National College of Ireland

# A Predictive Model for Predicting Blood Pressure Levels Using Machine Learning Techniques

MSc Research Project
Data Analytics


Akinwale Sunday Obafemi
Student ID: 20200854

School of Computing
National College of Ireland


Supervisor: Jorge Basilio

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

National College of Ireland

| | |
|---|---|
| **Student Name:** | Akinwale Sunday Obafemi |
| **Student ID:** | 20200854 |
| **Programme:** | Data Analytics |
| **Year:** | 2022 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Jorge Basilio |
| **Submission Due Date:** | 15/08/2022 |
| **Project Title:** | A Predictive Model for Predicting Blood Pressure Levels Using Machine Learning Techniques |
| **Word Count:** | 7108 |
| **Page Count:** | 22 |

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:** | |
| **Date:** | 15th September 2022 |

## 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): | |

# A Predictive Model for Predicting Blood Pressure Levels Using Machine Learning Techniques

Akinwale Sunday Obafemi
20200854

**Abstract**

This study examined an efficient model for the accurate prediction of blood pressure levels by building on existing models that are useful for HBP prediction. The best model was selected based on RMSE and MAE evaluation metrics that were used to evaluate their performance. The RMSE metric was used to provide information by doing a term-by-term comparison and showing the value-performance relationship between selected models; while the MAE metric shows the average of absolute errors that the selected models are liable. The data for this study was sourced from Kaggle.com because it is publicly available and reduces the chance for ethical misconduct. Features of the data selected are useful for training, testing, and estimation of the study outcome. This was done through the processes of pre-cleaning, visualisation, transformation, engineering, and modelling. The CatBoost model outperformed other models with an RMSE score of 7.307459 and an MAE score of 5.790854; this was further confirmed after evaluation by the Grid search presented good evaluation scores where the RMSE and MAE scores were significantly reduced to 0.87773 and 0.12256 respectively. The study also shows that the most significant variable for blood pressure measurement is age while the least significant is the number of major vessels coloured by fluoroscopy (ca_4).

## 1 Introduction

Hypertension is a disease caused by high blood pressure (BP). This disease has grown to become a global menace and affects individuals from different races or backgrounds. It has become a leading cause of death, especially amongst the older generation. That being said, it does not only affect the older people but has also been noticed among the younger generation especially in recent times as over time there has been an increase in pressure levels of various kinds on the younger generation, which include peer pressure and the likes. All these sorts of pressures can directly or indirectly affect health. The human heart is tasked with supplying blood to the organs and tissues, so with every heartbeat, blood is being pumped around the body via the circulatory system and this puts pressure on the vessel walls. A sharp or drastic turn of events can lead to an increase in the rate of pumping which can then in turn lead to high blood pressure. The World Health Organisation (WHO) estimates that approximately 1.13 billion people are living with hypertension across the globe, but less than one in five of those individuals have their condition under control. Because of this, keeping close track of one's blood pressure (BP) levels consistently is necessary for the preservation of one's good health, as doing so will assist in the early detection of any deterioration in one's physical health. In addition,

it would lower the probability that the illness would worsen. Systolic blood pressure, also known as SBP, is the upper limit of the blood pressure range, while diastolic blood pressure, also known as DBP, is the lower limit Zheng and Yu (2021).

## 1.1 Background and Motivation

As high blood pressure serves as a preceding event for many other illnesses, its subtle way of affecting patients makes it important to monitor and find a way of early detection before it becomes a serious illness in the lives of patients. This study explains the common method of measuring blood pressure.

A growing percentage of teenagers and younger children are suffering from hypertension as a result of the sedentary lifestyle that many young people today lead. There is mounting evidence that high blood pressure in childhood is a precursor to high blood pressure in adulthood. Children's hypertension is not detected until it becomes life-threatening or until they reach adulthood, which is a tragedy Katamba et al. (2020). Alcohol use, obesity, a lack of physical activity (or an excessively sedentary lifestyle), and a very low intake of fruits and vegetables combined with a high-fat diet are some risk factors for hypertension that can be discovered to be frequent in our environment. Other elements that are typically seen as unchangeable risk factors include advanced age (over 65), underlying illnesses including diabetes, and a family history of hypertension. The BP can be measured using two readings which are described below:

- **Systolic Blood Pressure (SBP):** This is the first number and the one usually above. This number represents the pressure when the heart contracts to pump blood out to other parts. This number is considered normal when it is below 140mmHg but becomes high when the reading gets higher than that.

- **Diastolic Blood Pressure (DBP):** This reading represents the pressure when the heart muscles relax and get filled with blood. It is usually lower than SBP. It is considered normal when its value is below 90mmHg.

The BP should be measured regularly and should also be measured at different times during the day.

Due to both the long-term health effects of uncontrolled hypertension and its role in identifying many serious medical conditions, the importance of early and accurate diagnosis cannot be stressed in the case of pediatric hypertension. Furthermore, it is vital to treat and prevent hypertension-related cardiovascular issues in childhood and adolescence before any further clinical signs occur. However, it is predicted that there is difficulty in predicting the complications of hypertension from the medical point of view. According to the research conducted by Bard et al. (2019), the invasive approach and the cuff-based method are the most common and accurate techniques for measuring blood pressure. The device is nonetheless exclusive to hospitals and healthcare facilities. The invasive procedure involves implanting a BP sensor into a blood artery or the heart to measure arterial pressure. This technology has a significant drawback in that it is highly painful for patients and should generally be reserved for critically ill individuals. A superior alternative is a cuff-based method. However, the cuff-based method also has its limitation. Some of these include the sensitivity to option, backend application, and its application in the clinics Bard et al. (2019).

According to El-Hajj and Kyriacou (2020), barely one-third of hypertensive individuals have their blood pressure under control due to the dearth of easily available blood

pressure monitoring equipment (2020). This circumstance has necessitated study into the dissemination of BP measurement techniques that are simpler for the general people. This will ensure that there is no need to visit hospitals or schedule a doctor's appointment before measuring blood pressure.

This study aims to establish a more effective method for anticipating swings in blood pressure levels and analyze the elements that influence these changes over time. In recent years, there has been an increase in the collection of clinical data, which has aided in the mapping of patterns for addressing the majority of health issues. These data include age, smoking habit, the quantity of activity, weight or body mass index (BMI), stress level, cholesterol, and family history, among others, as physiological risk factors for a variety of disorders, including hypertension. Some hypertensive patients may exhibit symptoms such as headaches, chest pain, nosebleeds, and dizziness, but the majority do not exhibit any symptoms at all, making it difficult to diagnose and emphasizing the significance of the predictive model, particularly for lower-income countries that lack good or proactive healthcare facilities.

Ji et al. (2021) explained that the advancement of machine learning and artificial intelligence has provided fresh insights into the process of predicting the difficulties that can arise from hypertension. On the one hand, there has been significant development in data storage technology, and as a result, a significant quantity of medical data about individuals who suffer from hypertension has been collected. On the other hand, data mining tools have advanced to a more mature stage, which contributes to the article's research by providing further technological support. At present, there have been successful applications of machine learning in a variety of fields, including disease prediction, medical picture recognition, medical diagnostics, and others. In addition, a large number of specialists and academics have attempted to research the hypertension-related consequences using data mining techniques Lee et al. (2018).

This study is unique in that it aims to aid in the use of machine learning techniques to detect how blood pressure levels change bringing about some key factors which include age, gender and the level of exercise amongst others. It is essential to study how factors such as weight and others are measured, followed by an evaluation of the machine learning models to be employed in this research and the extent to which the models have contributed to the extension of the knowledge base of the research area. In addition, it is vital to describe how the models might be integrated or modified based on their previous applications to improve forecast precision.

Furthermore, the paper will examine the accuracy of these predictions and a distinction would be made between the algorithms investigated in this paper.

## 1.2 Research Question

- **Main RQ:** How well can blood pressure levels be predicted using Machine Learning techniques? This will provide insight into how different machine learning algorithms can be used to predict blood pressure. The best model will then be determined and suggested for modelling HBP.

- **Sub RQ:** How do we detect major factors that both directly or indirectly affect the fluctuations of BP? The prediction will reveal most vital factors that can cause or affect BP. This will help doctors, clinicians and patients determine an accurate course of action as it relates to BP.

## 1.3 Research Objectives

After this research project, a few goals should have been met, including the development of a model capable of accurately predicting blood pressure values from the many models used to forecast HBP. The most efficient will be selected based on their performance using the evaluation metrics: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). The model with the best RMSE and MAE scores when compared to the other models will be the indices of selection of the best model. Beyond selecting the best model, the correlation shown between the physiological factors and the blood pressure level as well as the extent of correlation would be determined. Each of the models will be implemented and evaluated and the results will be extracted. All these will be done to the effect that the results presented will assist in clinical diagnosis, prediction, and finding solutions to a blood pressure reading in a faster and more efficient manner.

The subsequent sections of this work will review related literature and critique the literature, the methodology, and design of this project, the results as well as viable conclusions. The implementation, evaluation, and results of the analysis for the detection of High Blood Pressure will also be assessed. Finally, conclusions and recommendations will be made while making suggestions for further studies.

# 2 Related Work

## 2.1 Introduction

The ultimate goal of this research is to apply machine learning approaches to effectively detect how variations in blood pressure levels are influenced by important elements, such as behavioural and lifestyle habits. It is crucial to evaluate how variables like age, maximum heart rate, and others impact blood pressure levels before evaluating the machine learning models that will be used in this research and how much, overall, the models have contributed to expanding the body of knowledge in the field in which the research is being conducted. To achieve improved accuracy, it is necessary to also describe how the models can be integrated or modified based on the previous usage. More information about this research would be given by the earlier literary works that would be reviewed here.

## 2.2 Blood Pressure Analysis

Monitoring and detecting BP are significant areas of medical research. In recent years, it has received a great deal of attention in the medical industry due to the need to reduce the rising incidence of hypertension. The number of persons diagnosed with cardiovascular diseases (CVDs) continues to rise since a significant proportion of them are either uninformed of the condition or unable to treat it sufficiently with medicines Schultz et al. (2022). Fuchs and Whelton (2020) claim that high blood pressure is a key cause of CVDs, based on clinical tests conducted in the research. Additionally, high blood pressure may result in disorders such as atrial fibrillation, aortic syndromes, and coronary artery disease. Continuous monitoring of blood pressure (BP) is an effective method for managing hypertension. Blood pressure is a changeable characteristic that varies with heart rate. In the 20th century, insurance data were utilized to demonstrate direct links between high blood pressure and death rate. From these data, it was also

determined that a person's blood pressure and weight increase proportionally with his age Zhou et al. (2021). Therefore, it can be altered by elements such as environmental noise, temperature, physical and mental activity, etc. The precision of blood pressure measurement is essential for diagnosing hypertension. Inaccuracy might result in incorrect diagnoses and unnecessary treatment Stergiou et al. (2021).

Correct and approved Blood Pressure Measuring Devices (BPMD) are essential for measuring blood pressure (BP) because they are the acknowledged product that is supposed to produce accurate readings. Major devices include automatic oscillometric BP measurement of the upper extremities and auscultatory BP measurement. Due to the difficulties encountered in certain parts of the world, particularly in Low- to Middle-Income Countries, the appropriate devices may vary by location (LMIC). According to Organization et al. (2020), the obstacles include a lack of affordability or availability of such equipment, maintenance capacity, and inadequate staff training.

There are two primary methods for monitoring blood pressure: invasive (direct) and non-invasive (indirect). The intrusive technique includes putting a catheter into an artery and then measuring the observed pressure wave on a monitor, according to Schumann et al. (2021). Typically utilized in hospitals and other healthcare facilities. El-Hajj and Kyriacou (2021) indicated that the non-invasive approach for measuring blood pressure entails applying pressure to the arm's outside. The use of photoplethysmogram (PPG) signals is a new advance in the assessment of blood pressure. This approach is non-invasive. It can be accomplished by collecting two PPG signals from two anatomical sites and examining their waveforms. Observing the waveform obtained from a PPG signal and an electrocardiogram can also be employed with this method (ECG) Mejía-Mejía et al. (2021).

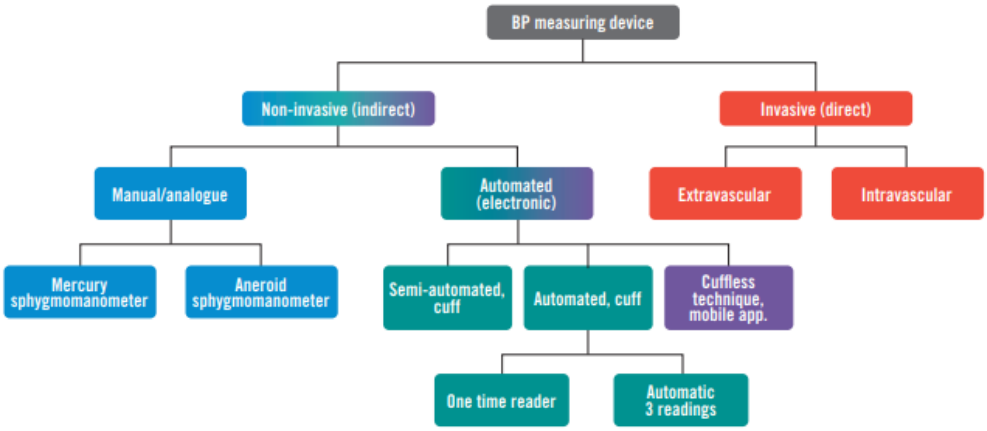


Figure 1: Blood Pressure Measurement Methods

## 2.3 ML Models for Hypertension Detection

Several previous investigations have already demonstrated the significance of hypertension detection in humans. The importance of controlling high blood pressure cannot be emphasized enough, as failure to do so can result in dire repercussions, including death.

According to Martinez-Ríos et al. (2021), in 2013 around 45 per cent of fatalities from heart disease and 51 per cent of deaths from stroke were caused by high blood pressure. These types of figures necessitate an investigation into the causes of such a surge. This project would investigate in depth the efficiency with which machine learning may be utilized to detect BP level variations. In addition, the healthcare industry has developed a culture of storing pertinent data that is beneficial for this reason, which has made it easier for data analysts to conduct their analyses. Existing research has also demonstrated how multiple machine learning algorithms have been employed and compared to determine which provides the highest degree of precision. Chowdhury et al. (2020) estimated BP values from PPG signals using machine learning (ML) approaches. Their investigation revealed that they were investigating feature selection approaches to make use of relevant features when creating their model and so limit the ML model's overfitting. Gaussian Process Regression (GPR) with the Relief feature selection technique surpasses other algorithms in terms of estimating SBP and DBP with RMSE values of 6.74 and 3.59, respectively.

Paviglianiti et al. (2021) subsequently compared Convolutional Neural Networks (CNNs) such as ResNet and WaveNet to Recurrent Neural Networks (RNNs). ResNet emerged as the model with the best performance after the investigation. AlKaabi et al. (2020) developed predictive models employing three machine learning (ML) techniques, including Logistic Regression, Decision Tree, and Random Forest. Based on the findings of the study, Random Forest and Decision Trees produced roughly the same level of accuracy with marginally superior outcomes than Logistic Regression. The body mass index (BMI), age, family history, and waist size are four important risk factors for hypertension, according to Zhao et al. (2021). This conclusion was reached after comparing four machine learning (ML) techniques, including Random Forest (RF), CatBoost, MLP Neural Network, and Logical Regression (LR). With an accuracy of 0.82, a sensitivity of 0.83, and a specificity of 0.81, RF was deemed to have outperformed the previous three models.

## 2.4 Previous Studies on Prediction of Blood Pressure

The study of Golino et al. (2014) introduced and applied the Classification and Regression Machine Learning Techniques to predict blood pressure among respondents. CART was used based on its increasing popularity and ease of interpretation. The results were based on fifteen variables used in the training sample. The study showed that men have higher systolic blood pressure. The study also showed that systolic pressure correlated with anthropometric variables. It was also revealed from the study that for women, BMI, WC, and WHR were the best predictors combined with the lowest deviance. However, for men, BMI, WC, HC, and WHC showed the best prediction with the lowest deviance (57.25). The result finally predicted that the classification tree analysis was a better predictor than traditional logistic regression. This study provided a better insight into the prediction of blood pressure levels based on gender as there are physiological variations in clinical diagnosis based on gender. However, the study is limited in that it uses only a type of classification technique limiting the uniqueness of seeking the best possible predictive model.

LaFreniere et al. (2016) worked on the use of machine learning to predict hypertension from a given clinical dataset. This was done using an artificial neural network. Risk factors were identified and used to predict the hypertension risk of an individual. The neural network model presented had an accuracy of 82%. This works better than the

works of Golino as a larger sample size was used (185, 371). However, using a wider range of evidence-based factors in literature than those available in the Canadian database used will provide more predictive accuracy for the training network used for the study. The usability of the suggested model also needs to be tried to ascertain its effectiveness.

Aparna and Babu (2020) worked on Blood Pressure prediction using Machine Learning algorithms. The study used unsupervised reinforced learning. A regression algorithm was used to predict Systolic Blood Pressure using a dataset collected from sources. The created system may also be a fundamental machine learning algorithm for Systolic Blood level Sign Prediction. To give more accurate information on the SBP forecast, the system's accuracy is frequently upgraded. The code can also be turned into an executable file for Python, making it simple to install and use on any system. To further improve prediction abilities, the anomalies relating to invalid and/or less accurate values are frequently deleted.

Ji et al. (2021) explained that it is medically hard to forecast hypertension problems. The study used machine learning and data mining to predict hypertension problems. GB DT-based feature selection can screen out hypertension problems. LightGBM-based hypertension complications prediction model. The accuracy, F1, and AUC of the prediction models are 0.9189, 0.8888, and 0.9233 after 10-fold cross-validation and comparison analysis, which is much superior to other machine learning models. The proposed method can reliably identify hypertension problems, so clinicians can take preventive actions to decrease their impact. However, a supervised machine learning model and a bigger variation of models could serve as a better predictor.

Chai et al. (2022) examined if anthropometric measurement works better than machine learning. The anthropometric measurements have been non-invasive predictive models. An imbalanced dataset of 2461 samples was used for the study. The balanced nature of the data could be a limitation of the study. The training dataset was reduced to age, C index, ethnicity, gender, height, location, parental hypertension, and waist circumference. The target variable of the dataset used was balanced using SMOTE and random under-sampling. The appropriate hyper-parameter models were evaluated using accuracy, precision, sensitivity, specificity, F1-score, misclassification rate, and AUC. No model consistently topped all seven-performance metrics. LightGBM outperformed Decision Tree in all six categories except sensitivity. We used Bayes' Theorem to assess the models' applicability in the Sarawak teenage population. The top four models were LightGBM, Random Forest, XGBoost, and CatBoost, whereas the bottom four were Logistic Regression, LogitBoost, SVM, and Decision Tree. This study shows that machine learning models affect the prediction of results.

This study seeks to build on previous research by using novel data and other machine learning models (mostly ensemble models) that would accurately predict the resting blood pressure of patients.

# 3 Methodology

## 3.1 Research Methodology

This section explains the steps involved in answering the research questions. This described the steps and stages involved in providing answers to the research questions. The stages involved in the KDD methodology are described in details by Tengnah et al. (2019). The study employed the use of Knowledge Discovery in Databases (KDD) meth-

odology. The KDD methodology starts with the Data Selection and Integration Stage followed by Data Cleaning and Pre-Processing, then Data Transformation, Data Mining, and Evaluation and Interpretation Stages.

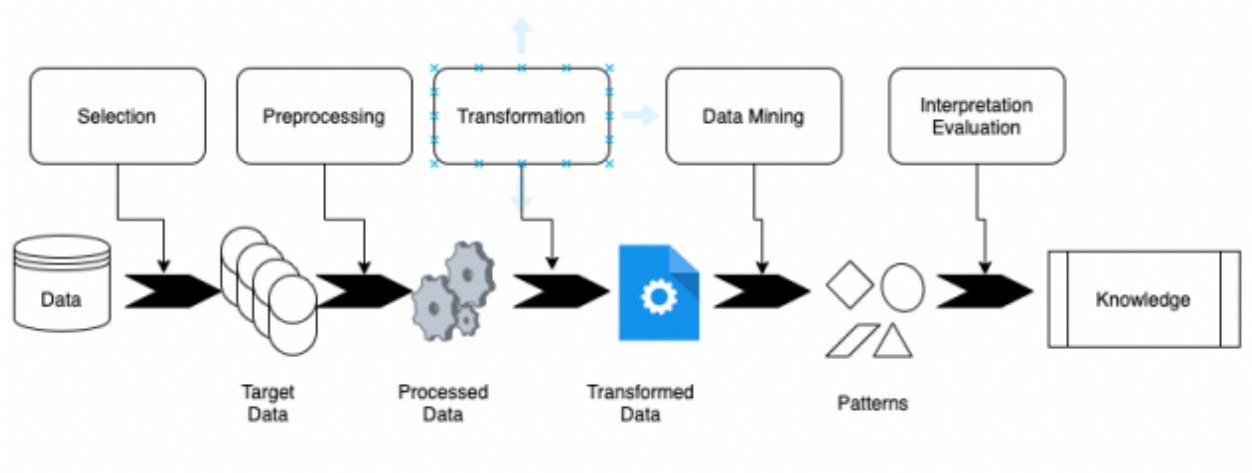


Figure 2: The Lifecycle of the KDD

The stage-by-stage description of the methodology will be discussed in this section. All the steps that go into gathering and selecting data, cleaning it, and preparing it for analysis are included here. The data would also be transformed to make it easier to analyse. Using the exploratory data analysis (EDA) that is part of the data mining process, patterns and insights can be gleaned after the data has been transformed.

## 3.2 Data Selection and Integration

The data was collected and sourced from Kaggle.com. The public availability of data reduces the cause for concern for any ethical bridge that could be created. The dataset used for the project included 14 features. Listed below are the features that have been selected for training, testing, and estimation:

- Age: As a person ages, his or her blood pressure is more likely to rise due to the constriction of blood vessels. This can also increase the risk of hypertension, indicating that older individuals are more susceptible to developing hypertension.
- Sex (or gender): This refers to a person's gender. In general, among individuals of the same age, men have a higher BP than women. This column is represented by binary values, with 1 representing Male and 0 representing Female.
- Type of Chest Pain (cp): This variable is translated to four values ranging from 0 to 3 and describes the sort of chest pain the individual typically suffers. 0 indicates conventional angina, 1 indicates atypical angina, 2 indicates non-angina discomfort, and 3 indicates asymptomatic.
- Resting blood pressure: A time of 3 to 5 minutes of rest is typically recommended before measuring blood pressure; hence, this column provides the measurement of the Systolic Blood Pressure value for each individual. This is our dependent variable.

- Chol: This indicates the serum cholesterol concentration in mg/dL.
- Fasting Blood Sugar (fbs): This characteristic describes the individual's fasting blood glucose level. It is called fasting blood sugar because a blood sample is collected following an overnight fast. If a person's fbs level is less than 100 mg/dL, it is regarded normal; if it is between 100 and 125 mg/dL, it is deemed pre-diabetes; and if, after two tests, it is greater than or equal to 126 mg/dL, he or she has diabetes. Now, these statistics simply indicate whether the fbs is larger than 120, i.e., whether the individual has diabetes. So, 1 = true; and 0 = false.
- Resting electrocardiographic (ECG) results: This section comprises the ECG results obtained from the patient while at rest. This records information regarding the resting heart rate and rhythm. 0 = normal, 1 = ST-T wave abnormalities, and 2 = probable or confirmed left ventricular hypertrophy according to the criteria of Estes.
- Thalach: This indicates the highest heart rate reached.
- Exang: This page provides information regarding exercise-induced angina. 1 = Yes; 0 = No.
- Oldpeak: This section offers information regarding ST depression generated by activity in comparison to rest.
- Slope: Here are the values for the peak exercise ST segment's slope. 0 Indicates upward slope, 1 = level, and 2 = downward sloping.
- Ca: The number of major blood arteries (0-3) that are coloured by fluoroscopy.
- Thal: This denotes the degree of defect observed in the individual based on the physiological and clinical tests described previously. 0 = normal, 1 = flaw fixed, and 2 = defect reversible.
- Target: This column indicates whether the individual has a cardiac ailment. 1 indicates that the individual has a cardiac disease, whereas 0 indicates that he does not.

The target feature for this experiment will be the Resting Blood Pressure column; therefore, as the data is being cleaned, the column heading for the final column, which is now labelled "Target," will be modified, as the data was previously used for a different analysis.

## 3.3 Data Pre-processing

The used data would be thoroughly reviewed and checked for null/missing values, and if any were discovered, they would be handled in a manner that facilitates their use in the proposed project. Additionally, the data would be prepared by doing the subsequent steps:

- Importing all essential libraries
- Reading the dataset into our IDE for coding

- Checking for missing or null values and if any are found then handle appropriately

This stage of the whole experiment is very essential as it helps to prepare the data in such a way that it becomes meaningful and improves the data quality for analytical purpose Sakr et al. (2018).

## 3.4 Data Exploration and Visualization

Exploring a dataset is one of the first steps towards engaging with it. Exploration is to grasp the data's global landscape and detect unique values. Exploration is used for data cleansing and editing, as well as visual discovery and the production of "hypotheses." Examining multiple data aggregations and summaries, both numerically and graphically, is one method for exploring data. This includes examining each variable separately as well as the relationships between variables. The goal is to find trends and exceptions. Data Visualisation or Visual Analytics refers to the process of exploring data through making charts and dashboards. We utilise histograms and box plots to learn about the distribution of numerical variables' values, locate outliers (extreme observations), and find additional information relevant to the analysis task. We can also examine scatter plots of pairs of numerical data to learn about potential links, the type of relationship, and to identify outliers. It was observed that:

- Sex: We have more males (1) than females (0).
- Chest pain type (Cp): We have more individuals with typical angina (0) and few individuals that are asymptomatic (3).
- Fbs: More individuals have their fasting blood sugar values greater than 120mg/dL and few individuals have their blood sugar values less than 120mg/dL.
- Restecg: We have individuals with more occurrences of ST-T wave abnormality followed by almost a matching number of individuals with normal ECG results. There are only a few cases of individuals showing probable or definite left ventricular hypertrophy by Estes' criteria.
- Exang: We have more cases of individuals with no exercise-induced angina as compared with individuals with exercise-induced angina.
- Slope: We have more individuals with occurrences of flat slope of peak exercise ST segment. Occurrences of individuals with down-sloping ST segment are very close to those having a flat slope but there are very few occurrences of individuals with up-sloping ST segment.
- Ca: The major vessel represented as 0 is the most prevalent among individuals followed by the ones represented as 1, 2, 3 and 4 respectively.
- Heart disease: We have more individuals (represented as 1) with cases of heart disease i.e with over 50% diameter narrowing.

The result showed that there are more females with average systolic blood pressure more than 120.

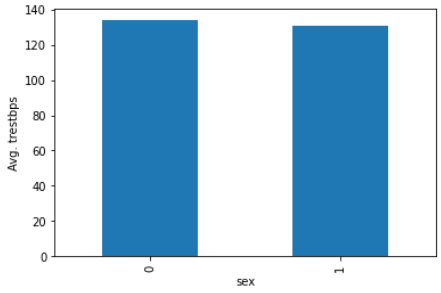


Figure 3: Average Resting BP for each sex

## 3.5 Data Transformation

The dataset structure would be modified at this point in order to facilitate analysis. In this case, an exploratory data analysis (EDA) process would be used to better comprehend the data's patterns. Structured data might be improved and possibly expanded to better explain the project by creating new features from existing columns to help with an effective EDA. It was observed that a linear relationship exists between age and systolic blood pressure. Positive linear relationship exists between how serum cholesterol might affect blood pressure (BP) levels at rest. It was also observed that negative linear relationship between age and maximum heart rate achieved which gives validity to the medical fact that: the decrease in maximum heart rate which occurs with ageing is a non-modifiable and inevitable consequence of ageing. The EDA shows different types of relationships from the different shapes (e.g., a linear relationship between trestbps and age). Along the diagonal, where just a single variable is involved, the frequency distribution for that variable is displayed.

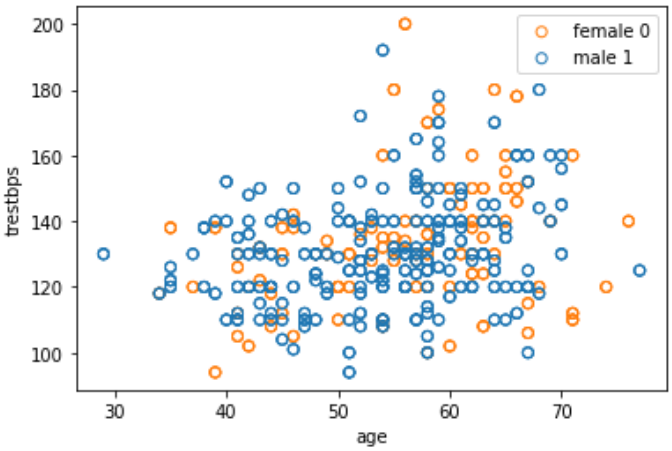


Figure 4: Scatterplot of Resting BP against Age

The measures of central tendency, shape and spread among selected predictors were evaluated. Some of the results are presented below. It was obvious that none of the variables are normally distributed. The algorithms which will be used for modelling (e.g Tree Based Method and Ensemble Models) do not make any assumptions over the distribution of the predictors and don't require feature scaling ( Standardisation and

Normalisation). Hence, making it easier to use the available data without too much consideration on data distribution, however, outlier variables were treated to prepare the training data for the model as most models perform better when they are noiseless so that the models will not be skewed in a negative direction.

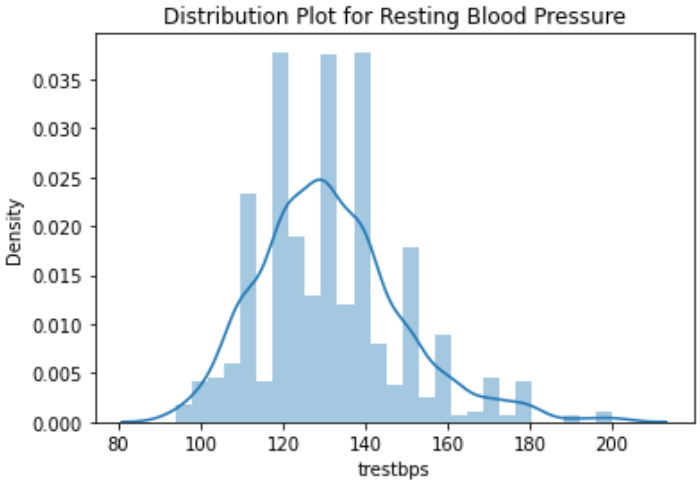


Figure 5: Distribution Plot for Resting BP

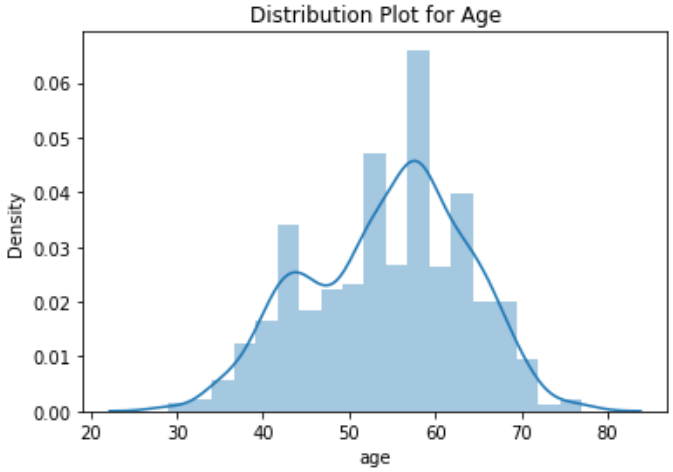


Figure 6: Distribution Plot for Age

## 3.6 Feature Engineering and Feature Selection

Feature engineering and feature selection helps to extract the most valuable set of variables for modelling. It helps to select or generate new predictors devoid of redundancy. Correlation matrix was used in this process. Two or more independent variables that have the same correlation values with the target variable were identified to reduce the effect of multicollinearity. Only one of such variables were selected as predictors. Multicollinearity is the presence of two or more predictors with the same linear relationship

with the outcome variable. The predictors that reduces multicollinearity were selected for modelling and they are age, trestbps, chol, thalach, oldpeak, sex, cp2, fbs, restecg, exang, slope, ca, thal, heart disease.

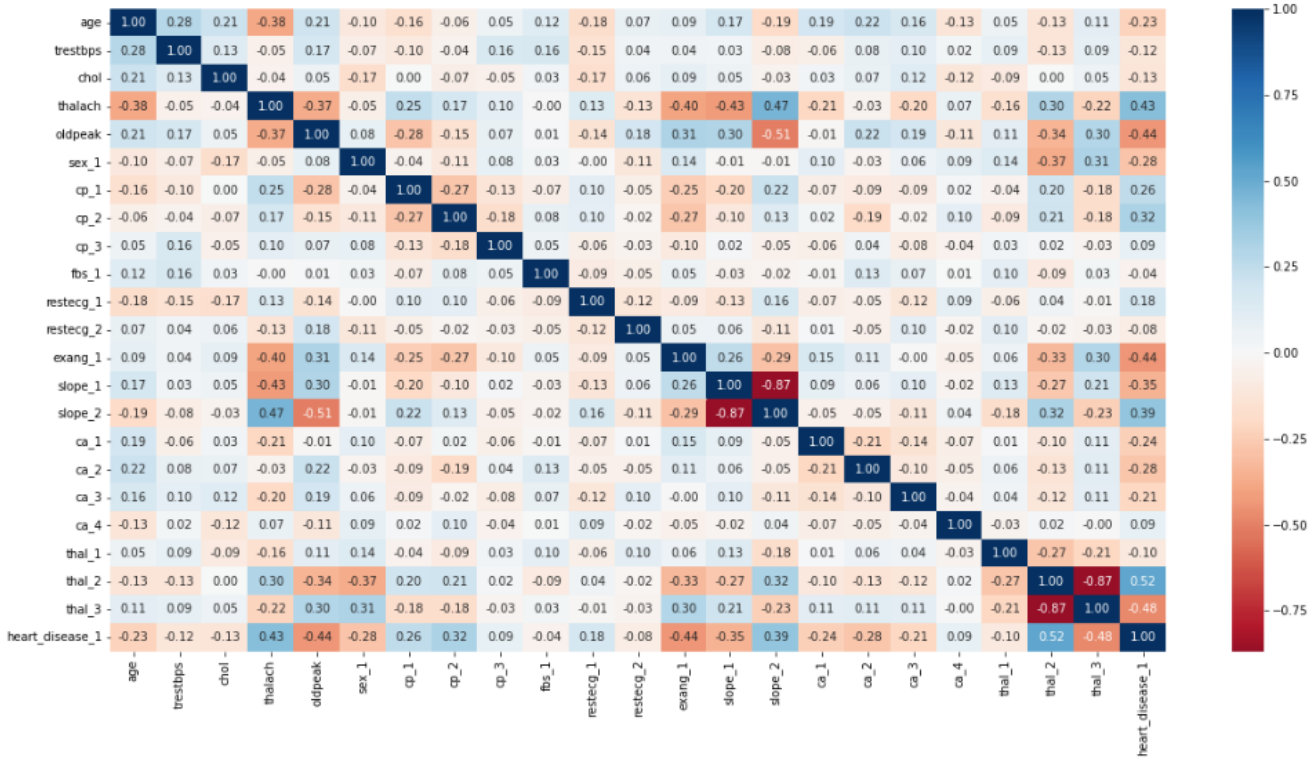


Figure 7: Correlation of Variables

## 3.7 Data Mining and Modelling

We'll be looking at how Decision Tree, Extra Trees, Random Forest, LightGBM, XGBoost models have been utilised in the past, and we will see if there are any ways in which the prior models may be enhanced to attain higher accuracy. Chowdhury et al. (2020) had also employed Gaussian Process Regression (GPR) as their best performing model amongst to get an RMSE value for the SBP to be 6.74, so we aim to observe a better performing model during this experiment.

The study also looks at the differences in outcomes between running the model locally on a machine and running it in the cloud. It is at this point that each model is produced, and each model is evaluated on both the localhost system and the remote host machine. Each model's performance would be evaluated using a set of measures that could be applied to both contexts.

When each model is constructed, the data prepared during the data pre-processing stage is input into the models for evaluation on the Google Collaboratory notebook. Using measures across both contexts, the accuracy of each model was tested.

# 4 Design Specification

The design specification of this project describes the architectural overview of the research. It shows the roadmap through the entire research. This helps to summarize the overall project architecture with details of the techniques, and tools involved.

The architectural design of this project involves three layers which are the Data layer, the Business Logic layer and the Client/ Presentation layer. Each layer is described below:

- **Data Layer:** This layer involves all the activities that revolve around the preparation of the data right from its raw form till it becomes suitable to be used for modelling. It consists of the cleaning of the data, treating of missing values transformation, and then providing it for the Business Logic layer.
- **Business Logic Layer:** This is the layer where the implementation of the modelling process takes place on the data. Here the classification models are run on the data so as to extract patterns and make predictions to generate insights from the data. The models are then evaluated by using some performance metrics which include Accuracy, Sensitivity, etc.
- **Client/Presentation Layer:**This is the layer where the insights derived from the business logic layer are presented to the clients. The presentations are shown via visualizations of the patterns obtained. The tool that was used in this layer in generating the visualizations for presentation to the client was the python programming language.

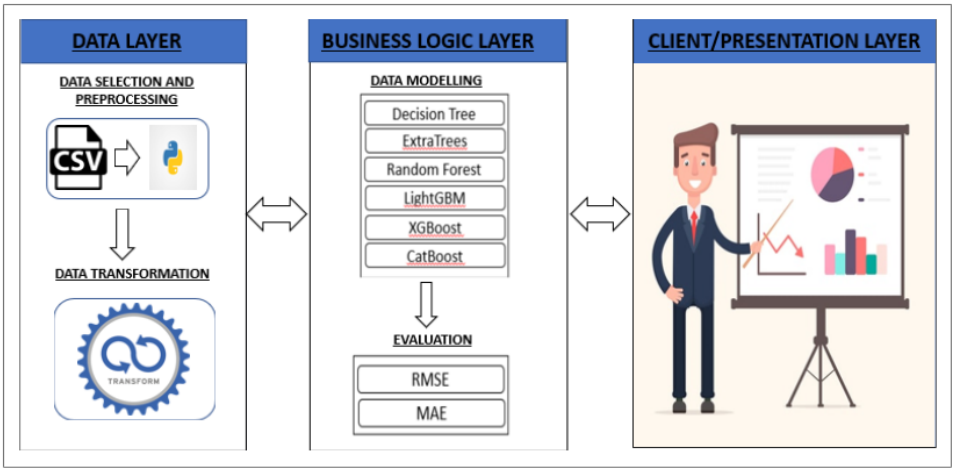


Figure 8: The Design Specification Flow

In conclusion, the design methodology and the design specification used for this particular research were chosen to achieve the given objectives of this research.

# 5 Implementation

## 5.1 Implementation Tools

The project was implemented and visualised using Anaconda Python distribution via Jupyter notebook. A Windows 10 Operating System was used having 16GB RAM and 500GB ROM. The model was also run on Google Colaboratory to check for consistency of results. This assisted in making predictions using different sampling techniques with the evaluation metrics for all implemented models.

## 5.2 Model Building

The algorithms that will be used for building the model are: "Decision Tree Regressor", "Extra-Trees Regressor", "Random Forest Regressor", "Light Gradient Boosting Machine" (LightGBM), "Extreme Gradient Boosting"(XGBoost) and "CatBoost". Default parameters were used for most of the models.

### 5.2.1 Implementation of Decision Tree Regressor

It creates tree-structured regression and classification models. It divides a dataset into smaller parts while developing a decision tree. The outcome is a decision-and-leaf tree. A decision node has two or more branches indicating attribute values. Leaf node represents the target decision. The uppermost decision node in a tree is called the root node. Decision trees handle categorical and numerical data. For the blood pressure prediction, Decision Tree Regressor with a maximum depth of 5 was utilised.

### 5.2.2 Implementation of Extra-Tree Regressor

This class implements a meta estimator that employs averaging to increase predicted accuracy and control over-fitting. Extra Trees Regressor with a maximum depth of 5 was used during the model building for the blood pressure prediction.

### 5.2.3 Implementation of Random Forest Regressor

It is a machine learning technique that uses numerous decision trees and a statistical technique called bagging to perform both regression and classification problems. Bagging and boosting are two of the most prominent ensemble strategies for dealing with excessive volatility and bias. Instead of simply averaging tree prediction, a Random Forest algorithm employs two fundamental elements that give it the name random: a) When building trees, use random selection of training observations. b) Random feature subsets for separating nodes. In other words, rather than depending on individual decision trees, Random Forest constructs numerous decision trees and merges their forecasts to get a more precise and reliable prediction. Random Forest Regressor with a maximum depth of 5 and 30 estimators was used during the model building for the blood pressure prediction.

### 5.2.4 Implementation of LightGBM

It is a gradient boosting framework that uses sophisticated tree-based learning techniques.It is a Fast-processing algorithm. LightGBM grows leaf-wise while other algorithms grow level-wise. LightGBM grows on leaves with high loss. It reduces leaf

loss more than a level-wise algorithm. LGBM Regressor with a maximum depth of 5 and 30 estimators was used during the model building for the blood pressure prediction.

#### 5.2.5 Implementation of XGBoost

This is an open-source gradient enhanced trees implementation. Gradient boosting combines the estimates of simpler, weaker models to forecast a target variable. The weak learners in gradient boosting for regression are regression trees, and each one transfers an input data point to a leaf with a continuous score. XGBoost minimises a regularised (L1 and L2) objective function composed of a convex loss function and a penalty for model complexity (in other words, the regression tree functions). The ultimate forecast is made by iteratively adding additional trees that anticipate the residuals or errors of earlier trees. Gradient boosting minimises model loss by using a gradient descent approach. XGB Regressor with a maximum depth of 5 and 30 estimators was used during the model building for the blood pressure prediction.

#### 5.2.6 Implementation of CatBoost

CatBoost combines decision trees and gradient boosting. Boosting combines numerous weak models to generate a powerful predictive model using greedy search. Gradient boosting fits decision trees progressively, so the fitted trees learn from previous failures and reduce errors. It continues adding functions until the selected loss function is no longer minimised. CatBoost doesn't use gradient boosting to create decision trees. CatBoost grows trees where all nodes at the same level test the same predictor with the same condition, hence a leaf index can be predicted using bitwise operations. The oblivious tree process is simple and efficient on CPUs, while the tree structure finds an optimal solution and prevents overfitting. For the blood pressure prediction, CatBoost Regressor with a maximum depth of 5, 30 estimators and learning rate 0.5 was used.

# 6 Evaluation of Results

The performance of each of the models are evaluated using the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). In their 2014 paper, Chai and Draxler (2014) refute the concept that MAE should be the only metric used to measure normal errors, arguing instead for RMSE as the best option. Willmott et al. (2009), on the other hand, were correct that MAE is more resilient, though there are superior alternatives. Most crucially, neither side gives a theoretical basis for either metric. There is no evidence to support the claim that RMSE is better than MAE, but rather that a variety of metrics should be utilised to evaluate model performance. Models with multiple facets necessitate multi-faceted evaluation, hence, justifying the use of both evaluation performances for this work.

## 6.1 Evaluation of each model used

#### 6.1.1 Evaluation of Decision Tree Regressor

The RMSE score for Decision Tree Regressor was 13.433 while that for MAE score was 10.71. The MAE score performed lower.

| | Model | RMSE Score | MAE Score |
|---|---|---|---|
| 0 | DecisionTree | 13.433263 | 10.706429 |
| 1 | ExtraTrees | 12.387637 | 10.324089 |
| 2 | RandomForest | 11.822367 | 9.671689 |
| 3 | LightGbm | 10.846500 | 8.790167 |
| 4 | XGBoost | 11.799123 | 9.327126 |
| 5 | CatBoost | 7.307459 | 5.790854 |

Figure 9: Models Summary

### 6.1.2 Evaluation of Extra-Tree Regressor

The RMSE score for Extra Tree Regressor was 12.42 while that for MAE score was 10.33. The MAE score performed lower.

### 6.1.3 Evaluation of Random Forest Regressor

The RMSE score for Random Forest Regressor was 12.36 while that for MAE score was 10.00. The MAE score performed lower. The two sampling scores were lower than the Decision Tree and Extra Tree Regressors.

### 6.1.4 Evaluation of LightGBM Regressor

The RMSE score for LightGBM was 10.84 while that for MAE score was 8.79. The MAE score performed lower. The two sampling scores were lower than the Decision Tree, Random Forest and Extra Tree Regressors.

### 6.1.5 Evaluation of XGBoost Regressor

The RMSE score for XGBoost was 11.80 while that for MAE score was 9.33. The MAE score performed lower. The sampling score for XGBoost was higher than that of Light-GBM but lower than the other previous models and same can be accounted for the MAE scores.

### 6.1.6 Evaluation of CatBoost Regressor

The RMSE score for CatBoost was 7.30 while that for MAE score was 5.79. The MAE score performed lower. The sampling score for CatBoost model was the lowest. As it has been established that the closer to zero the better the performance of the model.

## 6.2 Hyperparameter Optimization

After the implementation of Grid Search CV for the model (CatBoost) with the best result among the initial algorithms used, the results obtained for the important CatBoost parameters after executing the search were presented: With Grid Search CV, the hyper

optimised parameters are: learning_rate is 0.3, max_depth is 9, n_estimators is 200, and random_seed is 60. Iter od_type was used to prevent overfitting the model.

After applying these optimised parameters to the CatBoost algorithm, we got improved evaluation scores from our model as against the initial scores when parameters were chosen at random. Our RMSE and MAE scorers have significantly reduced with values 0.88 and 0.12 respectively. The closer the scores are to zero, the better the performance of our model Chai and Draxler (2014).

RMSE: The RMSE statistic gives information regarding short-term model performance by permitting a term-by-term comparison of the actual difference between estimated and measured values. To calculate the RMSE, it is assumed that the errors are unbiased and distributed normally. Hence, the standard error (SE) and the root mean square error (RMSE) can be used to obtain a more full picture of the distribution of errors. When there are more samples, RMSEs will be even more accurate in reconstructing the error distribution. As the value gets smaller, so does its performance. (Chai and Draxler (2014), Savage et al. (2013))

MAE: Absolute error in machine learning refers to the difference between the predicted value of an observation and the actual value of that observation. For a group of predictions and observations, MAE considers the average of absolute errors as a measure of the size of mistakes.

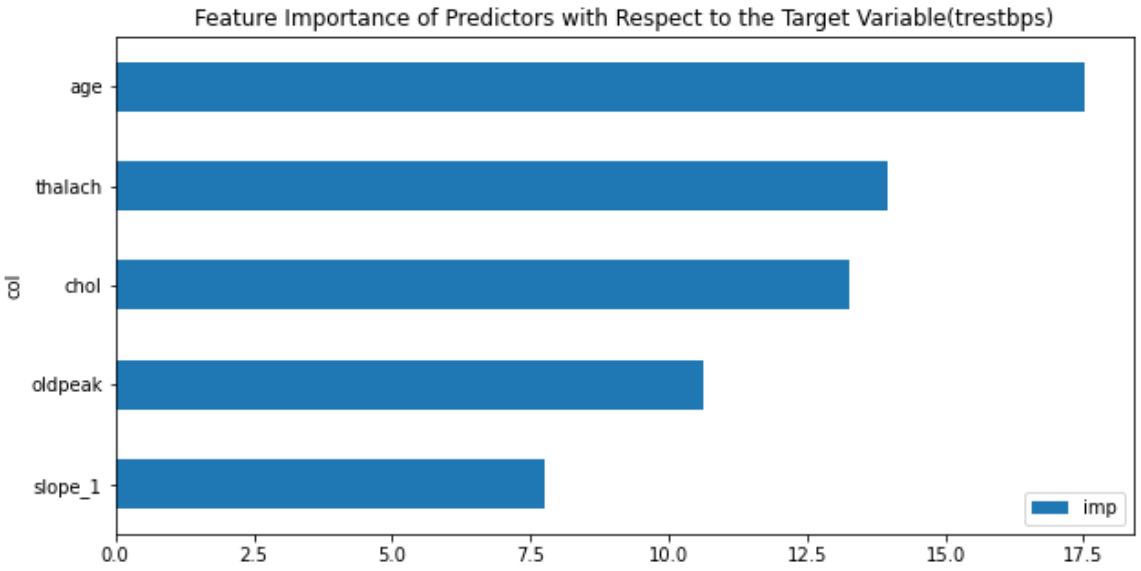


Figure 10: Feature Importance of Predictors

Using the independent variables for prediction with respect to the target variable, the chart above shows that age is the most important variable for predicting blood pressure, followed by the maximum heart rate achieved(thalach) and the serum cholesterol level(chol).

Because age is the most important predictor of blood pressure levels, older persons should prioritise guidelines such as frequent exercise and healthy eating. Frequent medical checkups should be performed to monitor maximum heart rate obtained, which is the second most important predictor, and high cholesterol meals should be avoided, which is the third most important predictor of blood pressure levels (Bosu (2016)).

We recommend that policymakers work with stakeholders to establish strategies emphasising the relevance of lifestyle changes and adherence coaching in order to effectively control blood pressure. Furthermore, we recommend that researchers conduct additional longitudinal studies involving physical and biochemical measurements in order to identify the most important variables associated with blood pressure control and to investigate the cause and effect relationship between variables and blood pressure control (Okoro and Ngong (2012))

# 7 Conclusion and Future Work

This study aimed to define an efficient model for the accurate prediction of blood pressure levels by building on existing models that are useful for HBP prediction. These models were selected based on RMSE and MAE evaluation metrics that were used to evaluate their performance. The RMSE metric was used to provide information by doing a term-by-term comparison and shows the value-performance relationship between selected models; while the MAE metric shows the average of absolute errors that the selected models are liable to. The data for this study was sourced from Kaggle.com because it is publicly available and reduces the chance for ethical misconduct. Features of the data selected is useful for training, testing, and estimation of the study outcome. This was done through the processes of pre-cleaning, visualisation, transformation, feature engineering, and modelling.

The CatBoost model outperformed other models with RMSE score of 7.307459 and MAE score of 5.790854; these scores was further improved after performing a Grid Search which gave good evaluation scores where the RMSE and MAE scores were reduced greatly. The study also shows that the most significant variable for blood pressure measurement is age while the least significant is the number of major vessels coloured by fluoroscopy (ca_4).

The study provided accurate prediction for patients' hypertension complication, it will also help to effectively analyse factors affecting complication so that medical personnel can control these factors or advise on managing potential factors, hence, improving survival rate. However, for further studies, large dataset can be used to ensure effective predictions as this gives better predictive variables and abilities. Data from other sources and regions can also be used or compared to enhance generalised prediction models and factors affecting HBP across various divides.

## 7.1 Acknowledgement

I would like to send my profound gratitude to my supervisor, Jorge Basilio who gave all needed and necessary guidance and knowledge throughout the research work. I would also like to thank my family for their moral support throughout the time of this research.

# References

AlKaabi, L. A., Ahmed, L. S., Al Attiyah, M. F. and Abdel-Rahman, M. E. (2020). Predicting hypertension using machine learning: Findings from qatar biobank study, *Plos one* **15**(10): e0240370.

Bard, J. A., Bashore, C., Dong, K. C. and Martin, A. (2019). The 26s proteasome utilizes a kinetic gateway to prioritize substrate degradation, *Cell* **177**(2): 286–298.

Bosu, W. K. (2016). Determinants of mean blood pressure and hypertension among workers in west africa, *International Journal of Hypertension* **2016**.

Chai, S. S., Goh, K. L., Cheah, W. L., Chang, Y. H. R. and Ng, G. W. (2022). Hypertension prediction in adolescents using anthropometric measurements: Do machine learning models perform equally well?, *Applied Sciences* **12**(3): 1600.

Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature, *Geoscientific model development* **7**(3): 1247–1250.

Chowdhury, M. H., Shuzan, M. N. I., Chowdhury, M. E., Mahbub, Z. B., Uddin, M. M., Khandakar, A. and Reaz, M. B. I. (2020). Estimating blood pressure from the photoplethysmogram signal and demographic features using machine learning techniques, *Sensors* **20**(11): 3127.

El-Hajj, C. and Kyriacou, P. A. (2020). A review of machine learning techniques in photoplethysmography for the non-invasive cuff-less measurement of blood pressure, *Biomedical Signal Processing and Control* **58**: 101870.

El-Hajj, C. and Kyriacou, P. A. (2021). Deep learning models for cuffless blood pressure monitoring from ppg signals using attention mechanism, *Biomedical Signal Processing and Control* **65**: 102301.

Fuchs, F. D. and Whelton, P. K. (2020). High blood pressure and cardiovascular disease, *Hypertension* **75**(2): 285–292.

Golino, H. F., Amaral, L. S. d. B., Duarte, S. F. P., Gomes, C. M. A., Soares, T. d. J., Reis, L. A. d. and Santos, J. (2014). Predicting increased blood pressure using machine learning, *Journal of obesity* **2014**.

Ji, X., Chang, W., Zhang, Y., Liu, H., Chen, B., Xiao, Y. and Zhou, S. (2021). Prediction model of hypertension complications based on gbdt and lightgbm, *Journal of Physics: Conference Series*, Vol. 1813, IOP Publishing, p. 012008.

Katamba, G., Agaba, D. C., Migisha, R., Namaganda, A., Namayanja, R. and Turyakira, E. (2020). Prevalence of hypertension in relation to anthropometric indices among secondary adolescents in mbarara, southwestern uganda, *Italian Journal of Pediatrics* **46**(1): 1–7.

LaFreniere, D., Zulkernine, F., Barber, D. and Martin, K. (2016). Using machine learning to predict hypertension from a clinical dataset, *2016 IEEE symposium series on computational intelligence (SSCI)*, IEEE, pp. 1–7.

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Karlsson Linnér, R. et al. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals, *Nature genetics* **50**(8): 1112–1121.

Martinez-Ríos, E., Montesinos, L., Alfaro-Ponce, M. and Pecchia, L. (2021). A review of machine learning in hypertension detection and blood pressure estimation based on clinical and physiological data, *Biomedical Signal Processing and Control* **68**: 102813.

Mejía-Mejía, E., May, J. M., Elgendi, M. and Kyriacou, P. A. (2021). Classification of blood pressure in critically ill patients using photoplethysmography and machine learning, *Computer Methods and Programs in Biomedicine* **208**: 106222.

Okoro, R. N. and Ngong, C. K. (2012). Assessment of patient's antihypertensive medication adherence level in non-comorbid hypertension in a tertiary hospital in nigeria, *Int J Pharm Biomed Sci* **3**(2): 47–54.

Organization, W. H. et al. (2020). Who technical specifications for automated non-invasive blood pressure measuring devices with cuff. geneva, switzerland: World health organization (who); 2020.

Paviglianiti, A., Randazzo, V., Villata, S., Cirrincione, G. and Pasero, E. (2021). A comparison of deep learning techniques for arterial blood pressure prediction, *Cognitive Computation* pp. 1–22.

Sakr, S., Elshawi, R., Ahmed, A., Qureshi, W. T., Brawner, C., Keteyian, S., Blaha, M. J. and Al-Mallah, M. H. (2018). Using machine learning on cardiorespiratory fitness data for predicting hypertension: The henry ford exercise testing (fit) project, *PLoS One* **13**(4): e0195344.

Savage, N., Agnew, P., Davis, L., Ordóñez, C., Thorpe, R., Johnson, C., O'Connor, F. and Dalvi, M. (2013). Air quality modelling using the met office unified model (aqum os24-26): model description and initial evaluation, *Geoscientific Model Development* **6**(2): 353–372.

Schultz, M. G., Currie, K. D., Hedman, K., Climie, R. E., Maiorana, A., Coombes, J. S. and Sharman, J. E. (2022). The identification and management of high blood pressure using exercise blood pressure: Current evidence and practical guidance, *International journal of environmental research and public health* **19**(5): 2819.

Schumann, R., Meidert, A. S., Bonney, I., Koutentis, C., Wesselink, W., Kouz, K. and Saugel, B. (2021). Intraoperative blood pressure monitoring in obese patientsarterial catheter, finger cuff, and oscillometry, *Anesthesiology* **134**(2): 179–188.

Stergiou, G. S., Palatini, P., Parati, G., O'Brien, E., Januszewicz, A., Lurbe, E., Persu, A., Mancia, G., Kreutz, R. et al. (2021). 2021 european society of hypertension practice guidelines for office and out-of-office blood pressure measurement, *Journal of Hypertension* **39**(7): 1293–1302.

Tengnah, M. A. J., Sooklall, R. and Nagowah, S. D. (2019). A predictive model for hypertension diagnosis using machine learning techniques, *Telemedicine Technologies*, Elsevier, pp. 139–152.

Willmott, C. J., Matsuura, K. and Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics, *Atmospheric Environment* **43**(3): 749–752.

Zhao, H., Zhang, X., Xu, Y., Gao, L., Ma, Z. and Sun, Y. (2021). Predicting the risk of hypertension based on several easy-to-collect risk factors: a machine learning method, *Frontiers in Public Health* p. 1395.

Zheng, J. and Yu, Z. (2021). A novel machine learning-based systolic blood pressure predicting model, *Journal of Nanomaterials* **2021**.

Zhou, B., Perel, P., Mensah, G. A. and Ezzati, M. (2021). Global epidemiology, health burden and effective interventions for elevated blood pressure and hypertension, *Nature Reviews Cardiology* **18**(11): 785–802.