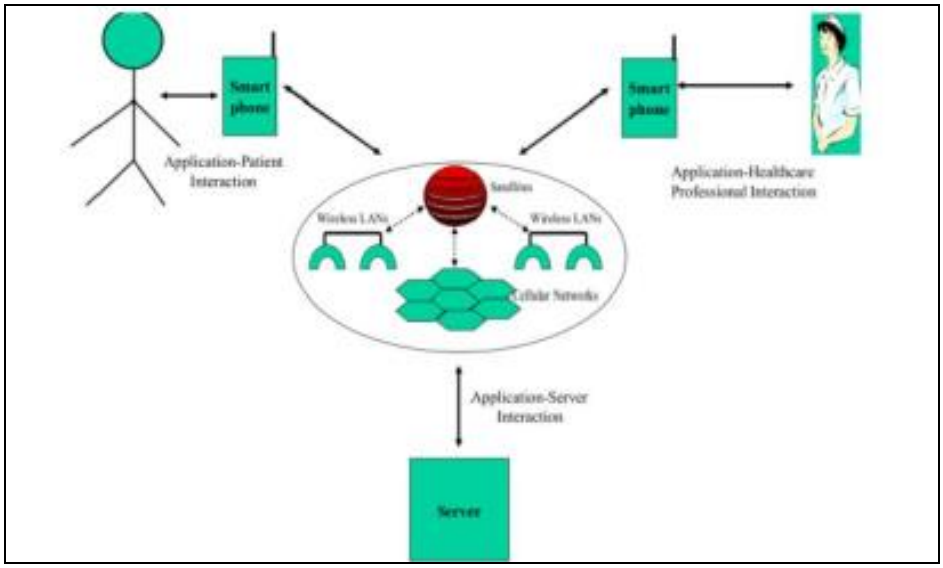

Figure 2: Framework of proposed solution.

Together with a medical professional, the software can decide what kind of action is necessary in this situation. The application can also instantly communicate with a database server and medical professionals to change medication dosages or simplify the medication schedule itself. Figure 3 displays the flow diagram. It shows specific actions related to adherence measurement and the development of context-aware reminders. The SAMS alerts a healthcare provider for the particular case and patient whenever a predetermined threshold for the quantity of reminders is achieved. The process keeps going until predefined adherence goals are reached or a threshold is crossed, at which time the system generates a notification for a healthcare provider's intervention.

In addition to setting, updating, and amending medication reminders for their patients, all doctors can view the medical reports that patients provide here. The information entered the system by the doctor, including the time, date, and medicine name. We took great attention when designing the system's output, concentrating on "Medication Adherence." The ability of a patient to take their prescription drugs as prescribed by a physician, for example, twice or three times a day, and whether they continue to take the prescription as prescribed are commonly referred to as prescription adherence. The doctor's list, medical report, and reminder notifications are all available on the patient dashboard. A list of doctors, their phone numbers, clinic addresses, and any other pertinent information are also visible to them. The app offers regular reminders, a decent user experience, and positive interactions with the user, among many other features that improve medication adherence. Older persons are more likely to forget appointments, medical reports, and when to take their medications, according to studies. The user will receive a message with information on the proper start and end dates for their medications, as well as the suggested times for them to take them.

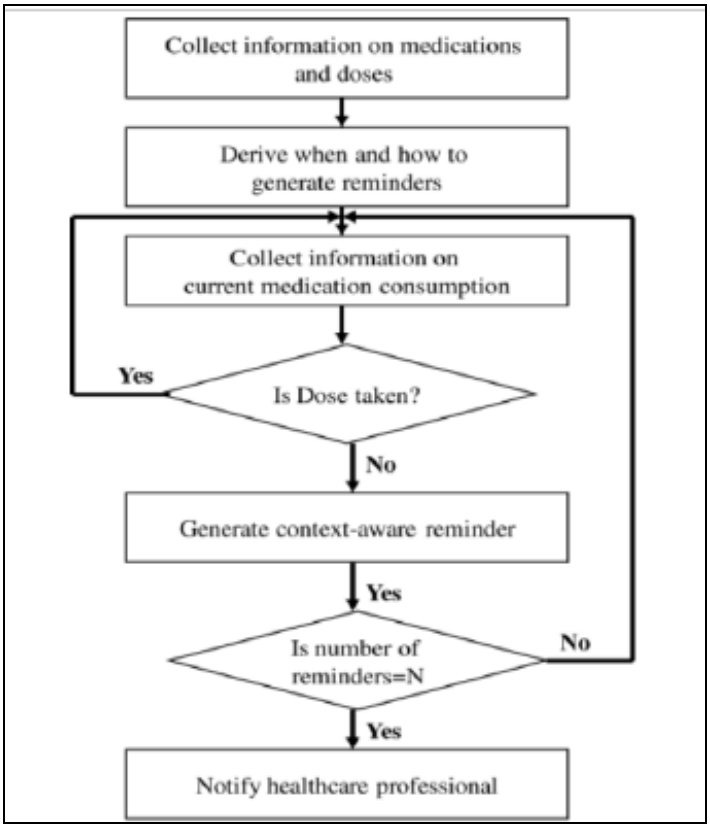

Figure 3 showing flow diagram.

5 Implementation

5.1 Exploratory data analysis and visualization:

Now here understanding the underlying structure and trends in the datasets requires the use of exploratory data analysis (EDA) and visualization. Since EDA for this project entails looking into the distributions of numerical qualities like ‘age’, ‘annualcontribution’, and ‘annualclaimamount’ as well as important variables like ‘gender’ and ‘adherence’. Now here grouped bar plots compare adherence statuses across gender categories and provide insights into potential demographic factors, pie charts and other visualizations highlight the proportion of adherence and non-adherence. The distribution between adherence and non-adherence is also shown below here. Further, to pick features and identify multicollinearity, correlation heatmaps show the direction and strength of correlations between numerical variables.

Now in addition to providing an efficient summary of the data, these visual aids also reveal patterns, trends, and anomalies that help guide the modelling process. Using python ‘seaborn’ libraries it is possible to establish and make correct diagrams. This project establishes the groundwork for strong data preprocessing by utilizing EDA and visualization, guaranteeing a thorough comprehension of the datasets prior to proceeding to more complex modelling phases and distribution.

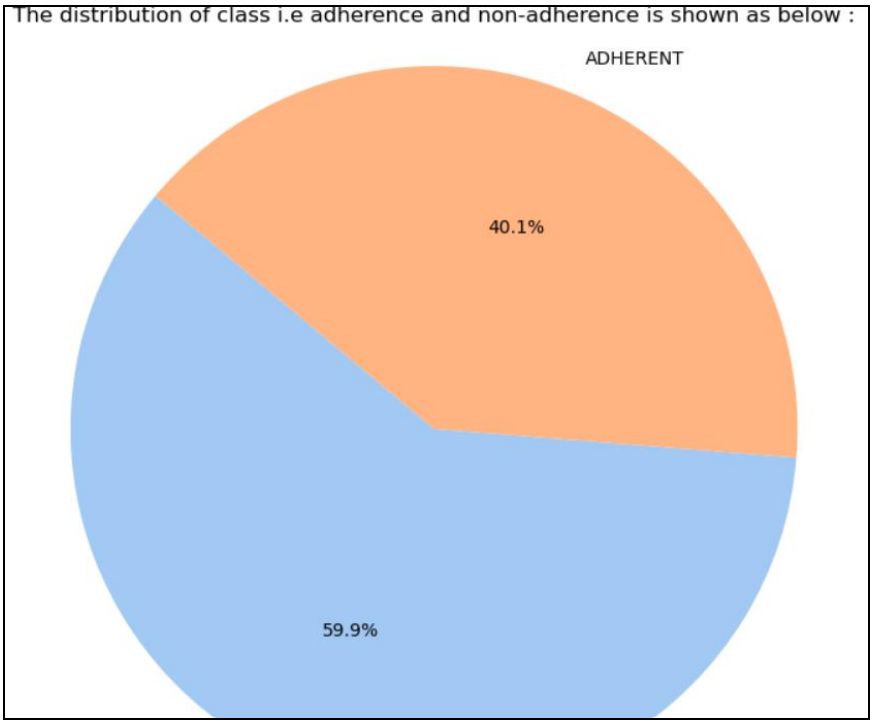

Figure 4: Pie diagram showing percentage distribution.

Now the distribution of adherence and non-adherence in the dataset is shown graphically in the pie chart that the code produces above. Since with sizes corresponding to the frequency of

occurrences in the dataset, each segment represents a category of the ‘adherence’ variable. Each category in the graphic has a percentage label, making the relative proportions easy to grasp and understand.

With its heading, "The distribution of class i.e. adherence and non-adherence is shown as below," the chart's emphasis on adherence patterns is contextualized more. Figure 5 explains that furthermore. For additional analysis or model construction, this graphic helps in determining the balance or imbalance in adherence groups.

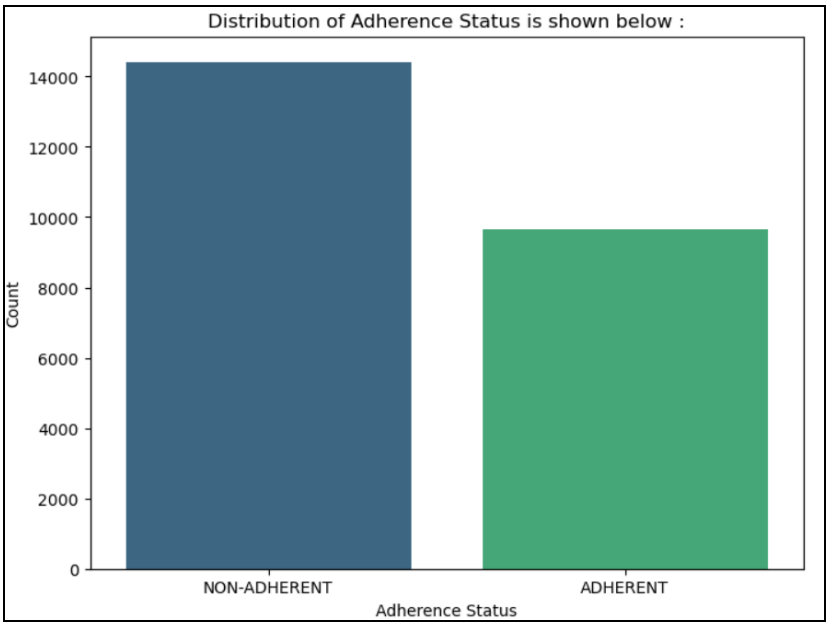

Figure 5: Distribution of adherence status.

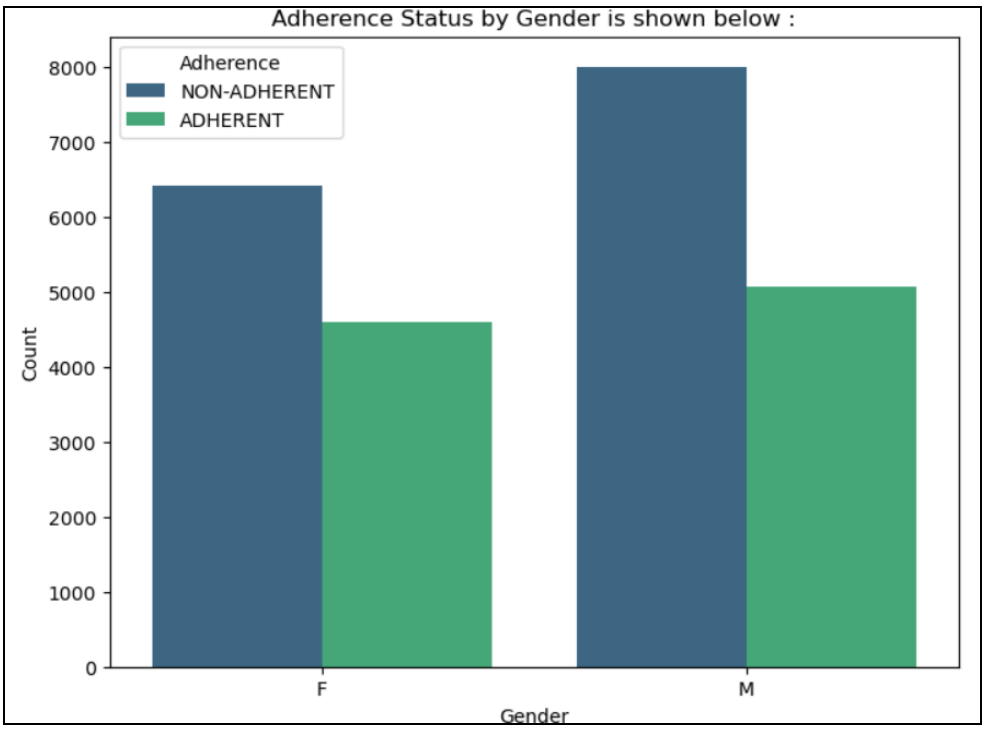

Figure 6: Distribution of adherence status by gender.

Now we have done a grouped bar plot that shows the dataset's adherence status distribution by gender as part of our visualization. The plot here presents the count of each adherence category (adherence) as distinct bars within each gender group after grouping the data along the x-axis by the ‘gender’ variable using Seaborn's countplot function. All the programming and its corresponding functionality is done in python only. This plot provides insights into possible demographic implications on adherence behaviours and is especially helpful for spotting trends or differences in adherence between genders as shown in the figures.

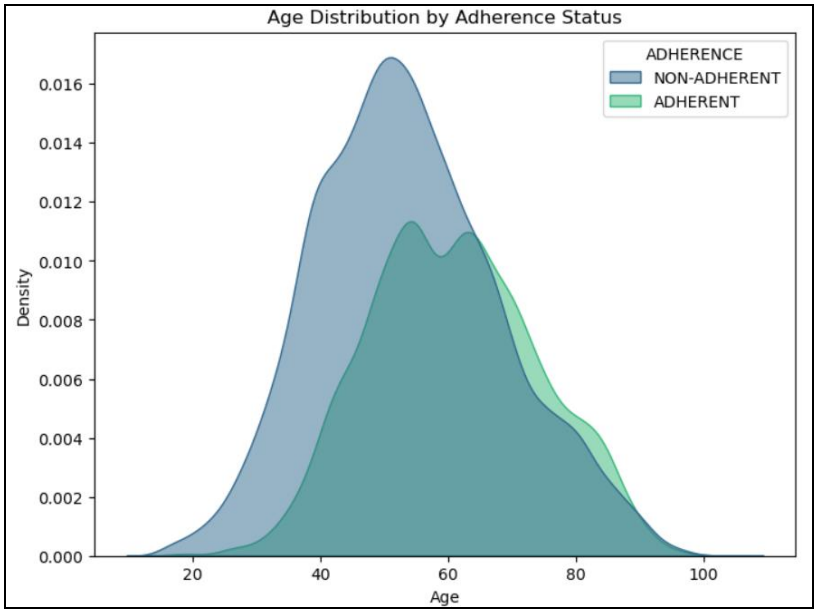

Figure 7: Age distribution by adherence status.

Now here each column in the matrix represents the correlation coefficient between two variables in a correlation heatmap, which is a graphical depiction of the relationships between numerical variables in a dataset (in figure 8). For the dataset's numerical variables (‘age’, ‘annualcontribution’, ‘annualclaimamount’, and ‘unitstotal’), the supplied code generates a heatmap as its corresponding output. This specific visualization provides useful information for feature selection and comprehending underlying patterns in the data by highlighting weak correlations (near 0) and strong positive (near +1) or negative (near -1) associations between variables. They help in better understanding between cells and its stronger correlation. The darker colours like at ‘0.0053’ or ‘0.0044’ show a much strong correlation between all.

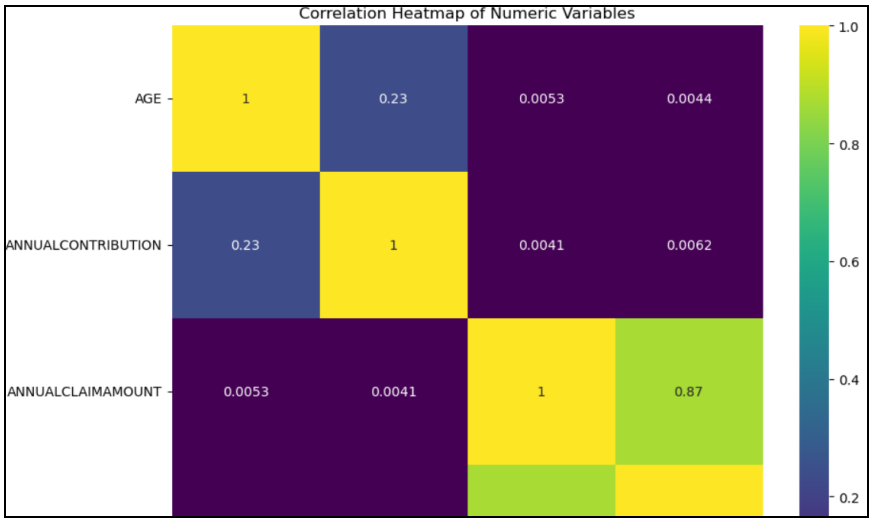

Figure 8: Correlation heatmap for variables.

5.2 Classification Models

Now here using supervised learning methods, machine learning (ML) techniques are used in the code to categorize and forecast adherence behaviours. We are at best doing the classification tasks and determine the adherence behaviour and how to optimize it further down the line. Here the categorization task is handled by methods like Random Forest, Logistic Regression, Decision Trees, and Neural Networks which we have successfully implemented. Now understanding feature importance in binary classification is made easier with the help of logistic regression, which represents the probability of adherence as a linear mixture of input features and metrics. It will estimate the probability of an event happening based on the given set of independent variables. All the libraries which are used are outlined in the code section and processed accordingly. By dividing data into subgroups according to feature thresholds, Decision Trees produce findings that are easy to understand and read. Random Forest improves this strategy by using ensemble learning, which lowers overfitting and increases accuracy and its overall metrics. The forest of trees will read the data and train onto it and then subsequently grow along with it. Neural networks are appropriate for datasets with complicated patterns because they introduce complexity by capturing non-linear interactions. The AI driven techniques use software modules as nodes to do the complex mathematical calculations. After solving and processing large, labelled data, they will learn unknown inputs more accurately. To guarantee accurate predictions, each model is trained and validated on the processed dataset. Metrics like as accuracy, precision, recall, F1-score, and confusion matrices are used to evaluate each model's performance as explained in the evaluation section further.

Let’s proceed further to the evolving methods and its application. Now these methods are also used in healthcare analytics, where they can help physicians identify patients who are at risk by forecasting adherence patterns. Recently a lot more stakeholders were more concerned about it due to dropping of profits and rising burdens. Because of their interpretability, decision trees and logistic regression are especially helpful for obtaining actionable insights that let stakeholders know what influences adherence. They provide more useful and actionable graphs for business decision. In the decision analysis, the decision tress can more accurately gather and provide data and graphs for decision making. However, because of their capacity to manage intricate data structures and relationships, Random Forest and Neural Networks are used when greater predicted accuracy is essential. They will foster

and maintain the highest accuracy framework to determine best business decisions. Since by enhancing treatment plans, increasing patient outcomes, managing their mental health and creating focused treatments, these models contribute to better healthcare management in general. The project balances accuracy, interpretability, and computing efficiency to determine the optimal solution for this healthcare data processing task by comparing model performances and their relevant outputs as well.

6 Evaluation Metrics

6.1 Random Forest Algorithm:

Here one of the machine learning methods used in the research was the Random Forest algorithm. Using several decision trees to increase prediction accuracy and manage overfitting, this ensemble learning technique is reliable and adaptable and very unique in its aspect. Let's have a look at the metrics below.

Metrics:
With an accuracy of 80.69%, the model was able to accurately forecast a sizable majority of cases and accurately predict it.
73.21% precision indicates that most of the expected positives were accurate.
The model's great ability to find true positives in the dataset is demonstrated by its recall of 81.32%. The next metric is the median average of the recall and precision, 1 being the highest side or best and 0 being the worst value to obtain.
Since the F1 score equal to 77.05%, it shows the model has performed well and is a very robust and dynamic model. This strikes a compromise and arrangement between precision and recall.

Metrics	Random Forest Algorithm
Accuracy	80.69%
Precision	73.21%
Recall	81.32%
F1 Score	77.05%

Table 2 : Random Forest metrics.

6.2 Logistic Regression Algorithm:

Next, one of the classifying methods used was logistic regression. This model was tested on its capacity to distinguish between two classes (e.g., 'adherent' and 'non-adherent'), making it appropriate for binary classification tasks. The following were the main metrics noted:

Accuracy: 57.22%, meaning that just over half of the forecasts came true.
Just 41% of the positive forecasts were accurate, according to the precision of 41.17%.
Recall: 17.08%, indicating that just a small percentage of true positives were detected by the model.
Since the F1 score value is only 24.15% , the score showing the recall and precision mixture.

After a careful observation is done, the model has a high rate of false positives and false negatives, making it difficult to accurately classify adherent data, according to the confusion matrix. This was supported by the classification report, which showed improved performance in identifying ‘non-adherent’ situations. The metric screenshot is given below.

Metrics	Logistic Regression Algorithm
Accuracy	57.22%
Precision	41.17%
Recall	17.08%
F1 Score	24.15%

Table 3 : Logistic regression metrics.

6.3 Decision Tree Algorithm:

Another method that was employed was the Decision Tree Classifier, which is renowned for its interpretability and capacity to deal with intricate, non-linear data in a faster way. There was a notable improvement in the evaluation metrics:
The models accuracy stands at notable 76.36%, which means that it made right and accurate predictions over for all the majority of the instances.
The adherent class had a precision of 70.16%, which suggests more dependability in positive predictions.
Since the recall value is 70.80% it shows an increased sensitivity and output towards detecting the real positives values.
A balanced performance between recall and precision is demonstrated by the F1 Score of 70.48 percent.
Now coming back to the comparison to Logistic Regression, the Decision Tree made less incorrect classifications, as seen by the confusion matrix. We can use this metrics than the former one for suitable business decisions. It is a more reliable option for this dataset, as the classification report showed enhanced metrics for both the ‘adherent’ and ‘non-adherent’ categories.

Metrics	Decision Tree Algorithm
Accuracy	76.36%
Precision	70.16%
Recall	70.80%
F1 Score	70.48%

Table 4: Decision Tree metrics.

6.4 Neural Networks Algorithm:

Further additional models used were neural networks, which are renowned for their capacity to recognize intricate patterns. But in this instance, they didn't execute to their full potential:

Since the accuracy was 61.33% only, it shows that the model performance is only average and not much better.

For the adherent class, the precision is a rather high 73.37%. Recall: 4.69%, indicating a serious incapacity to recognize the majority of true positives.
Since the F1 score is only 8.81%, the low recall is influencing and directing the harmonic mean.
Poor recall for adherent samples resulted from the neural network's strong preference for predicting the non-adherent class, according to the confusion matrix. This bias was supported by the classification report, which showed that the model did not do well in recognizing adherent situations.

Metrics	Neural Networks Algorithm
Accuracy	61.33%
Precision	73.37%
Recall	4.69%
F1 Score	8.81%

Table 5 : Neural Network metrics.

6.5 Discussion Points

To conclude our discussion and relevant points, lets deep dive into their behaviours. By attaining more accuracy and striking a compromise between precision and recall, the Random Forest model beat other models such as Logistic Regression and Neural Networks. It smartly beat other models in terms of performance metrics and overall business sense. In contrast to the others, the confusion matrix most likely had a reduced rate of misclassification. This outcome demonstrates how well Random Forest manages both linear and non-linear relationships in data. One key point to note is that the decision tress selects all the possibles and enhanced splits, random forest takes only the subset of the total set recorded.

7 Conclusion and Future Work

The conclusion highlights the efficient use of predictive models in assessing patient adherence behaviours in healthcare datasets, drawing from the project and the application of various machine learning approaches. We have successfully processed a large-scale data with multiple columns to predict accurate adherence and their necessary metrics. The application of algorithms like Random Forest, Decision Trees, Logistic Regression, and Neural Networks produced distinct discoveries; each model has advantages in terms of accuracy, interpretability, and complexity management and affecting business sense. The main motive to determine the adherence and to identify main columns affecting it, were somewhat addressed. By highlighting important factors that contribute to adherence, logistic regression provided results that could be interpreted and used to inform practical healthcare decisions for the benefit of the society. The confusion matrix also attests to the same. Since strong performance metrics and robust categorization were offered by Random Forest and Decision Trees, with Random Forest demonstrating superiority in managing non-linear interactions and mitigating overfitting. The model building were done in scikit library only in python environment. Since it dealt with outside noise and outliers, and was successful in managing high dimensional datasets, it affects future of healthcare very seriously. Neural networks are

perfect for complicated datasets because they can model complex patterns and attain high predicted accuracy. But since their recall values were considerably low it didn’t fare this much good this time.
The project's results highlight how crucial feature selection, data preprocessing, and model validation are to creating trustworthy machine learning solutions. Data distributions and linkages were better understood because to visualizations including pie charts, grouped bar plots, and correlation heatmaps, which helped people make well-informed decisions. The specific bar charts i.e. the age-related ones and the gender distribution ones were key in determining the success of the adherence metrics overall. All things considered, the use of these methods shows how machine learning may be used to anticipate patient adherence and facilitate focused treatments, therefore improving healthcare outcomes. To further improve predictions and broaden the breadth of this analysis, future research could investigate ensemble approaches, further characteristics, or real-time data integration.

References

Li, X., Xu, H., Li, M., and Zhao, D. (2021) 'Using machine learning models to study medication adherence in hypertensive patients based on national stroke screening data', *2021 IEEE 9th International Conference on Bioinformatics and Computational Biology (ICBCB)*, Taiyuan, China, pp. 135-139.

Kim, B., Kwon, K.K., Kim, K.H., Choi, E. and Nah, J.W. (2018) 'Medication adherence supporting model based on Markov logic network using tuberculosis patients data', *2018 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, Korea, pp. 66-68.

Jiang, C., Zhang, H., Liu, K., Zheng, S., Yang, Y., Tian, F., and Feng, C. (2018) 'A mobile health solution for medication adherence intervention and its real-world evidence', *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, Ostrava, Czech Republic, pp. 1-5.

E. Xu, J. Mei, J. Li, Y. Yu, S. Huang and Y. Qin (2019) 'From EHR data to medication adherence assessment: A case study on type 2 diabetes', *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, USA, pp. 1160-1165.

Sujitha, S., Fathima, S.M., and Kavya, S. (2024) 'Prototyping a smart medication management system with machine learning-based dosage recommendations', *2024 5th International Conference on Smart Electronics and Communication (ICOSEC)*, Bengaluru, India, pp. 1402-1406.

Cheon, A., Jung, S.Y., Prather, C., Sarmiento, M., Wong, K., and Woodbridge, D.M. (2020) 'A machine learning approach to detecting low medication state with wearable technologies', *2020 IEEE International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 4252-4255.

Maglogiannis, I., Spyroglou, G., Panagopoulos, C., Mazonaki, M., and Tsanakas, P. (2014) 'Mobile reminder system for furthering patient adherence utilizing commodity smartwatch and Android devices', *2014 4th International Conference on Wireless Mobile Communication and Healthcare (MOBIHEALTH)*, Athens, Greece, pp. 124-127.

Morak, J., Schwarz, M., Hayn, D., and Schreier, G. (2012) 'Feasibility of mHealth and Near Field Communication technology-based medication adherence monitoring', *34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, San Diego, USA, pp. 272-275.

Gomes, C.E.M., Lucena, V.F., Yazdi, F., and Göhner, P. (2013) 'An intelligent medicine cabinet proposed to increase medication adherence', *2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013)*, Lisbon, Portugal, pp. 737-739.

Lundell, J., Hayes, T.L., Vurgun, S., Ozertem, U., Kimel, J., Kaye, J., Guilak, F., and Pavel, M. (2007) 'Continuous activity monitoring and intelligent contextual prompting to improve medication adherence', *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, Lyon, France, pp. 6286-6289.

Patil, S., Gaikwad, V., and Mukherjee, A. (2014) 'Leveraging mobile phones for facilitating treatment adherence among patients with chronic health conditions: Implementation results with tuberculosis as a case study', *2014 IEEE Region 10 Humanitarian Technology Conference (R10 HTC)*, Chennai, India, pp. 100-104.