Journal of Basic Physical Research ISSN: 2141-8403 Prints, 2141-8411 Online
Vol.13, August, 2024, pp. 42 - 60. Available online at www.jbasicphyres-unizik.org

HARNESSING THE POTENTIALS OF MACHINE LEARNING ALGORITHMS IN INFORMATION TECHNOLOGY FOR PREDICTIVE HEALTHCARE ANALYTICS

[1*]Okechukwu, O. P.; [2]Ekwealor, O.U.; [3]Paul, R.U.
[1,2,3] Department of Computer Science, Nnamdi Azikiwe University, Awka, Anambra State
[1]op.okechukwu@unizik.edu.ng [2]ou.ekwealor@unizik.edu.ng*
[3]ru.paul@unizik.edu.ng

Abstract
This research investigates the fundamental human right to access high-quality healthcare by leveraging machine learning algorithms within information technology systems for predictive healthcare analytics. Specifically, it focuses on utilizing healthcare data to accurately predict disease risks, overcoming the challenges posed by incompatible IT systems. By delving into various aspects of machine learning processes, including data acquisition, preprocessing, model selection, and evaluation within the context of chronic disease prediction—diabetes in this instance—the study demonstrates the potential of integrating machine learning algorithms into information technology to enhance clinical decision-making, optimize operational efficiency, and improve patient outcomes. The methodology employed follows the Feature-Driven Development (FDD) approach, a subset of Agile Methodology, an approach that guides the development process. A Stack Ensemble Technique, combining multiple machine learning models, is employed for enhanced predictive accuracy. The Logistic Regression model achieved an accuracy of 74.54%, a precision of 74.56%, a recall of 74.54%, and an F1-score of 74.53%. The Decision Tree model attained an accuracy of 66.61%, a precision of 66.63%, a recall of 66.61%, and an F1-score of 66.6%. The Random Forest model demonstrated superior performance with an accuracy of 79.76%, a precision of 79.85%, a recall of 79.76%, and an F1-score of 79.75%. Furthermore, Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were employed for comprehensive evaluation. This research underscores the significant potential of integrating machine learning and IT systems to enhance healthcare delivery. By effectively predicting disease risks, optimizing resource allocation, and improving clinical decision-making, this approach contributes to a more efficient and effective healthcare system. The findings from this study provide a foundation for future research to explore the effectiveness of other ensemble techniques, incorporate additional features, and delve into interpretability methods to gain deeper insights into the model's decision-making process.
Keywords: *Healthcare; Machine Learning; Information Technology; Prediction; Data acquisition.*

Introduction

The healthcare organizations are producing heaps of data at alarming rate. This data comprises of medical records, genome-omics data, image scan or wearable medico device data that presents immense advantages and challenges at the same time. Data Science has emerged as a powerful approach in the field of Health Informatics, leveraging advanced analytical techniques to extract valuable insights from vast and diverse healthcare data. With the advent of electronic health records, wearable devices, and other digital health technologies, healthcare systems are generating an unprecedented volume of data, often referred to as Big Data. The incorporation of machine learning into healthcare has transformed the landscape of disease detection, allowing for a paradigm shift from reactive to proactive approaches. Khan (2023) put it this way that data science in Health Informatics involves the integration of computational, statistical, and machine learning methods to analyze and interpret this data, facilitating evidence-based decision making, personalized medicine, and improved patient outcomes. This paper provides an overview of the applications of Data Science in Health Informatics, highlighting its potential to transform healthcare delivery, disease surveillance, and biomedical research.

According to Rasool *et al.,* (2023), early disease detection not only has the potential to enhance patient outcomes, but also reduces the burden on healthcare systems by a substantial

amount. Hence the integration of machine learning has ushered in a new era of predictive potential, transforming the landscape of early disease detection.

According to Najjar (2023) healthcare systems grapple with the increasing complexity of patient data, the integration of ML algorithms becomes instrumental in extracting meaningful patterns, predicting outcomes, and optimizing resource allocation. Therefore, the most significant benefits of Machine Learning Algorithms in Information Technology for Predictive Healthcare Analytics are its ability to simultaneously consider a large number of variables. Traditional diagnostic methods may concentrate on a small number of indicators, but machine learning algorithms can incorporate a vast array of variables, including genetic information, medical history, lifestyle choices, and environmental conditions. Harnessing the potential of ML algorithms within the realm of IT for predictive healthcare analytics signifies a paradigm shift, offering a proactive approach to healthcare management and decision-making.

In the information age, healthcare systems are grappling with the increasing complexity of patient data, generated at an unprecedented rate. Personal and medical data within the healthcare industry are exceptionally sensitive due to their confidential nature. Safeguarding patient privacy through secure data storage, transmission, and access controls is paramount. The core concept of harnessing the potential of Machine Learning (ML) algorithms within Information Technology (IT) for predictive healthcare analytics is to address these challenges. The quality, heterogeneity, and interoperability of healthcare data pose significant obstacles. Integrating diverse data sources, ranging from electronic health records to wearable devices, while maintaining accuracy and reliability is a formidable task. Furthermore, the utilization of patient data for predictive healthcare analytics raises ethical concerns regarding privacy and consent. The scarcity of skilled professionals capable of interpreting and leveraging analytics insights in healthcare settings, particularly in developing nations, is another critical issue.

The aim of this research is to investigate and demonstrate the transformative impact of integrating Machine Learning (ML) algorithms within Information Technology (IT) frameworks for predictive healthcare analytics. This involves evaluating the scope, challenges, and opportunities in leveraging ML for predictive analytics in healthcare, assessing how predictive analytics can enhance the efficient utilization of healthcare resources, including personnel, equipment, and facilities, leading to cost-effective and sustainable healthcare practices, developing strategies to address privacy concerns, ensure informed consent, and uphold ethical standards in the collection, processing, and utilization of patient data for predictive healthcare purposes, and exploring how predictive analytics can contribute to the timely identification of health issues, enabling proactive intervention and improved patient outcomes.

Related Works

The convergence of information technology and utilization of machine learning algorithms within the realm of information technology has paved the way for transformative advancements in predictive healthcare analytics with enhancement in the accuracy of healthcare predictions, optimize resource allocation, and improve patient outcomes. Various forms of machine learning algorithms, such as supervised, unsupervised, and semi-supervised learning methods have shown remarkable performance in early disease identification. Examples of these methods are support vector machines (SVMs) and deep neural networks (DNNs). These algorithms are particularly good at learning from labeled data, which makes them suitable for jobs like predicting the risk of diabetes, cardiovascular disease and diagnosing cancer etc.

Rajkomar, *et al.* (2019) reviewed the diverse applications of machine learning in healthcare, emphasizing predictive analytics. They explore predictive models for disease onset, progression, and treatment outcomes. The study highlights the potential of machine learning in improving patient outcomes through early identification of health risks. Many efforts are done to cope with the explosion of medical data on one hand, and to obtain useful knowledge from it on the other hand. This prompted researchers to apply all the technical innovations like big data analytics, predictive analytics, machine learning and learning algorithms in order to extract useful knowledge and help in making decisions.

Boukenze *et al.* (2016) in their study "Predictive Analytics in Healthcare System Using Data Mining Techniques" present an overview on the evolution of big data in healthcare system explored the role of predictive analytics in healthcare. They opined that machine learning gives us the power to face diseases earlier that threaten the human being; child, young and old people, through the anticipation of cure and helping in decision- making.

Muniasamy *et al.* (2020) used Deep Learning for Predictive Analytics in Healthcare. According to their work Health data predictive analytics is emerging as a transformative tool that can enable more proactive and preventative treatment options.

In the work of Wadhwa and Babber (2020), Predictive Analysis on Diabetes Using Machine Learning Algorithms uses large volume of multimodal patient data to perform correlations between Body Mass Index, Blood Pressure, Glucose levels, Diabetes Pedigree Function and Skin Thickness of people in different age groups with diabetes. The results indicate a strong relationship between Blood Pressure, BMI and Glucose levels of people with diabetes. This presents immense advantages by applying effective artificial intelligence tools.

The Rising Need for Technology in Healthcare

Technology is indeed changing the way patients interact with healthcare and how the healthcare system understands the patients, making the system more like a consumer market. Bohr and Memarzadeh (2020) puts it this way that with increasing healthcare spending, a rise in a health-conscious population has also been observed, which can provide opportunities for better health management and reduced healthcare spending. This includes consumer goods and services to increase health and wellness such as healthy nutrition, fitness, and meditation retreats. Other increasing trends include wearable and mobile health technologies.

Several numbers of changes and trends are observed in the healthcare industry over the recent years to comply with the changing environment and the rapid technological developments. The rise of electronic medical records (EMRs), personalized genomics, lifestyle and health data, and the capacity for better and faster analysis of data, digital trends are profoundly changing the healthcare system. Many digital stakeholders are seeking to disrupt the healthcare system by taking a technological approach to healthcare data while companies target better patient outcomes by harnessing analytics, machine learning, and other digital tools. According to Bohr et, al (2020) google is building systems biology programs and analytical tools and applying these in areas such as digital pathology while companies like Garmin, Fitbit (Google), and Apple are using information including heart rate and sleep data from their smart watches to predict an overall state of health of an individual. Moreso digitization of medical records, the EMRs provides a systematized, digital collection of patient health information, which can be shared across healthcare settings. This includes notes and information collected by the clinicians in the office, clinic or hospital and contains the patient's medical history, diagnoses, and treatment plans

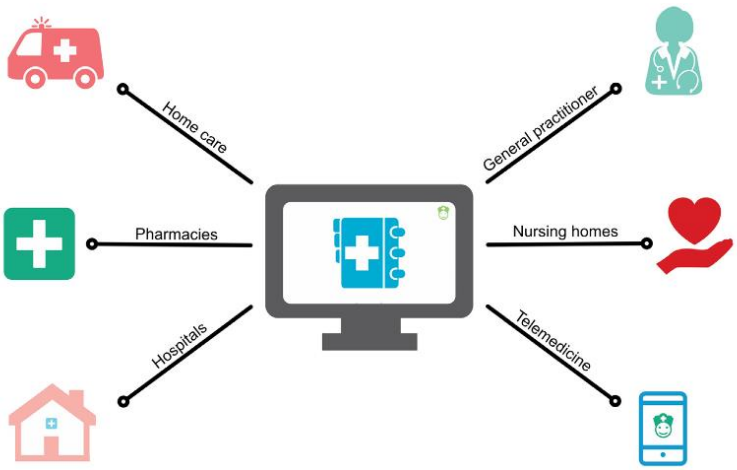


Figure 1: *Electronic medical records can be acquired from multiple sources (Bohr and Memarzadeh, (2020).*

Each of these sources provides a significant amount of data that could immensely improve the healthcare system overall. EMRs can make this process faster, more accurate, and precise. These substantial amount of healthcare data accumulated over the course of each patient life can potentially be used to obtain a better understanding of medical conditions, diagnoses, and treatments.

Fundamental Concepts of Disease Detection with Predictive Models

Disease detection has witnessed a radical transformation in the era of data-driven healthcare. Fundamental to this revolution are machine learning powered predictive models. These models have revolutionized our approach to early disease detection by providing insights and forecasts that were previously inconceivable. To comprehend their significance, it is necessary to venture into the fundamental concepts of predictive models and comprehend how they are altering the healthcare landscape. Rasool *et al.* (2023) put it this way that a predictive model is, at its core, a mathematical representation that uses input data to generate an output, typically in the form of a prediction or classification. The ability to identify patterns and relationships within data forms the basis of predictive models.

In the context of disease detection, predictive models are intended to use a variety of factors, spanning from patient characteristics to environmental variables, to predict the likelihood of a disease's presence, progression, or response to treatment. This predictive capability has the potential to transform healthcare by facilitating proactive interventions and individualized treatments (Liu *et al.*, 2021).

Owing to the intricate and multidimensional nature of medical data, discerning patterns may prove elusive for human observers. Yet, machine learning algorithms excel in processing and scrutinizing this data, revealing intricate correlations that could function as early indicators of disease. By uncovering these patterns, predictive models play a pivotal role in promptly and accurately detecting diseases, enabling medical professionals to intervene proactively before conditions worsen. Regression models are specifically designed to forecast continuous numerical values. In the realm of disease detection, these models may project parameters like blood glucose levels or tumor size based on a range of input variables. Common methods for modeling relationships between variables and outcomes encompass linear regression and logistic regression (Leong *et al.*, 2022). According to Zitnik *et al.*, (2019) classification models designate data points to categories or classes that have been predefined. Using input features such as symptoms, test results, and medical history, these models are used in

healthcare to classify patients as either having a particular disease or being healthy. Decision trees, support vector machines, and random forests are prevalent classification algorithms.

It is important to note that irrespective of enormous potential of predictive models, its effectiveness hinges on the quality and quantity of the training data they receive. Rasool, *et al*., (2023) asserted that is essential to use high-quality, representative datasets to ensure that models can generalize well to new, unseen data. In addition, data privacy and ethical considerations must be given top priority, particularly as patient information becomes the foundation of model training.

Predictive models powered by machine learning are reshaping the early disease detection landscape. These models provide invaluable insights into the presence of disease, its progression, and response to treatment by revealing intricate patterns within complex datasets. As we continue to embrace the data-driven future of healthcare, medical practitioners and data scientists must grasp the fundamental concepts of predictive models. Collaboration between these disciplines is necessary to realize the full potential of predictive models, which will ultimately result in better patient outcomes and more proactive healthcare strategies (DeGregory *et al.*, 2018).

AI and Decision Making in Health Systems

Effective management of health systems, like the provision of public health or health care, is in essence a lattice of information processing tasks (Panch *et al.*, 2018)). Policymakers adapt the functioning of healthcare systems, including organization and governance, financing, and resource management, with the aim of achieving desired healthcare system outcomes. The provision of healthcare encompasses two essential information processing tasks. The first task is screening and diagnosis, which involves classifying cases based on their history, examination, and investigations. The second task is treatment and monitoring, which entails planning, implementing, and monitoring a multistep process to achieve a desired outcome in the future.

Across the domains of health system management and care provision, these processes follow a fundamental structure of hypothesis generation, hypothesis testing, and action. Machine learning has the potential to enhance hypothesis generation and testing within a health system by uncovering hidden trends in data. As a result, it can have a significant impact on both individual patient outcomes and the overall healthcare system. Beam and Kohane (2018) in their study asserted that Machine learning builds upon traditional statistical techniques by employing methods that do not rely on preconceived assumptions about data distribution. It has the ability to uncover patterns within the data, which can then be used to formulate hypotheses and hypothesis tests.

Impacts of Using Machine Learning in Healthcare

Using machine learning in healthcare can have several potential impacts. Here are two examples:

i. Improved diagnostic accuracy: Machine learning algorithms can analyze vast amounts of patient data, including medical records, imaging results, and genetic information, to assist in diagnosing diseases. By identifying patterns and correlations in the data, these algorithms can provide more accurate and timely diagnoses, potentially leading to earlier interventions and improved patient outcomes.
ii. Personalized treatment plans: Machine learning algorithms can analyze patient data, such as medical history, genetic profiles, and treatment outcomes, to develop personalized treatment plans. By considering individual patient characteristics and predicting treatment responses, these algorithms can help healthcare providers tailor

interventions to each patient's specific needs, potentially leading to more effective and efficient treatments.

iii. Predictive analytics for proactive care: Machine learning algorithms can analyze real-time patient data, such as vital signs, wearable device data, and electronic health records, to predict potential health risks or complications. This enables healthcare providers to intervene proactively, offering timely interventions and preventive measures to improve patient outcomes and reduce hospital readmissions.

iv. Streamlined administrative processes: Machine learning algorithms can automate administrative tasks, such as coding and billing, appointment scheduling, and resource allocation. By reducing manual work and improving efficiency, healthcare organizations can allocate more time and resources to patient care, ultimately enhancing the overall healthcare experience. It's important to note that while machine learning has the potential to bring significant benefits to healthcare, it should always be used in conjunction with clinical expertise and ethical considerations. Collaboration between healthcare professionals, data scientists, and IT experts is crucial to ensure the responsible and effective implementation of machine learning in healthcare settings.

Machine Learning's Role in Predictions

According to Nithya and Ilango (2017) Machine Learning is a division of artificial intelligence that practices a variety of statistical, probabilistic and optimization techniques that allows computers to learn from prior examples and to detect hard-to-discern patterns from huge, noisy or complex data sets. Here a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

It is a method of data analysis that automates analytical model building. Through procedures that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed.

Machine learning methodologies are intended to find out the chance to optimize the decisions, based on the predictive value of large-scale data sets. Machine learning task/processes can be broken into the following.

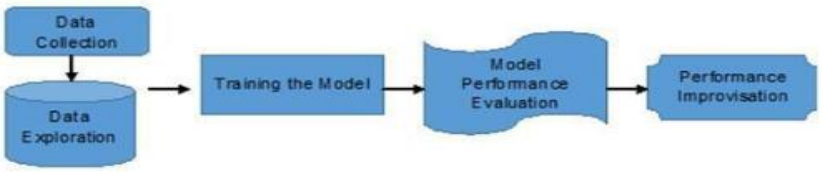


Figure 2: *Machine Learning Process (Nithya and Ilango, 2017)*

i. *Collecting data:* Whether the data is written on paper, recorded in text files and spreadsheets, or stored in an SQL database, the data need to be gathered in an electronic format suitable for analysis. This data will serve as the learning material an algorithm uses to generate actionable information.

ii. *Exploring and preparing the data:* The quality of any machine learning project is based largely on the quality of data it uses. This stage in the machine learning process tends to require a great deal of human intervention. An often-cited statistic suggests that 80 percent of the effort in machine learning is dedicated to data. Much of this time is spent learning more about the data and its nuances throughout a practice called data exploration.

iii. *Training a model on the data:* The specific machine learning task will inform the selection of an appropriate algorithm, and the algorithm will represent the data in the form of a model.
iv. *Evaluating model performance:* It is very important to estimate how well the algorithm learned from its past experience, since each machine learning model results in a biased solution to the learning problem. The accuracy of the model can be evaluated using a test dataset, depending on the type of model used.
v. *Improving model performance:* It is necessary to utilize the advanced strategies to augment the performance of the model, if better performance is needed. Every now and then, it may be required to change to a different type of model overall.

After these steps have been accomplished, if the model appears to be performing acceptably, it can be deployed for its intended task. The model may be applied to provide score data for predictions, for projections of financial data, to generate suitable insight for marketing or research, or to automate tasks. The successes and failures of the deployed model might even provide additional data to train the next generation model.

Materials and Methods

The methodology adopted for the development this study is the Agile Methodology. Recognized for its customer-centric, iterative progress, giving room for changes all through the development process.

The technique adopted for this project is the **Stack Ensemble Technique of Analysis**. Stacking is an ensemble learning technique that combines the predictions of multiple base models to improve the overall performance of the model. It works by training a meta-model on the predictions of the base models, rather than on the original data. This allows the meta-model to learn how to best combine the predictions of the base models to make more accurate predictions.

Actually, Ensemble techniques are the methods that use multiple learning algorithms or models to produce one optimal predictive model. The model produced has better performance than the base learners taken alone. Other applications of ensemble learning also include selecting the important features, data fusion, etc. Ensemble techniques can be primarily classified into **Bagging**, **Boosting**, and **Stacking** (Yash Kandelwal, 2021).

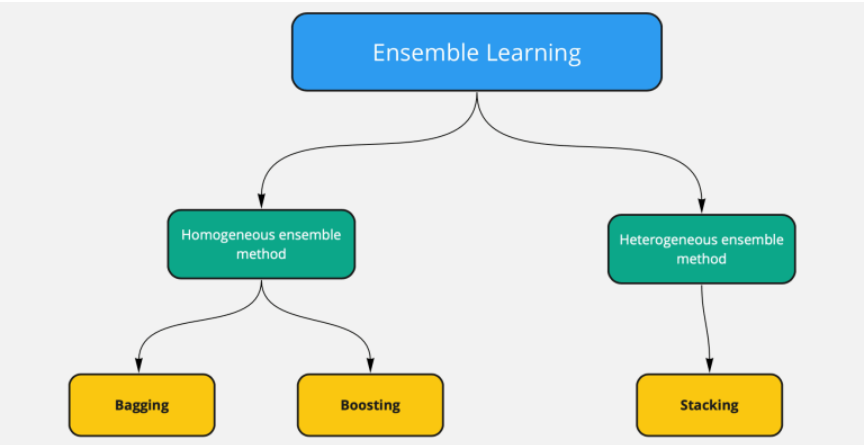


***Figure 3**: Ensemble Technique (Sharma, 2021).*

Bagging, also known as bootstrap aggregating, is an ensemble learning technique that aims to improve the accuracy and stability of machine learning models. It works by creating multiple subsets of the training data, training a base model on each subset, and then combining the

predictions of the base models to make a final prediction. The key idea behind bagging is that by training multiple models on different subsets of the data, we can reduce the variance of the overall model. This is because each base model is trained on a slightly different dataset, so it is less likely to overfit to the training data.

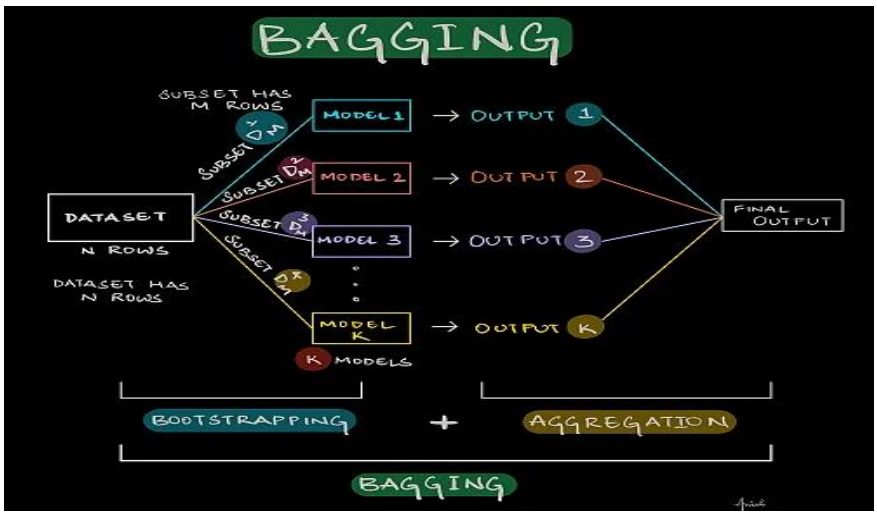


***Figure 4**: Bagging method in ensemble learning (Shah, 2022).*

Boosting is an ensemble learning technique that reduces the bias of a machine learning model by training multiple models sequentially. Each model is trained to correct the errors of the previous model. This means that each model is focused on learning the parts of the training data that the previous models were unable to learn correctly.

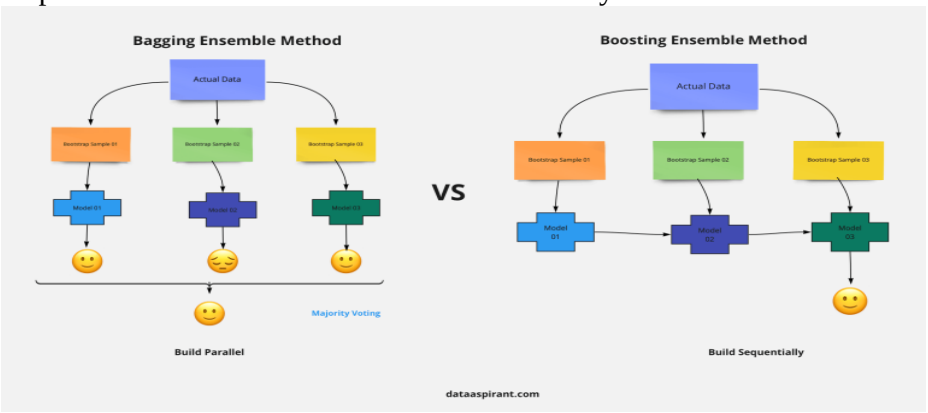


***Figure 5**: Diagrammatical difference between Bagging and Boosting (Polamuri, 2023).*

An example of the stack ensemble technique is in the field of fraud detection. Fraud detection is a challenging task because fraudsters are constantly developing new methods to defraud businesses and individuals. Stacking can be used to combine the predictions of multiple fraud detection models to improve the overall detection rate. For example, a stack ensemble for fraud detection could include the following base models: a logistic regression model trained on historical fraud data, a decision tree model trained on transaction data, a random forest model trained on customer data.

The predictions of these base models would then be stacked into a meta-model, such as a logistic regression model, to make the final prediction. Stacking is a powerful ensemble learning technique that can be used to improve the accuracy and robustness of machine

learning models. It is a relatively complex technique, but it is often worth the effort to implement, especially for problems where high accuracy is required.

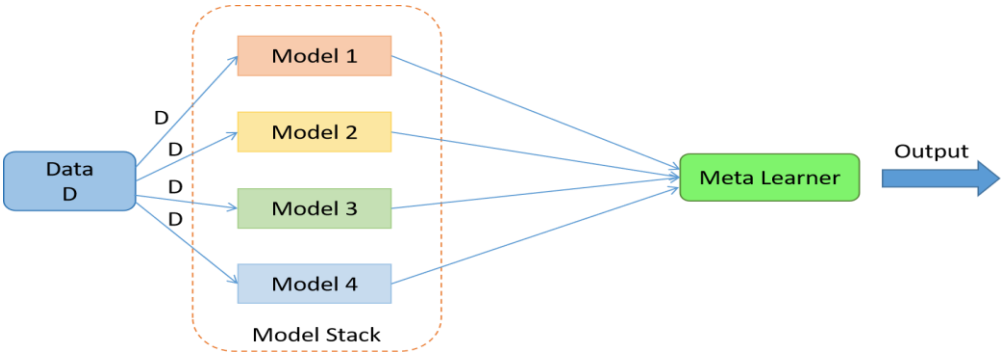


***Figure 6**: Stack ensemble method (Verma, 2023).*

The steps involved in the stack ensemble technique are as follows:

i. Split the training data into folds. This is typically done using a cross-validation technique, such as k-fold cross-validation.
ii. Train a set of base models on each fold of the training data. The base models can be any type of machine learning model, but it is often recommended to use heterogeneous models, meaning that the base models are of different types.
iii. Generate predictions from the base models on the training data. This will create a new training dataset, where each data point contains the predictions of the base models as features.
iv. Train a meta-model on the new training dataset. The meta-model is responsible for combining the predictions of the base models to make the final prediction.
v. Use the meta-model to make predictions on the test data.

Stacking can combine the predictions of any type of machine learning model, including heterogeneous models (models that are of different types). This is important for mitigating and preventing gender disparity in academic achievement because there is no single model that can perfectly predict academic success. By combining the predictions of multiple models, stacking can reduce the overall error and improve the accuracy of the predictions.

Stacking can help to reduce bias in machine learning models. This is important for mitigating and preventing gender disparity in academic achievement because machine learning models can be biased against certain groups of people, such as girls and women. Stacking can help to reduce bias by combining the predictions of multiple models that are trained on different data sets and using different algorithms.

Stacking can make machine learning models more interpretable. This is important for mitigating and preventing gender disparity in academic achievement because it allows us to understand how the model is making predictions and to identify any potential biases. Stacking models are more interpretable than boosting and bagging models because they use a meta-model to combine the predictions of the base models. The meta-model can be interpreted to understand how the different factors influence the final prediction. The above reasons are why “Stacking” was chosen as the best ensemble technique for this project.

Dataset used for the Analysis

The dataset utilized in this research was sourced from Kaggle.com and is titled "diabetes_data.csv". It comprises a set of variables (columns) that play a crucial role in the analysis and development of the diagnostic and recommendation system for hypertension.

Each variable provides valuable insights into different aspects of patient health and cardiovascular characteristics. The dataset was contributed by Kaggle user " PRASAD SHINGARE" and can be accessed at (https://www.kaggle.com/code/prasadshingare/diabetes-hypertension-and-stroke-prediction/input?select=hypertension_data.csv). The dataset is called the Diabetes prediction by Prasad Shingare, a Kaggle expert in India.

This file contains the following basic information: (Age, Sex, HighChol, CholCheck, BMI, Smoker, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, GenHlth, MentHlth, PhysHlth, DiffWalk, Stroke, HighBP, Diabetes) about the patients. Here is an overview of the key variables included in the dataset:

i. **Age**: This column represents the patient's age in years.
ii. **Sex**: This column indicates the patient's gender (coded in binary).
iii. **HighChol**: This column indicates whether the patient has high cholesterol (coded as 1/0).
iv. **CholCheck**: This column might represent how often the patient gets their cholesterol checked (e.g., annually, biannually).
v. **BMI**: This stands for Body Mass Index, a measure of body fat based on height and weight.
vi. **Smoker**: This column likely indicates whether the patient is a smoker (coded as 1/0).
vii. **Heart Disease or Attack**: This column indicates whether the patient has a history of heart disease or heart attack (coded as 1/0).
viii. **PhysActivity**: This column represents the patient's level of physical activity.
ix. **Fruits**: This column represent the patient's daily or weekly fruit intake.
x. **Veggies**: This column whether the patient eats vegetables.
xi. **HvyAlcohol Consump**: This column indicates whether the patient has heavy alcohol consumption (possibly coded as 1/0).
xii. **GenHlth**: This is an abbreviation for General Health, potentially a rating of the patient's overall health.
xiii. **MentHlth**: This is an abbreviation for Mental Health, potentially a rating of the patient's mental well-being.
xiv. **PhysHlth**: This is an abbreviation for Physical Health, potentially a rating of the patient's physical health.
xv. DiffWalk: The meaning of this column is unclear without further context. It indicates whether the patient finds it difficult walking.
xvi. **Stroke**: This column indicates whether the patient has a history of stroke (coded as 1/0).
xvii. **HighBP**: This is an abbreviation for High Blood Pressure, indicating whether the patient has high blood pressure (coded as 1/0).
xviii. **Diabetes**: This column is the target variable, indicating whether the patient has diabetes (coded as 1/0).

Understanding and analyzing these variables collectively contribute to the creation of a robust diagnostic and recommendation system, enabling a comprehensive approach to hypertension management and patient care.
The dataset is formatted as a CSV file and can be downloaded from Kaggle. It includes a readme file that provides more information about the data collection process and the format of the data. To analyze the data, Pandas' read_csv() function was first used to import the relevant information from the provided CSV file. Leveraging the power of Pandas, it was possible to efficiently import the dataset stored in a CSV file using the read_csv() method.

	A	B	C	D	E	F	G	H	I	J
1	Age	Sex	HighChol	CholChecl	BMI	Smoker	HeartDise	PhysActiv	Fruits	Veggies
2	4	1	0	1	26	0	0	1	0	1
3	12	1	1	1	26	1	0	0	1	0
4	13	1	0	1	26	0	0	1	1	1
5	11	1	1	1	28	1	0	1	1	1
6	8	0	0	1	29	1	0	1	1	1
7	1	0	0	1	18	0	0	1	1	1
8	13	1	1	1	26	1	0	1	1	1
9	6	1	0	1	31	1	0	0	1	1
10	3	0	0	1	32	0	0	1	1	1
11	6	1	0	1	27	1	0	0	1	1
12	12	0	1	1	24	1	1	1	1	1
13	4	1	0	1	21	0	0	1	1	1
14	7	1	1	1	27	0	0	1	1	1
15	10	1	0	1	58	0	0	0	1	1
16	10	0	1	1	29	1	0	1	1	0
17	10	0	0	1	18	1	0	1	1	0

***Figure 7**: A section of the dataset used for the analysis.*

The dataset is balanced. This means that the different in size of the negative cases and the positive cases is too little to cause significant bias in the study.

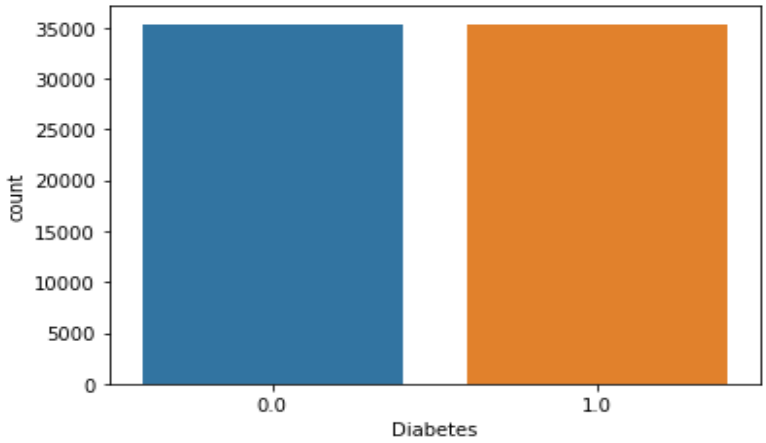


***Figure 8**: Bar showing the balanced distribution of the dataset.*

The dataset underwent feature scaling using the **StandardScaler** module from the **sklearn.preprocessing** library. This preprocessing step standardized the numerical features, aligning them with a mean of 0 and a standard deviation of 1. The utilization of **StandardScaler** aimed to normalize feature values, addressing varying scales and optimizing the dataset for subsequent stages of the project. The figure below illustrates this.

	0	1	2	3	4
0	-0.172579	1.00029	-0.934922	-0.0925932	-2.22607
1	0.0250976	1.00029	-0.934922	0.0206979	0.00751971
2	0.552235	-0.999712	-0.934922	-0.545758	0.758469
3	-0.699716	-0.999712	-0.934922	-0.659049	-1.12853
4	-1.02918	-0.999712	-0.934922	-1.56538	-0.743429
5	1.93597	-0.999712	0.0432152	-0.432467	0.277091
6	0.815804	1.00029	-0.934922	0.0206979	0.00751971
7	-0.238471	-0.999712	-0.934922	-0.659049	-0.935981
8	0.684019	-0.999712	1.02135	-1.2255	-1.37885
9	-0.0407946	1.00029	-0.934922	-0.659049	-1.12853

***Figure 9**: A section of the output of the StandardScaler on the dataset*

Algorithm

In pursuit of revolutionizing hypertension management, this project leverages advanced algorithms to harness insights from a comprehensive dataset. These algorithms form the backbone of a sophisticated diagnostic and recommendation system, aiming to provide personalized and effective healthcare solutions. Let's explore the key algorithms integral to the success of this innovative approach.

The Correlation Matrix

In data analytics, the corr() method is a fundamental tool employed to calculate the correlation matrix, revealing the relationships between variables in a dataset. Correlation measures the statistical association between two or more variables, indicating the degree to which changes in one variable correspond to changes in another. The corr() method is typically applied to a DataFrame in Python, often using libraries such as Pandas. The corr() method operates on a Pandas DataFrame, with each column representing a different variable. It computes pairwise correlation coefficients for all variable combinations in the DataFrame. The empty cells indicate very insignificant correlation values.

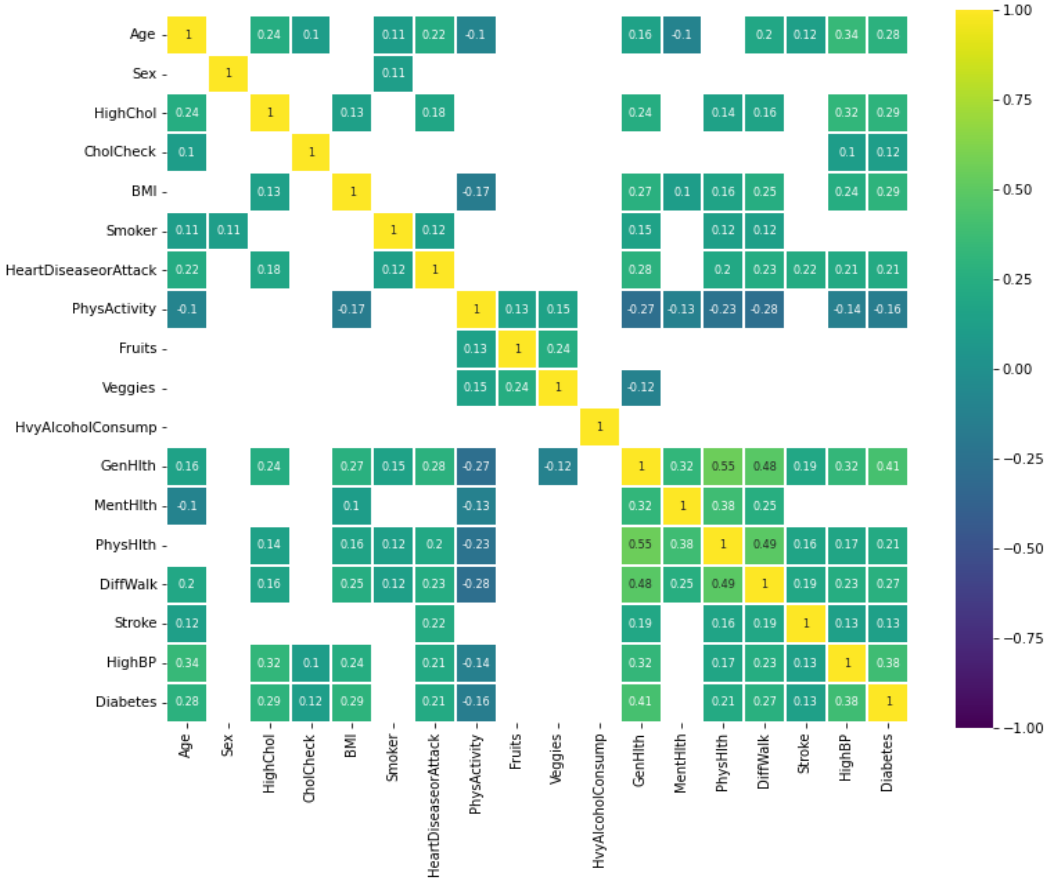


***Figure 10**: The correlation among the health variables under study.*

By applying the corr() method, data analysts can gain valuable insights into the interdependencies among variables, aiding in feature selection, identifying multicollinearity, and informing subsequent analyses in various data science and machine learning tasks.

Confusion Matrix

A confusion matrix is a fundamental tool in evaluating the performance of a classification model. It provides a comprehensive and detailed summary of the model's predictions, breaking down the outcomes into four categories: True Positive (TP), True Negative (TN), False

Positive (FP), and False Negative (FN). These components are crucial for assessing the effectiveness of a model across various metrics. The following are the components of the Confusion Matrix:

i. **True Positive (TP):** Instances where the model correctly predicts the positive class. For example, correctly identifying patients with hypertension.
ii. **True Negative (TN):** Instances where the model correctly predicts the negative class. For example, correctly identifying patients without hypertension.
iii. **False Positive (FP)**: Instances where the model predicts the positive class incorrectly. Also known as a Type I error. For example, predicting hypertension in a patient who does not have it.
iv. **False Negative (FN):** Instances where the model predicts the negative class incorrectly. Also known as a Type II error. For example, failing to predict hypertension in a patient who actually has it.

Results and Discussion

Logistic Regression

Logistic Regression is a statistical method commonly used in data analytics and machine learning for binary classification problems. Despite its name, it is employed for predicting the probability of an instance belonging to a particular category rather than predicting a continuous outcome. Logistic Regression is well-suited for problems where the dependent variable is binary, meaning it has only two possible outcomes (e.g., 0 or 1, True or False). It applies the sigmoid (logistic) function to transform a linear combination of input features into a value between 0 and 1. This transformed value represents the probability of belonging to the positive class.

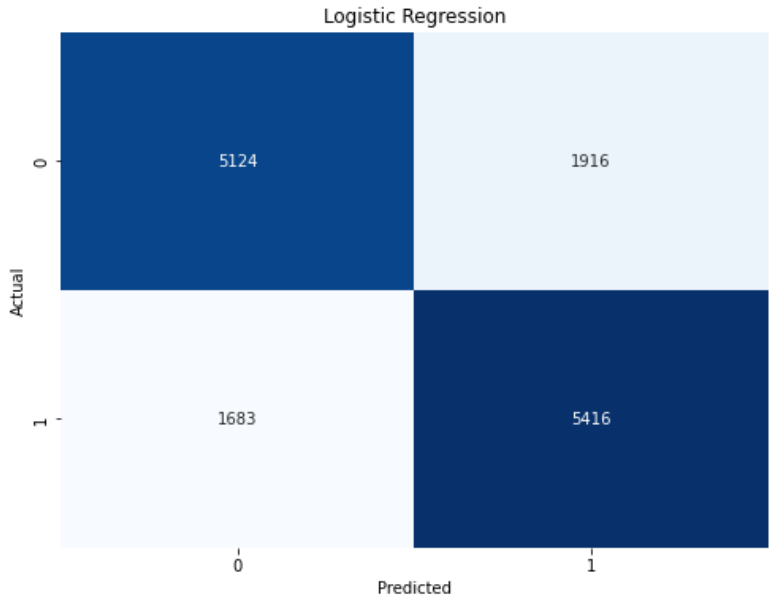


Figure 11*: Confusion matrix showing the performance of the Logistic Regression model on the test dataset.*

Logistic Regression establishes a decision boundary based on the calculated probabilities. For binary classification, a common threshold is set at 0.5. If the predicted probability is above 0.5, the instance is classified as the positive class; otherwise, it belongs to the negative class. From the above Confusion matrix evaluation of the Logistic Regression model, the accuracy of 74.54%, Precision score of 74.56%, recall score of 74.54% and F1 score value of 74.53% was achieved.

Decision Tree

A Decision Tree is a powerful and interpretable machine learning algorithm used for both classification and regression tasks. It is a tree-like model where each internal node represents a decision based on a particular feature, each branch represents the outcome of the decision, and each leaf node represents the final decision or the predicted output. Decision Trees are widely used due to their simplicity, ability to handle both numerical and categorical data, and their capability to capture complex relationships in the data. Decision Trees have a hierarchical structure with nodes representing decisions or tests based on specific features. The tree structure flows from the root node to the leaf nodes, where the final decisions or predictions are made. Decision Trees employ various splitting criteria, such as Gini impurity or mean squared error, to determine how well a particular feature separates the data into distinct classes or groups. From the Confusion matrix evaluation of the Decision Tree model given below, the accuracy of 66.61%, Precision score of 66.63%, recall score of 66.61% and F1 score value of 66.6% was achieved.

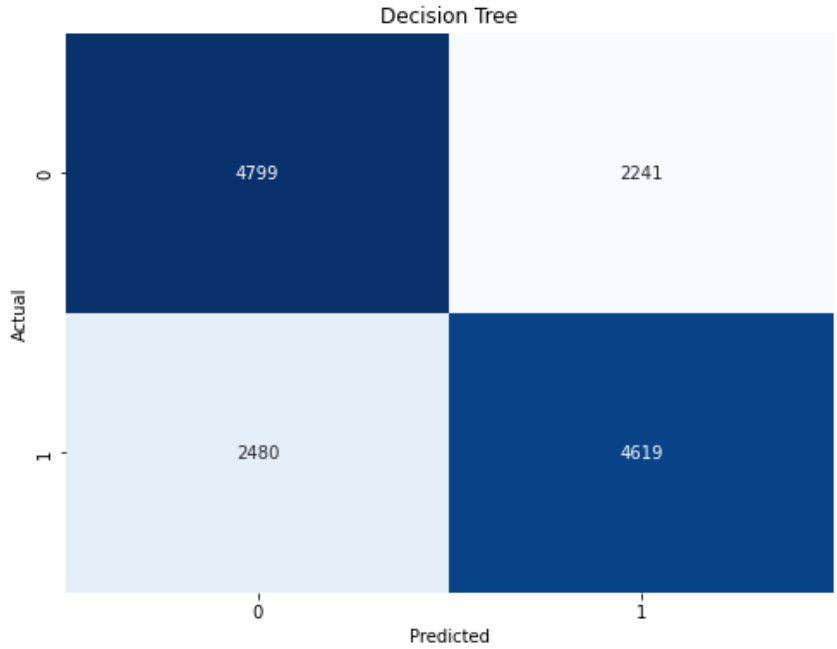


***Figure 12**: Confusion Matrix showing the good performance of the Decision Tree model*

K Nearest Neighbor

K-Nearest Neighbors (KNN) is a versatile and straightforward algorithm used for both classification and regression tasks in data analytics and machine learning. It is a non-parametric, instance-based learning algorithm, meaning it doesn't make assumptions about the underlying data distribution and makes predictions based on the similarity of new instances to existing data points. It makes predictions based on the majority class or average value of the k-nearest data points to a given instance. The value of k is a user-defined parameter. KNN relies on a distance metric (commonly Euclidean distance) to measure the similarity between instances. The algorithm identifies the k-nearest neighbors by finding the data points with the smallest distances to the target instance. KNN can be computationally expensive, especially with large datasets, as it requires calculating distances for each prediction. From the Confusion matrix evaluation of the KNN model, the accuracy of 71.19%, Precision score of 71.24%, recall score of 71.19% and F1 score value of 71.17% was achieved.

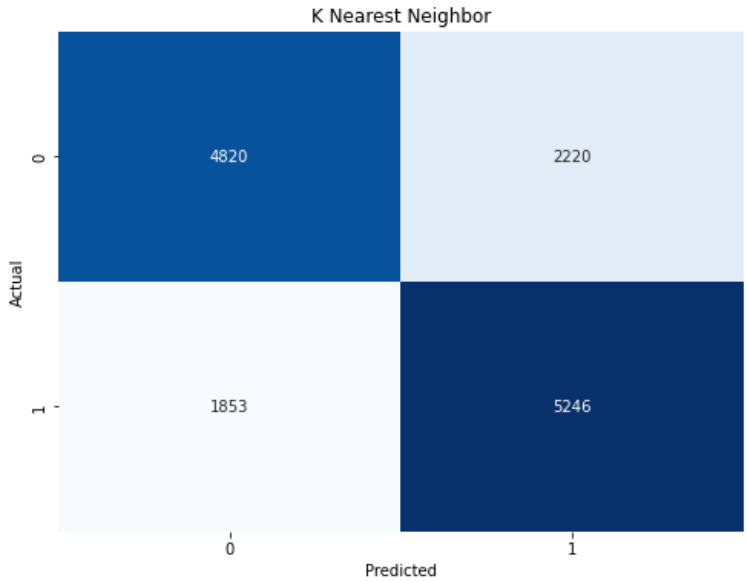


***Figure 13**: Confusion matrix showing the performance of the KNN model.*

Random Forest

For this study the Random Forest model was used as the META model of the stack ensembled system, which comprises of the three aforementioned algorithms: Logistic Regression, K Nearest Neighbor and Decision Tree. Random Forest is a potent ensemble learning algorithm applicable to both classification and regression tasks. It constructs multiple decision trees during training and combines their predictions, providing a robust and versatile solution in data analytics and machine learning. Random Forest builds multiple decision trees and aggregates their predictions to enhance accuracy and stability. From the Confusion matrix evaluation of the Random Forest model given below, the accuracy of 79.76%, Precision score of 79.85%, recall score of 79.76% and F1 score value of 79.75% was achieved.

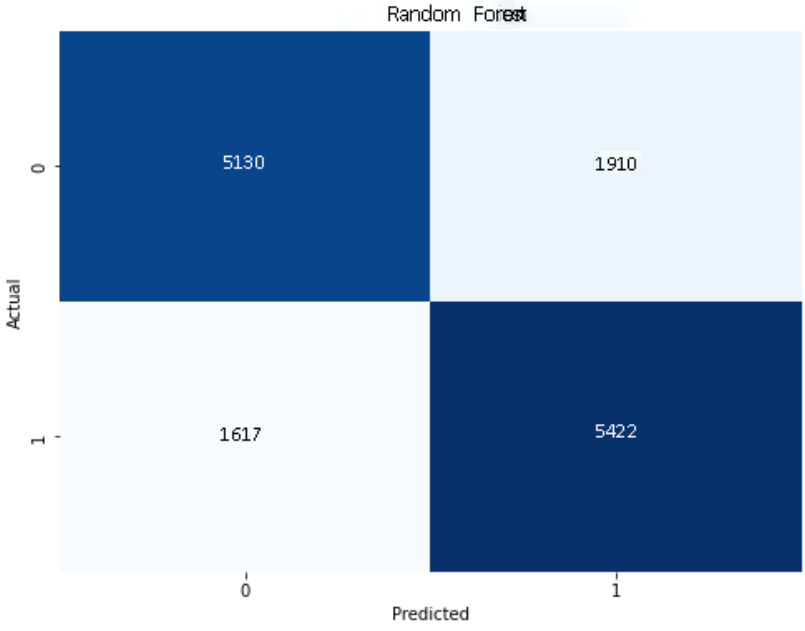


***Figure 14**: Confusion matrix showing the performance of the Random Forest Model.*

Evaluation Metrics

As we navigate the landscape of model evaluation in data analytics and machine learning, it becomes crucial to employ metrics that provide nuanced insights into the performance of our

models. Among these, the Receiver Operating Characteristic (ROC) curve and Precision-Recall (PR) curve stand as invaluable tools. These metrics go beyond simple accuracy, offering a deeper understanding of a model's ability to discriminate between classes and balance precision and recall. In the upcoming discussion, we delve into the significance and interpretation of the ROC curve and Precision-Recall curve, unraveling their role in assessing the efficacy and reliability of our predictive models.

Receiver Operating Characteristic (ROC) curve

The Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the trade-off between true positive rate (sensitivity) and false positive rate across various classification thresholds. A perfect ROC curve would be one where the true positive rate is always 1 (100%), and the false positive rate is always 0 (0%). A perfect ROC curve starts from the bottom-left corner and ascends vertically to the top-left corner. This indicates that the model is achieving a true positive rate of 1 without incurring any false positives just like the one we have in Figure 4.8.

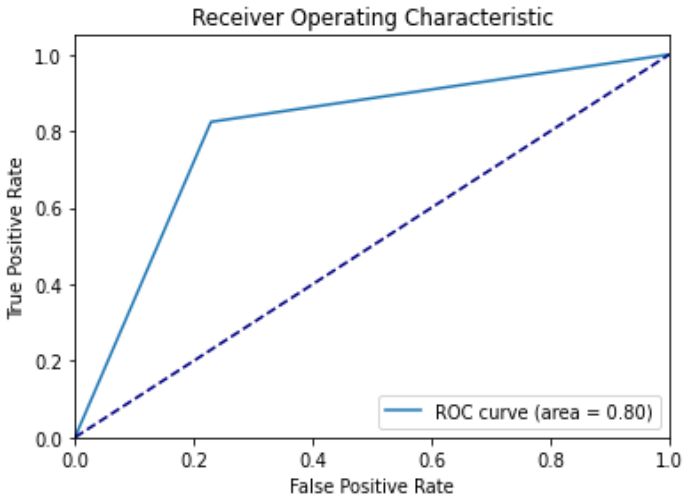


Figure 15*: Figure showing the ROC curve of the meta model (Random Forest model).*

Precision-Recall (PR) curve

The Precision-Recall (PR) curve is a valuable metric for evaluating classification models, particularly in scenarios where class imbalances exist. It showcases the trade-off between precision and recall at various decision thresholds. A perfect PR curve indicates a model with flawless precision and recall across all thresholds. A perfect PR curve starts from the bottom-right corner and ascends vertically to the top-right corner. This signifies that the model achieves both perfect precision and recall. At the top-right corner, precision is maximized, indicating that every positive prediction made by the model is indeed correct. Simultaneously, recall is maximized, implying that the model captures every positive instance in the dataset.

The area under the Precision-Recall curve (AUC-PR) quantifies the overall performance of the model. In the case of a perfect curve, the AUC-PR value would be 1, reflecting flawless precision and recall trade-offs. Similar to the ROC curve interpretation, a random classifier would follow the 45-degree diagonal line (the line of no-discrimination) from the bottom-right to the top-left. A perfect model deviates significantly from this diagonal, reaching the top-right corner.

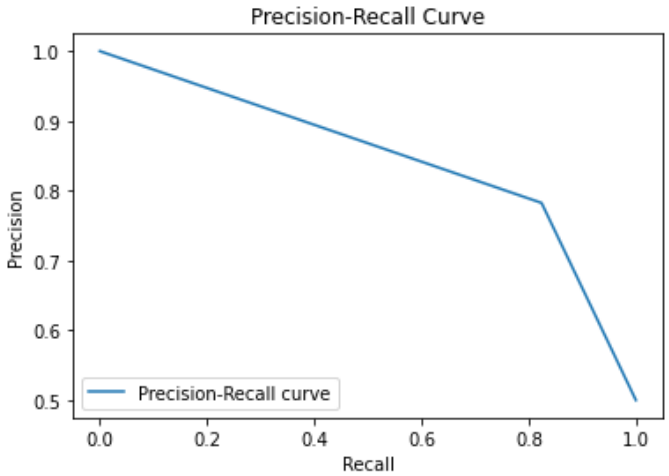


Figure 16: Figure showing the Precision-Recall Curve of the Random Forest model

Conclusion

Predictive analytics in healthcare has change the way of how medical researchers and practitioners gain insights from medical data and take decisions. In this study, the algorithms that formed the base models are: Logistic Regression, K Nearest Neighbor and Decision Tree. While the Random Forest classifier was used as the meta model. The meta model achieved an accuracy of 79.76%. The meta models surely out-performs every single base model when used on the test data. This analysis promises to transform patient care and healthcare operations. This advanced technology leverages data-driven insights and algorithms to enhance medical decision-making, accelerate diagnosis, optimize treatment plans, and improve patient outcomes. The figure below shows the summary of the performances of all the models used for this study.



Figure 16: Result summary of the models

Recommendations

Some limitations of this study are the size of dataset and missing attribute values. To build a prediction model for diabetes with 99.99% accuracy, the research recommends the use of thousands of records with zero missing values.

References

Beam, A. L., & Kohane, I. S. (2018, April 3). Big Data and Machine Learning in Health Care. *JAMA, 319*(13), pp. 1317-1318. doi:https:// jama.jamanetwork.com/ article.aspx?doi= 10.1001/ jama.2017.18391&utm_ca10.1001/jama.2017.18391

Bohr, A., & Memarzadeh, K. (2020). Chapter 1 - Current Healthcare, Big Data, and Machine Learning. In A. Bohr, & K. Memarzadeh, *Artificial Intelligence in Healthcare* (pp. 1-24). Academic Press. doi:https://doi.org/10.1016/B978-0-12-818438-7.00001-0.

Boukenze, B., Mousannif, H., & Haqiq, A. (2016). Predictive Analytics in Healthcare System Using Data Mining Techniques. *Computer Science & Information Technology (CS & IT)*, 1-9. doi:10.5121/csit.2016.60501.

DeGregory, K. W., Kuiper, P., DeSilvio, T., Pleuss, J. D., Miller, R., Roginski, J. W., . . . Thomas, D. (2018). A Review of Machine Learning in Obesity. *Obesity Reviews, 19*(5). doi:doi: 10.1111/obr.12667.

Khan, M. (2023). Data Science in Health Informatics: Harnessing Big Data for Healthcare. *OSF Preprints*. Retrieved September 21, 2023

Khandelwal, Y. (2021, August 13). *Home: Intermediate.* Retrieved September 15, 2023, from Analytics Vidhya Web site: https://www.analyticsvidhya.com/blog/2021/08/ensemble-stacking-for-machine-learning-and-deep-learning

Leong, Y. X., Tan, E. X., Leong, S. X., Koh, C. S., Nguyen, L. B., Chen, J. R., . . . Ling, X. Y. (2022). Where Nanosensors Meet Machine Learning: Prospects and Challenges in Detecting Disease X. *ACS Nano, 16*(9), 13279-13293. doi:10.1021/acsnano.2c05731

Liu, J. T., Glaser, A. K., Bera, K., True, L. D., Reder, N. P., Eliceiri, K. W., & Madabhushi, A. (2021). Harnessing Non-Destructive 3D Pathology. *Nature Biomedical Engineering, 5*(3), 203-218. doi:10.1038/s41551-020-00681-x

Muniasamy, A., Tabassam, S., Hussain, M. A., Sultana, H., Muniasamy, V., & Bhatnagar, R. (2020). Deep Learning for Predictive Analytics in Healthcare. In A. E. al. (Ed.), *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019)* (pp. 32-42). Springer. doi:doi.org/10.1007/978-3-030-14118-9_4

Najjar, R. (2023). Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. *Diagnostics, 13*(2760), 1-25. doi:https://doi.org/ 10.3390/ diagnostics13172760.

Nithya, B., & Ilango, V. (2017). Predictive Analytics in Health Care Using Machine Learning Tools and Techniques. *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 492-499). Madurai, INDIA: IEEE. doi:10.1109 /ICCONS.2017.8250771.

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial Intelligence, Machine Learning and Health Systems. *Journal of Global Health, 8*(2), 1-8. doi:doi: 10.7189/jogh.08.020303

Polamuri, S. (2023, October 10). *Dataaspirnant Machine Learning.* Retrieved from A dataaspirant web site: https://dataaspirant.com/bagging-algorithms

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *The New England Journal of Medicine, 380*(14), 1347-1358. doi:10.1056/NEJMra1814259.

Rasool, S., Husnain, A., Saeed, A., Gill, A. Y., & Hussain, H. K. (2023). Harnessing Predictive Power: Exploring the Crucial Role of Machine Learning in Early Disease Detection. JURIHUM. *JURIHUM : Jurnal Inovasi dan Humaniora, 1*(2), 302-315.

Shah, A. (2022, July 20). *A Medium Website.* Retrieved 23 September, 2023, from https://medium.com/@jwbtmf/decision-tree-and-ensemble-learning-algorithms-in-machine-learning-ea27b4429d85

Sharma, S. (2021, August 24). Retrieved October 10, 2023, from Medium Website: https://sid-sharma1990.medium.com/ensemble-techniques-in-machine-learning-23f6a55faa17

Shingare, P. (2022, December 30). *Kaggle.* Retrieved September 22, 2023, from A Kaggle Web site: https://www.kaggle.com/code/prasadshingare/diabetes-hypertension-and-stroke-prediction/input?select=hypertension_data.csv.

Verma, A. (2023, November 28). *Medium.* Retrieved from Medium Website: https://medium.com/@ajayverma23/mastering-complexity-the-comprehensive-guide-to-stacking-ensemble-models-7c0ef4876eda.

Wadhwa, S., & Babber, K. (2020). Artificial Intelligence in Health Care: Predictive Analysis on Diabetes Using Machine Learning Algorithms. *20th International Conference on Computational Science and Its Applicaions- ICCSA* (pp. 354-366). Cagliari: Springer.

Zitnika, M., Nguyenb, F., Bo, W., Leskoveca, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities. *Information Fusion, 50*, 71-91. doi:10.1016/j.inffus.2018.09.012