Application of machine learning techniques to predict the probability of various chronic diseases.

**Abstract:** This research study focuses on contributing to the healthcare industry through the proposal of an e-diagnosis or prediction system which functions through the implementation of machine learning models. This study focuses on the analysis of chronic diseases using machine learning techniques such as , utilising datasets provided online, that is, from Mendeley Data and Kaggle. Datasets related to Diabetes and Heart failure are employed to build reliable prediction models. The system aims to process the datasets and develop predictions after splitting the data to train and test the preprocessed datasets. The predictive models encompass in specific, Diabetes and Cardiovascular disease(resulting in Heart failure), on which following techniques are employed; logistic regression, random forest, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), naive bayes (Bernouilli) and Neural Network. The performance of each model in terms of its accuracy, precision, F1 score, etc. have been assessed in order to evaluate their overall efficiency in dealing with the pre-processed datasets and to further enhance the models applied.

**Keywords:** e-diagnosis, diabetes, cardiovascular,random forest, SVM, KNN, naive bayes, neural network

**1.INTRODUCTION**

All over the world, chronic diseases are a critical issue in the healthcare domain. According to the medical statement, due to chronic diseases, the death rate of humans increased up to 67% in the past few years. The treatments given for this disease consume over 70% of the patient’s income. Hence, it is highly essential to minimise the patient’s risk factor that leads to death. This can be done in different ways but the focus in this study is to use machine learning models to identify patterns among patients and their medical results using supervised learning, that is, by training the model using datasets with the outcome mentioned as well. Such models could allow the early detection of chronic diseases such as diabetes and heart related issues, which is our primary focus in this research project.

The advancement in medical research makes health-related data collection easier The healthcare data includes the demographics and medical analysis reports. This platform combines machine learning methods to explore the commonly used computing methods in cardiology and diabetes, such as logistic regression, random forests, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes (Bernoulli), AdaBoost and Neural networks. By comparing the accuracy of seven algorithms for predicting chronic diseases, we aim to improve the accuracy of chronic disease prediction, in particular diabetes and cardiac issues by providing the best medical product and service in terms of early diagnosis of such medical issues.

**2.MACHINE LEARNING**

Endowing machines with the ability to learn like humans is akin to a dream, as machines lack inherent intelligence. The disparity between human and machine capabilities lies in intelligence, where humans can learn from past experiences, a capacity absent in machines that must be programmed with specific instructions. Presently, machine learning enables computers to learn from experiences, a departure from traditional computational algorithms that relied on hard-coded instructions explicitly provided for problem-solving. Thus, this study aims to make use of machine learning to accurately make predictions for early diagnosis of diabetes and heart failure.

**Machine Learning applications in healthcare:**

Machine learning (ML) has played a pivotal role in revolutionising healthcare, streamlining intricate tasks and enhancing overall efficiency. These technologies not only reduce costs but also expedite drug discovery and enhance therapeutic outcomes, garnering significant attention and investment from healthcare stakeholders while also presenting unprecedented opportunities in the healthcare industry

ML applications in healthcare can be prove to be extremely beneficial to the area of concern, this can be categorised into three main groups:

1. Improving Available Medical Structures**:**

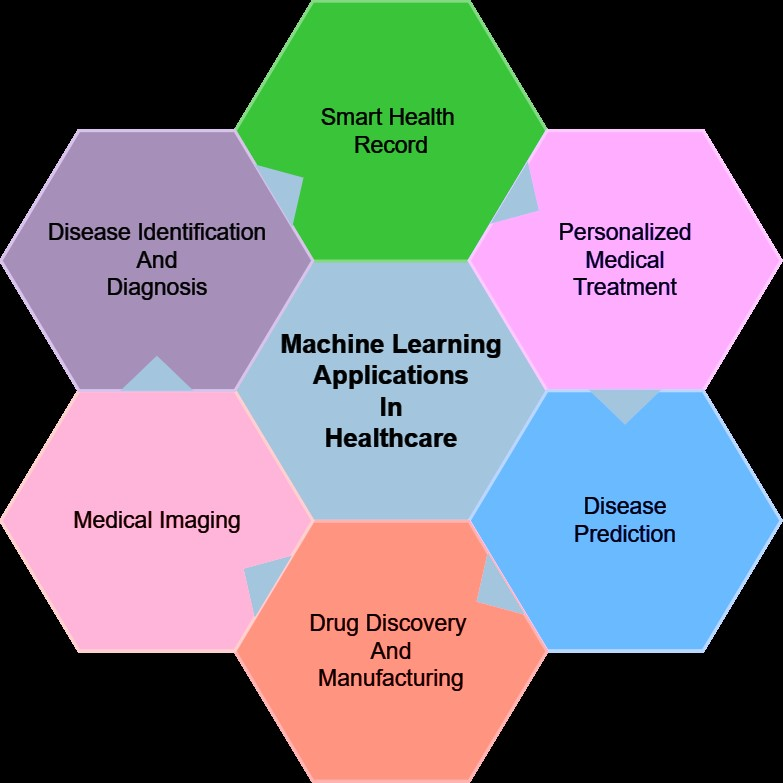
These applications enhance the performance of existing medical structures. ML-based technologies define specific rule-based tasks, such as classifying digital medical images, improving the accuracy of traditional image processing techniques. For instance, Aindra, a medical company utilizing artificial intelligence and ML, employs a platform for the accurate and swift diagnosis of cancers through medical image classification.

2. Upgrading Medical Structures:

This category focuses on providing medical structures with new capabilities, moving towards personalization. Precision medicine is a notable ML application, tailoring medical treatment to an individual's specific characteristics, such as genetic makeup. Companies like iCarbonx leverage large datasets, biotechnology, and artificial intelligence to advance personalised healthcare services.

3. Independent Medical Structures:

This emerging category involves creating ML-based models capable of independent action based on predefined goals. An intriguing prospect in healthcare is the development of hospitals without physicians, where robots autonomously handle healthcare processes from diagnosis to surgery. While this technology is currently in use for surgeries in some developed countries, it is evolving and undergoing rigorous testing to meet various standards. The Mayo Clinic, for example, is progressing towards a doctor-less hospital, utilising robotic assistance to enhance surgical procedures. Though this technology has imperfections, ongoing advancements indicate its promising future in reshaping healthcare processes.

Figure 1: Machine learning applications in healthcare

**Framework for Designing a Learning Model in Medicine**

This section outlines the different stages involved in crafting a learning model specifically tailored for the healthcare domain. The intent is to provide researchers with insights into the process of designing a learning model in medicine. We encourage researchers to delve further into this domain to gain a comprehensive understanding and knowledge of learning models. Designing a learning model in healthcare entails five essential phases: problem definition, dataset selection, data preprocessing, development of machine learning (ML) models, and evaluation. Figure 3 illustrates these phases. The subsequent sections elaborate on each of these phases in detail.

Problem

Definition

Data

Data

Preprocessing

Model Development

Evaluation of Results

Figure 2: Framework for designing a machine learning model

**Literature Review**

Victor Chang et al (2022) focused primarily on diabetes for which they made use of the Pima Indian Diabetes Dataset and used techniques such as, PCA, K-means clustering and importance ranking for data pre-processing. Models implemented included Naive Bayes classifier, random forest classifier and J48 decision tree models. In addition to this, feature selection was made use of wherein 3 and 5 factor feature selection was used. They obtained an accuracy of 80-90% using the aforementioned models and techniques.

Ahmad Shaker Abdalrada et al’s (2022) study aims to identify people with the co-occurrence of diabetes and cardiovascular diseases. This is a detailed study and analysis which was done using data collected on 200+ variables from more than 2000 patients. Logistic regression and Evimp functions were made use of in multivariate adaptive regression splines models for interpreting the common causes for the co-occurrences of the chronic diseases.

An Dinh et al (2019) make use of different variables and time frames to identify patients with diabetes and cardiovascular diseases. Logistic regression, support vector machines, random forest, and gradient boosting are the models which are made use of in their research study which were then combined to make a weighted ensemble model, which further increased the accuracy of their models. The accuracy came to be around 80% on implementation of the techniques.

Hang Lai et al (2021) made use of 8 features to develop predictive models and used logistic regression along with gradient boosting machine and used adjusted threshold method and class weight method to improve sensitivity. The accuracy of the models came around 70=80% and was compared with machine learning techniques such as decision tree and random forest, to which it performed better than.

**Problem definition**

According to the National Institute of Health, most people with type 2 diabetes do not have it diagnosed till very late, that is, they are unaware of such asymptomatic (or negligibly symptomatic) problems for up to 4-7 years. In order to tackle the issue, this study focuses on implementation of machine learning models for accurate prediction and diagnosis of underlying diseases. The developed solution would also help in getting early treatment for such non communicable and chronic diseases which would lead to better results in a shorter amount of time.

**Dataset and Exploratory Data Analysis**

In this research project, the main focus has remained not only on the analysis and development of machine learning models but also on the reliability of the results produced. Being considerate of the diversity and impartial results plays an important role in further enriching Science and our knowledge of it, especially in the healthcare domain. Thus, this research project has covered all the grounds by evaluating two datasets each, for cardiac problems and diabetes, in case of data being collected from a specific group of people sharing similar backgrounds.:

1. Diabetes Dataset - Mendeley Data - (<https://data.mendeley.com/datasets/wj9rwkp9c2/1>): This dataset is available online via CC by 4.0 This dataset covers data collected from thousand patients making it very usable for this study. This dataset has features in common and different from those covered in the Pima Indian dataset, which would train the model better by covering all grounds of causes of diabetes.
2. Pima Indians Diabetes Dataset -

(<https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/code>) - Consists of data collected from about 800 patients with 8 features apart from the outcome. Good variance of data and relevant features considered, focusing and pertaining to diabetes.

1. Heart failure dataset (<https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data>) - Consists of data collected from 300 patients, which is not very large but sufficient for training of data and getting good results. Removal of anomalies and outliers is not encouraged, as the size of the dataset is not big.
2. Heart disease dataset (<https://www.kaggle.com/datasets/yasserh/heart-disease-dataset/data>) - This dataset too, has data accumulated from 300 patients and all features are relevant to the outcome.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1:** Mendeley Data - Data summary   |  |  |  |  | | --- | --- | --- | --- | | **Feature** | **Description** | **Data Type** | **Range** | | ID | Identification Number of Patient | Numeric | [0,800] | | No\_Pation | Number of Patient | Numeric |  | | Gender | Gender (Male/Female) | Characters (M/F) | {M, F} | | Age | Age of Patient | Numeric | [20,79] | | Urea | Urea levels | Numeric | [0.5,38.9] | | Cr | Creatinine ratio | Numeric | [6,800] | | HbA1c | Glycated Haemoglobin - Blood glucose level | Numeric | [0.9,16] | | Chol | Cholesterol levels | Numeric | [0,10.3] | | TG | Triglycerides levels | Numeric | [0.3,13.8] | | HDL | High-Density Lipoprotein Cholesterol | Numeric | [0.2,9.9] | | LDL | Low-Density Lipoprotein Cholesterol | Numeric | [0.3,9.9] | | VLDL | Very Low Density Lipoprotein | Numeric | [0.1,35] | | BMI | Body Mass Index (Weight in Kg/Height in m) | Numeric | [19,47.75] | | CLASS | Yes, No or Predicted Diabetes | Characters | {Y,N,P} | | **Table 2:** Pima Indians Diabetes Dataset - Summary   |  |  |  |  | | --- | --- | --- | --- | | **Feature** | **Description** | **Data Type** | **Range** | | Pregnancies | No. of Pregnancies | Numeric | [0,17] | | Glucose | Glucose levels | Numeric | [0,199] | | Blood Pressure | Blood Pressure levels | Numeric | [0,122] | | Skin Thickness | Thickness of Skin | Numeric | [0,99] | | Insulin | Insulin levels | Numeric | [0,846] | | BMI | Body Mass Index | Numeric | [0,67.1] | | Diabetes Pedigree Function | Diabetes likelihood | Numeric | [0.078,2.42] | | Age | Age of the Patient | Numeric | [21,81] | | Outcome | 1 - Diabetic  0 - Non Diabetic | Boolean | {1,0} | |
| **Table 3:** Heart Failure Dataset - Data summary   |  |  |  |  | | --- | --- | --- | --- | | **Feature** | **Description** | **Data Type** | **Range** | | age | Age | Numeric | [40, 95] | | anaemia | Decrease of red blood cells or haemoglobin | Numeric | {0,1} | | creatinine\_phosphokinase | Level of the CPK enzyme in the blood (mcg/L) | Numeric | [23,7861] | | diabetes | Diabetic or Non Diabetic | Numeric | {0,1} | | ejection\_fraction | Percentage of blood leaving the heart at each contraction (percentage) | Numeric | [14,80] | | high\_blood\_pressure | High blood pressure - Hypertension | Numeric | {0,1} | | platelets | Platelets in blood (kilo platelets/mL) | Numeric | [25100,850000] | | serum\_creatinine | Level of serum creatinine in the blood (mg/dL) | Numeric | [0.5.9.4] | | serum\_sodium | Level of serum sodium in the blood (mEq/L) | Numeric | [113,148] | | sex | gender | Numeric | {0,1} | | smoking | Smokes or not | Numeric | {0,1} | | time | Follow-up period (days) | Numeric | [4,285] | | DEATH\_EVENT | If the patient deceased during the follow-up period | Boolean | {0,1} | | **Table 4:** Heart Disease Dataset - Data summary   |  |  |  |  | | --- | --- | --- | --- | | **Feature** | **Description** | **Data Type** | **Range** | | age | Age | Numeric | [29,77] | | sex | Gender | Boolean | {0,1} | | cp | CP levels | Numeric | {0,1,2,3} | | trestbps | Resting blood pressure | Numeric | [94,200] | | chol | Cholesterol levels | Numeric | [126,564] | | fbs | Fasting blood sugar | Boolean | {0,1} | | restecg | Patient's Resting ECG Levels | Numeric | {0,1,2} | | thalach | Maximum heart rate achieved | Numeric | [71,202] | | exang | Patient's Exang Levels | Boolean | {0,1} | | oldpeak | Patient's Old Peak History Recorded | Numeric | [0,6.2] | | slope | Slope levels | Numeric | {0,1,2} | | ca | CA levels | Numeric | {0,1,2,3,4} | | thal | Thal levels | Numeric | {0,1,2,3} | | target | 0 - Healthy  1 - Heart disease patient | Boolean | {0,1} | |
|  |  |

As may be seen from the data summaries given above, the dataset consisted mostly of numerical values except for a few rows, such as outcome and gender. Working with such inconsistencies in terms of the data type would lead to tedious amounts of work and codes to handle the data. This is why the data was then transformed by converting the binary outcomes, such as M/F or N/Y were made 0s and 1s respectively. The description of the dataset after data transformation has been shown below.

**Table 5:** Mendeley Data - Data description

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **ID** | **No\_Pation** | **Gender** | **Age** | **Urea** | **Cr** | **HbA1c** | **Chol** | **TG** | **HDL** | **LDL** | **VLDL** | **BMI** | **CLASS** |
| count | 1688 | 1688 | 1688 | 1688 | 1688 | 1688 | 1688 | 1688 | 1688 | 1688 | 1688 | 1688 | 1688 | 1688 |
| mean | 359.1665 | 862624.2612559241 | 0.490521 | 50.06694 | 5.03095 | 66.73934 | 6.716268 | 4.597168 | 2.055219 | 1.206872 | 2.628104 | 1.495438 | 26.59782 | 0.5 |
| std | 247.8909 | 7528272 | 0.500058 | 9.868652 | 2.868203 | 50.39081 | 2.769452 | 1.311082 | 1.347467 | 0.591841 | 1.058108 | 3.005186 | 5.295642 | 0.500148 |
| min | 1 | 123 | 0 | 20 | 0.5 | 6 | 0.9 | 0 | 0.3 | 0.2 | 0.3 | 0.1 | 19 | 0 |
| 25% | 143 | 34231.75 | 0 | 44 | 3.6 | 46 | 4.7 | 3.8 | 1.2 | 0.9 | 1.8 | 0.6 | 22 | 0 |
| 50% | 315 | 34327.5 | 0 | 51 | 4.5 | 58 | 5.5 | 4.4 | 1.7 | 1.1 | 2.5 | 0.8 | 24 | 0.5 |
| 75% | 600.25 | 45392 | 1 | 56 | 5.7 | 73 | 8.8 | 5.2 | 2.4 | 1.3 | 3.3 | 1.2 | 30 | 1 |
| max | 800 | 75435657 | 1 | 79 | 38.9 | 800 | 16 | 10.3 | 13.8 | 9.9 | 9.9 | 35 | 47.75 | 1 |

**Table 6:** Pima Indians Diabetes Dataset - Data summary

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Pregnancies** | **Glucose** | **BloodPressure** | **SkinThickness** | **Insulin** | **BMI** | **DPF** | **Age** | **Outcome** |
| count | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 |
| mean | 3.845052 | 120.894531 | 69.105469 | 20.536458 | 79.799479 | 31.992578 | 0.471876 | 33.240885 | 0.348958 |
| std | 3.369578 | 31.972618 | 19.355807 | 15.952218 | 115.244002 | 7.884160 | 0.331329 | 11.760232 | 0.476951 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.078000 | 21.000000 | 0.000000 |
| 25% | 1.000000 | 99.000000 | 62.000000 | 0.000000 | 0.000000 | 27.300000 | 0.243750 | 24.000000 | 0.000000 |
| 50% | 3.000000 | 117.000000 | 72.000000 | 23.000000 | 30.500000 | 32.000000 | 0.372500 | 29.000000 | 0.000000 |
| 75% | 6.000000 | 140.250000 | 80.000000 | 32.000000 | 127.250000 | 36.600000 | 0.626250 | 41.000000 | 1.000000 |
| max | 17.000000 | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 | 2.420000 | 81.000000 | 1.000000 |

**Table 7:** Heart Failure Dataset - Data summary

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **age** | **anaemia** | **creatinine\_phosphokinase** | **diabetes** | **ejection\_fraction** | **high\_blood\_pressure** | **platelets** | **serum\_creatinine** | **serum\_sodium** | **sex** | **smoking** | **time** | **DEATHEVENT** |
| count | 299.000000 | 299.000000 | 299.000000 | 299.000800 | 299.000000 | 299.000000 | 299.000000 | 299.00000 | 299.000000 | 299.000000 | 299.00000 | 299.000000 | 299.00000 |
| mean | 60.833893 | 0.431438 | 581.839465 | 0.418060 | 38.083612 | 0.351171 | 263358.029264 | 1.39388 | 136.625418 | 0.648829 | 0.32107 | 130.260870 | 0.32107 |
| std | 11.894809 | 0.496107 | 970.287881 | 0.494067 | 11.834841 | 0.478136 | 97804.236869 | 1.03451 | 4.412477 | 0.478136 | 0.46767 | 77.614208 | 0.46767 |
| min | 40.000000 | 0.000000 | 23.000000 | 0.000000 | 14.000000 | 0.000000 | 25100.000000 | 0.50000 | 113.000000 | 0.000000 | 0.00000 | 4.000000 | 0.00000 |
| 25% | 51.000000 | 0.000000 | 116.500000 | 0.000000 | 30.000000 | 0.000000 | 212500.000000 | 0.90000 | 134.000000 | 0.000000 | 0.00000 | 73.000000 | 0.00000 |
| 50% | 60.000000 | 0.000000 | 250.000000 | 0.000000 | 38.000000 | 0.000000 | 262000.000000 | 1.10000 | 137.000000 | 1.000000 | 0.00000 | 115.000000 | 0.00000 |
| 75% | 70.000000 | 1.000000 | 582.000000 | 1.000000 | 45.000000 | 1.000000 | 303500.000000 | 1.40000 | 140.000000 | 1.000000 | 1.00000 | 203.000000 | 1.00000 |
| max | 95.000000 | 1.000000 | 7861.000000 | 1.000000 | 80.000000 | 1.000000 | 850000.000000 | 9.40000 | 148.000000 | 1.000000 | 1.00000 | 285.000000 | 1.00000 |

**Table 8:** Heart Disease Dataset - Data summary

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **age** | **sex** | **cp** | **trestbps** | **chol** | **fbs** | **restecg** | **thalach** | **exang** | **oldpeak** | **slope** | **ca** | **thal** | **target** |
| count | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 |
| mean | 54.366337 | 0.683168 | 0.966997 | 131.623762 | 246.264026 | 0.148515 | 0.528053 | 149.646865 | 0.326733 | 1.039604 | 1.399340 | 0.729373 | 2.313531 | 0.544554 |
| std | 9.082101 | 0.466011 | 1.032052 | 17.538143 | 51.830751 | 0.356198 | 0.525860 | 22.905161 | 0.469794 | 1.161075 | 0.616226 | 1.022606 | 0.612277 | 0.498835 |
| min | 29.000000 | 0.000000 | 0.000000 | 94.000000 | 126.000000 | 0.000000 | 0.000000 | 71.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 47.500000 | 0.000000 | 0.000000 | 120.000000 | 211.000000 | 0.000000 | 0.000000 | 133.500000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 2.000000 | 0.000000 |
| 50% | 55.000000 | 1.000000 | 1.000000 | 130.000000 | 240.000000 | 0.000000 | 1.000000 | 153.000000 | 0.000000 | 0.800000 | 1.000000 | 0.000000 | 2.000000 | 1.000000 |
| 75% | 61.000000 | 1.000000 | 2.000000 | 140.000000 | 274.500000 | 0.000000 | 1.000000 | 166.000000 | 1.000000 | 1.600000 | 2.000000 | 1.000000 | 3.000000 | 1.000000 |
| max | 77.000000 | 1.000000 | 3.000000 | 200.000000 | 564.000000 | 1.000000 | 2.000000 | 202.000000 | 1.000000 | 6.200000 | 2.000000 | 4.000000 | 3.000000 | 1.000000 |

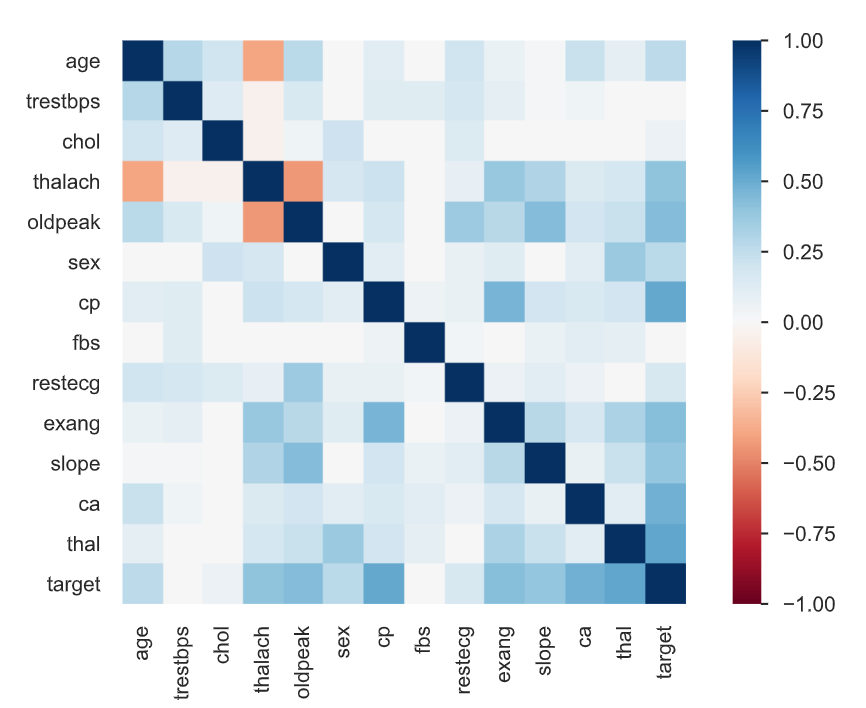
It can be said that the diabetes datasets consist of a good variety of data from both the datasets consisting of data collected from almost 2400 patients. Such a large amount of data can aid in evaluating the correct percentage of accuracy of the developed and implemented methods. The cardiac problems’ datasets, on the other hand, may not be as large in comparison to diabetes, but do consist of a good amount of data collected from patients of diverse origins.

The datasets also consist of 8-12 relevant features each apart from the outcome which is a good number of factors which should be analysed to aid in prediction of diabetes and cardiovascular diseases.

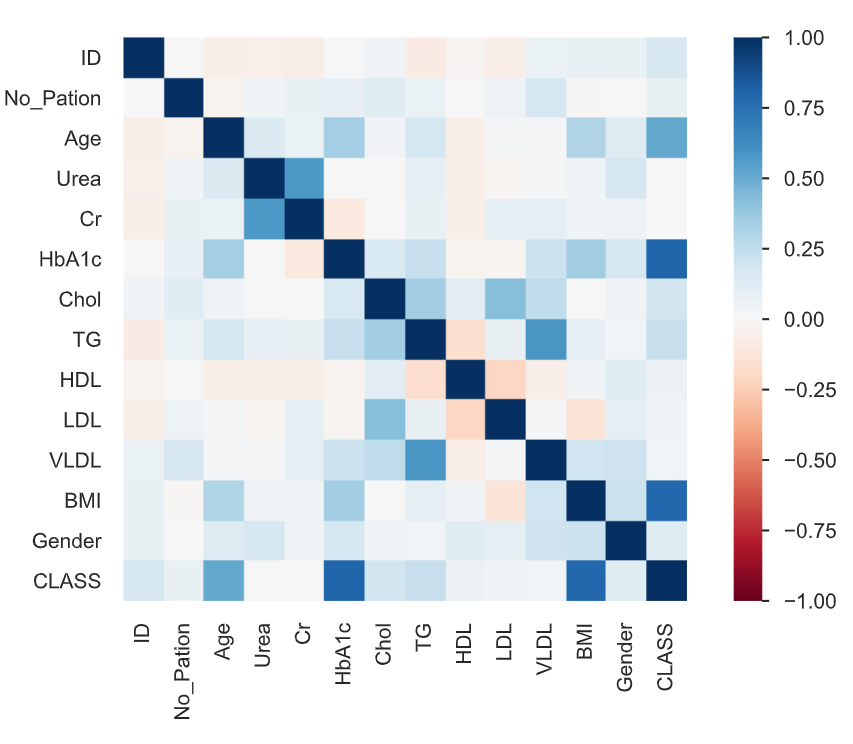
**Data Visualization**

In order to get a deeper and better understanding of the datasets and the correlations between various features, heatmaps, bar and dot plots have been used. Such plots of correlations between features and trends in dot plots, for instance, prove very useful in techniques such as feature selection, engineering, etc. The following visualisations were obtained of the datasets and its features:

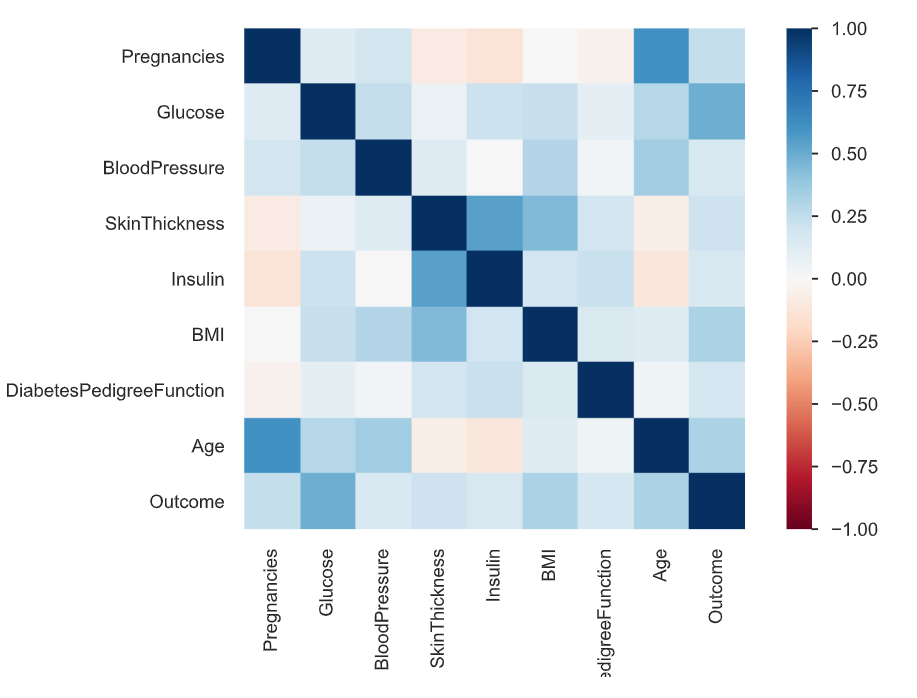




*Fig 6: Identifying correlations in the Heart disease dataset*



*Fig 3: Identifying correlations in Mendeley’s Diabetes Dataset*



*Fig 4: Identifying Correlations in Pima Indians Diabetes Dataset*

Identifying Correlations:

The datasets were plotted on a heat map to identify correlations among different features and the following was observed:

Age - CLASS High correlation

Urea - Cr High correlation

Cr - Urea High correlation

HbA1c - CLASS High correlation

TG - VLDL High correlation

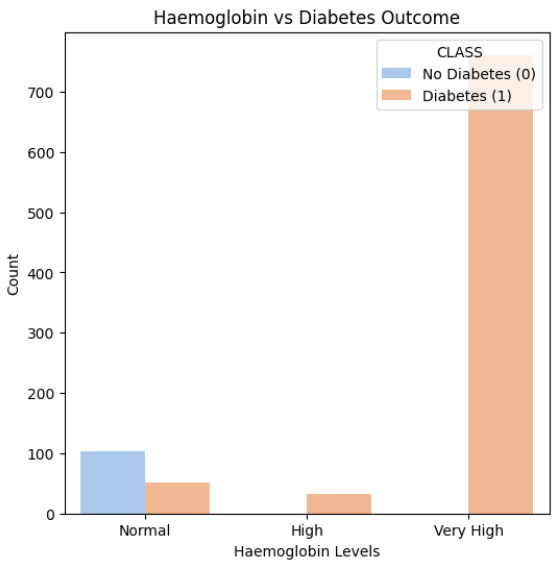
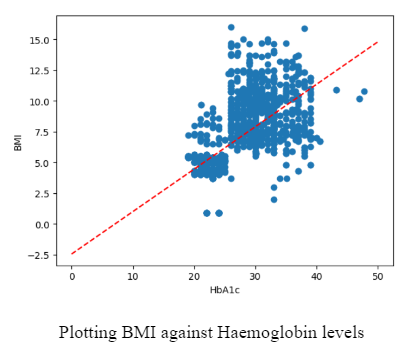
VLDL - TG High correlation

BMI - CLASS High correlation

CLASS - Age High correlation

Similarly, the correlation of the remaining datasets were noted as well.

The datasets were further graphed among different features to display and observe certain trends, such as the given figures obtained from plotting count of diabetic or non diabetic patients against haemoglobin levels as they have a high correlation as can be seen from the figure provided in the previous page. Such graphs are crucial in learning from the dataset in terms of the the range of data available, the extent of credibility and in determining the extent upto which techniques need to be applied to remove anomalies or noise of the data.

Fig 7: Bar plot of count of diabetic and non diabetic patients Fig 8: Dot plot of Haemoglobin levels against BMI

with their haemoglobin values (were highly correlated)

Identifying correlations in the Pima Indian Diabetes Dataset

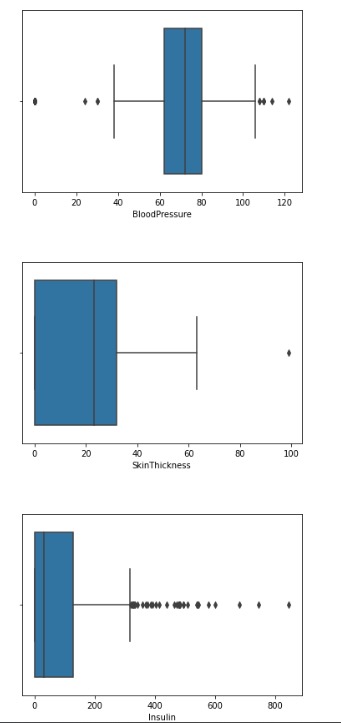
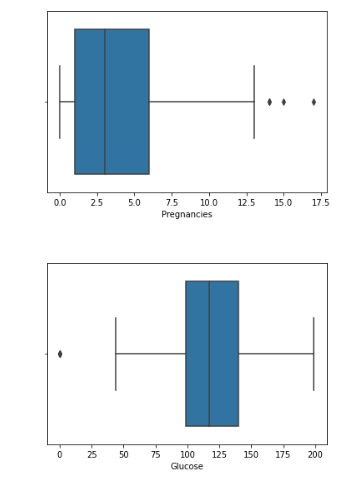
Anomalies remained another concern for which the following box plots were graphed,

Fig 9: Box plots of features in the Pima Indian Diabetes Dataset

**Data Pre-processing**

**1)Standard Scalar**: Firstly, Standard Scaler was used on the dataset, a data preprocessing technique used to standardise the features of a dataset. This approach scales the features so that they have a mean of zero and a standard deviation of one. This is useful when the features of the dataset have different scales. In the Mendeley dataset, the scales had variations in the ranges, which is why the Standardization approach was used as we wanted to bring them to the same scale.

*Where:*

*z is the standardised value of a data point.*

*x is the original value of the data point.*

*μ is the mean (average) of the feature.*

*σ is the standard deviation of the feature.*

**2)Principal Component Analysis (PCA)**: The next pre-processing technique used was Principal Component Analysis (PCA) which is a statistical technique used to reduce the dimensionality of a dataset while retaining most of the original information. This technique is widely used when the dataset is large and has large dimensionality as was in the Mendeley Dataset. It works by identifying a set of orthogonal axes, called principal components, that capture the maximum variance in the data. The principal components are linear combinations of the original variables in the dataset and are ordered in decreasing order of importance.

**3)Smoteen/SMOTE: SMOTE** - stands for Synthetic Minority Over-sampling Technique, and SMOTEEN (SMOTE-ENN) are two popular techniques used to handle imbalanced datasets in machine learning. Both techniques are oversampling methods that generate synthetic samples of the minority class to balance the class distribution. SMOTE generates synthetic samples by interpolating between existing minority class samples, while SMOTEEN combines SMOTE with ENN (Edited Nearest Neighbors) to remove noisy samples from the majority class.

**Selected Models:**

**1)Logistic Regression:** Logistic Regression is a supervised machine learning algorithm that is mainly used for classification tasks where the goal is to predict the probability that an instance belongs to a given class or not. It is a kind of statistical algorithm that analyses the relationship between a set of independent variables and the dependent binary variables. Logistic Regression is a powerful tool for decision-making, for example, email spam or not.

**2)** **Random Forest:** A Random Forest is an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time. The output of the random forest is the class selected by most trees.The random forest algorithm is a commonly-used machine learning algorithm that combines the output of multiple decision trees to reach a single result. It handles both classification and regression problems and is known for its ease of use and flexibility.

**3)SVM:** Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for classification or regression tasks. The main idea behind SVMs is to find a hyperplane that maximally separates the different classes in the training data. This is done by finding the hyperplane that has the largest margin, which is defined as the distance between the hyperplane and the closest data points from each class. Once the hyperplane is determined, new data can be classified by determining on which side of the hyperplane it falls. SVMs are particularly useful when the data has many features, and/or when there is a clear margin of separation in the data.

**4)KNN: K-Nearest Neighbour (KNN)** is a non-parametric machine learning algorithm that can be used for both classification and regression tasks. It is based on the principle of similarity, where similar data points are grouped together. The algorithm stores all available cases and classifies new cases based on their similarity to the available cases.The KNN algorithm works by selecting the k nearest neighbours to a given data point and classifying it based on the majority class of these neighbours. The value of k is usually chosen by cross-validation or other methods.

**5)Naive Bayes (Bernouilli):** Bernoulli Naive Bayes is a variant of the Naive Bayes algorithm that is used for discrete data and works on the Bernoulli distribution. The main feature of Bernoulli Naive Bayes is that it accepts features only as binary values like true or false, yes or no, success or failure, 0 or 1, and so on.The algorithm is based on the principle of conditional probability and assumes that the features are independent of each other. It is commonly used in text classification tasks such as spam filtering, sentiment analysis, and document categorization.

**6)AdaBoost** - AdaBoost is an ensemble learning technique that boosts the performance of weak learners by sequentially adjusting the sample weights, emphasising misclassified instances, and combining multiple weak classifiers into a strong classifier. It is effective in binary classification problems and applications like face detection and text categorization.

**7)Neural Network** - A neural network, or artificial neural network (ANN), is a machine learning model inspired by the brain's structure. It consists of interconnected neurons organised into layers with weighted connections and activation functions. Neural networks, especially deep learning models, are widely used for tasks such as image recognition, natural language processing, and reinforcement learning. They are trained to minimise the difference between predictions and actual targets, making them versatile and integral in various fields.

**Results of Machine Learning Algorithms**

In this research paper, out of the seven models implemented, the models which gave the best results were Logistic regression, Random forest, SVM and Neural Networks. In addition it may be mentioned that for all the models, the same dataset and pre-processing techniques were used as a control environment to compare the results from different models applied.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 9:** Results obtained from Mendeley’s Dataset   |  |  | | --- | --- | | **Machine Learning Algorithm** | **Accuracy** | | **Logistic Regression** | **1.00** | | Random Forest | 0.99 | | **SVM** | **1.00** | | K-Nearest Neighbour | 0.99 | | Naive-Bayes (Bernoulli) | 0.89 | | AdaBoost | 0.97 | | Neural Network | 0.98 | | **Table 10:** Results obtained from Pima Diabetes Dataset   |  |  | | --- | --- | | **Machine Learning Algorithm** | **Accuracy** | | Logistic Regression | 0.84 | | **Random Forest** | **0.91** | | SVM | 0.89 | | K-Nearest Neighbour | 0.85 | | Naive-Bayes (Bernoulli) | 0.73 | | AdaBoost | 0.80 | | Neural Network | 0.90 | |
| **Table 11:** Results obtained from Heart Failure Dataset   |  |  | | --- | --- | | **Machine Learning Algorithm** | **Accuracy** | | Logistic Regression | 0.90 | | Random Forest | 0.86 | | **SVM** | **0.95** | | K-Nearest Neighbour | 0.90 | | Naive-Bayes (Bernoulli) | 0.90 | | AdaBoost | 0.90 | | Neural Network | 0.90 | | **Table 12:** Results obtained from Heart Disease Dataset   |  |  | | --- | --- | | **Machine Learning Algorithm** | **Accuracy** | | Logistic Regression | 0.98 | | Random Forest | 0.93 | | SVM | 0.95 | | K-Nearest Neighbour | 0.98 | | Naive-Bayes (Bernoulli) | 0.95 | | AdaBoost | 0.98 | | **Neural Network** | **1.00** | |

The result will be further tested in terms of precision, recall, f1 Score and AUC:

**Accuracy:**

**Precision:**

**Recall:**

**F1 Score:**

**Precision** - Precision is a metric in machine learning that measures the accuracy of positive predictions made by a model. It quantifies the proportion of true positive predictions relative to all positive predictions (true positives and false positives). High precision indicates that when the model predicts a positive outcome, it is likely to be correct, making it particularly valuable when minimising false positives is important, such as in medical diagnoses.

**Recall** - Recall, also known as sensitivity or true positive rate, is a metric that assesses a model's ability to identify all positive instances in the dataset. It measures the proportion of true positives relative to all actual positive instances (true positives and false negatives). High recall suggests that the model is effective at capturing most of the positive cases, making it essential in scenarios where missing positives is costly, such as search and rescue operations.

**F1 Score** - The F1 Score is a metric that combines precision and recall into a single value. It provides a balanced measure of a model's performance by considering both false positives and false negatives. The F1 Score is the harmonic mean of precision and recall and is useful when you want to strike a balance between precision and recall.

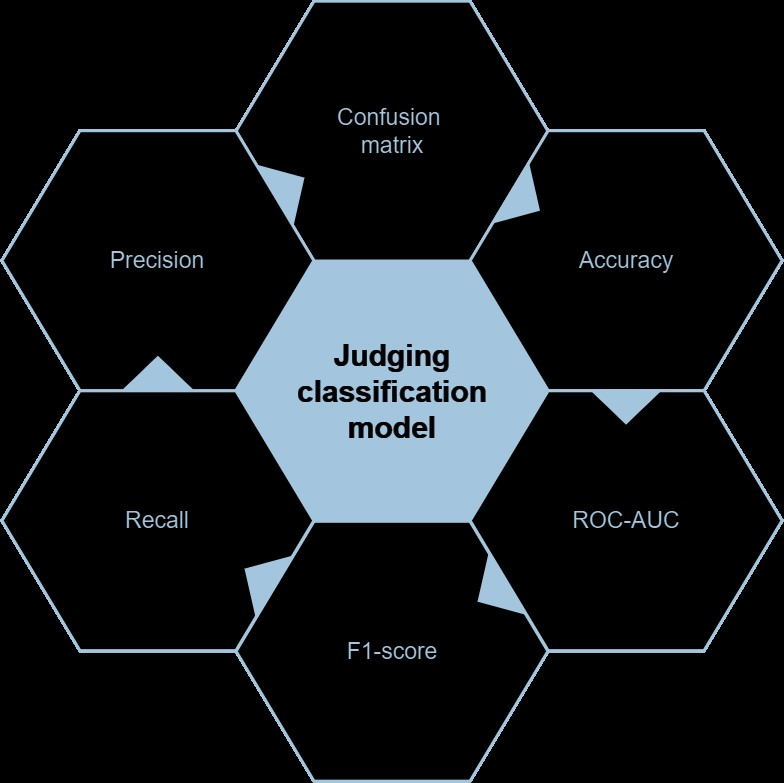
**ROC (Receiver Operating Characteristic)** - ROC is a graphical representation of a model's performance across different discrimination thresholds. It plots the true positive rate (recall) against the false positive rate as the threshold for classifying positive and negative instances varies. The area under the ROC curve (AUC-ROC) quantifies the overall performance of a model, with higher values indicating better discrimination. ROC curves are often used to evaluate and compare the performance of classification models, especially in scenarios where the trade-off between false positives and false negatives needs to be analyzed.

Fig 10: Evaluation models used

**Table 13:** Mendeley’s Diabetes Dataset Model Evaluation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy(%)** | **Precision(%)** | **Recall(%)** | **F1 Score(%)** | **ROC AUC(%)** |
| Logistic Regression | 100 | 100 | 100 | 100 | 100 |
| Random Forest | 99.09 | 99.38 | 98.76 | 99.07 | 99.08 |
| SVM | 99.70 | 99.38 | 100 | 99.69 | 99.71 |
| K-Nearest Neighbour | 99.09 | 99.38 | 98.76 | 99.07 | 99.08 |
| Naive Bayes(Bernoulli) | 89.12 | 99.21 | 78.26 | 87.50 | 88.84 |
| AdaBoost | 96.98 | 95.76 | 98.14 | 96.93 | 97.01 |
| Neural Network | 98.79 | 98.16 | 99.38 | 98.77 | 98.81 |
|  |  |  |  |  |  |

**Table 14:** PIMA Indian Diabetes Dataset Model Evaluation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy(%)** | **Precision(%)** | **Recall(%)** | **F1 Score(%)** | **ROC AUC(%)** |
| **Logistic Regression** | 83.81 | 89.83 | 82.81 | 86.18 | 84.09 |
| **Random Forest** | 90.48 | 89.71 | 95.31 | 92.42 | 89.12 |
| **SVM** | 88.57 | 88.24 | 93.75 | 90.91 | 87.12 |
| **K-Nearest Neighbour** | 84.76 | 85.29 | 90.62 | 87.88 | 83.12 |
| **Naive Bayes(Bernouilli)** | 73.33 | 84.62 | 68.75 | 75.86 | 74.62 |
| **AdaBoost** | 80.00 | 86.44 | 79.69 | 82.93 | 80.09 |
| **Neural Network** | 89.52 | 90.77 | 92.19 | 91.47 | 88.78 |

**Table 15:** Heart Failure Dataset Model Evaluation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy(%)** | **Precision(%)** | **Recall(%)** | **F1 Score(%)** | **ROC AUC(%)** |
| **Logistic Regression** | 90.48 | 84.62 | 100 | 91.67 | 90.00 |
| **Random Forest** | 85.71 | 78.57 | 100 | 88.00 | 85.00 |
| **SVM** | 95.24 | 91.67 | 100 | 95.65 | 95.00 |
| **K-Nearest Neighbour** | 90.48 | 84.62 | 100 | 91.67 | 90.00 |
| **Naive Bayes(Bernoulli)** | 90.48 | 86.62 | 100 | 91.67 | 90.00 |
| **AdaBoost** | 90.48 | 86.62 | 100 | 91.67 | 90.00 |
| **Neural Network** | 90.48 | 86.62 | 100 | 91.67 | 90.00 |

**Table 16:** Heart Disease Dataset Model Evaluation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy(%)** | **Precision(%)** | **Recall(%)** | **F1 Score(%)** | **ROC AUC(%)** |
| **Logistic Regression** | 97.73 | 100 | 95.65 | 97.78 | 97.83 |
| **Random Forest** | 95.45 | 100 | 91.30 | 95.45 | 95.65 |
| **SVM** | 95.45 | 100 | 91.30 | 95.45 | 95.65 |
| **K-Nearest Neighbour** | 97.73 | 100 | 95.65 | 97.78 | 97.83 |
| **Naive Bayes(Bernoulli)** | 95.45 | 100 | 91.30 | 95.45 | 95.65 |
| **AdaBoost** | 97.73 | 100 | 95.65 | 97.78 | 97.83 |
| **Neural Network** | 97.73 | 100 | 95.65 | 97.78 | 97.83 |

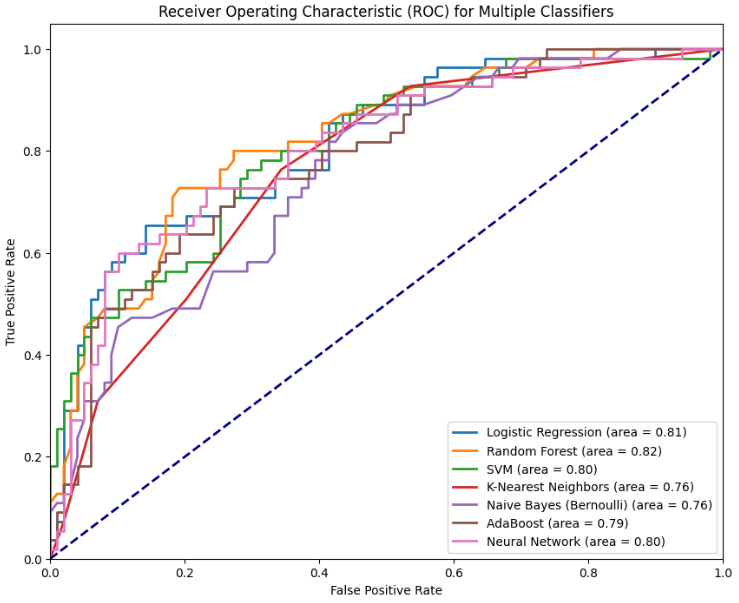
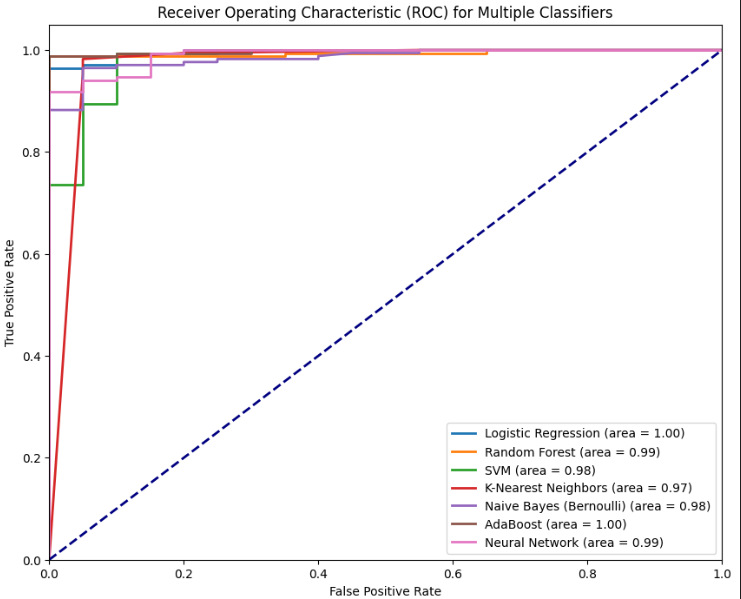
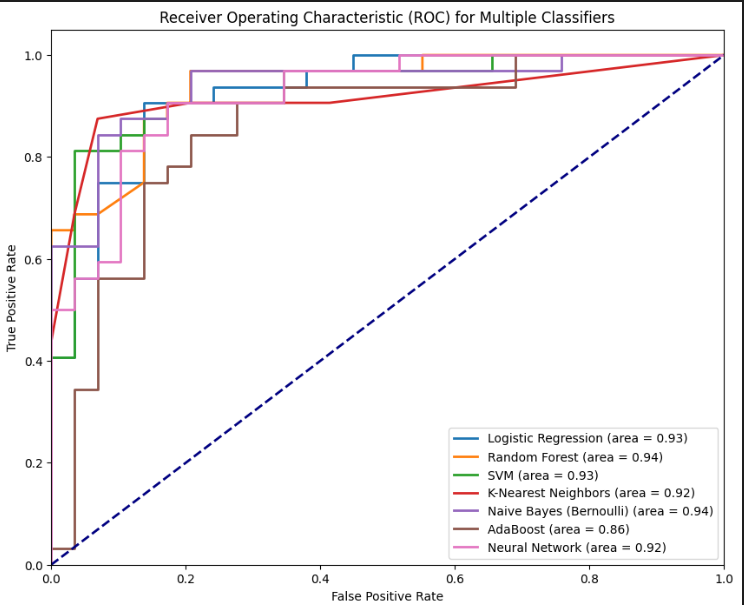


Fig 11: ROC Curve for Mendeley’s Diabetes Dataset Fig 12: ROC Curve for Pima Indian Diabetes Dataset 

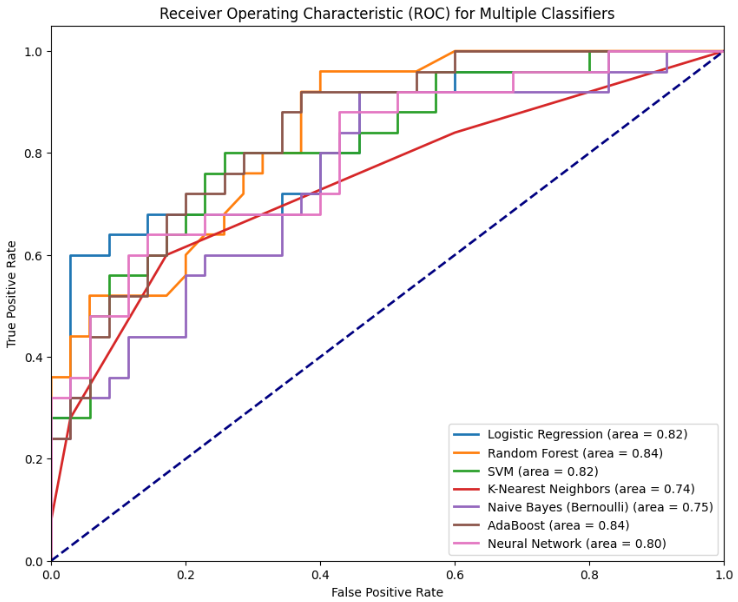


Fig 13: ROC Curve for Heart Failure Dataset Fig 14: ROC Curve for Heart Disease Dataset

A confusion matrix is a critical tool for evaluating the performance of classification models in machine learning. It presents a structured summary of the model's predictions by distinguishing between true positives (correct positive predictions), true negatives (correct negative predictions), false positives (incorrect positive predictions), and false negatives (incorrect negative predictions). This tabular representation allows for the calculation of essential performance metrics like accuracy, precision, recall, and specificity, helping assess the model's ability to classify data accurately and identify the types of errors it makes in a clear and concise manner.

**Table 17: Confusion Matrix Template:**

|  |  |  |
| --- | --- | --- |
|  | Actual Positive | Actual Negative |
| Predicted positive | True positive | False positive |
| Predicted negative | False negative | True negative |

The confusion matrix for each model used for each dataset is shown below:

Fig 15: Mendeley’s Diabetes Dataset Confusion Matrices

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | | | |

Fig 16: Pima Indians Diabetes Dataset Confusion Matrices

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | | | |

Fig 17: Heart Failure Dataset Confusion Matrices

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | | | |

Fig 18: Heart Diseases Dataset Confusion Matrices

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | | | |

Learning curves provide insights into a machine learning model's training and generalisation performance. They help assess bias, variance, convergence, data sufficiency, and guide decisions related to early stopping, hyperparameter tuning, and model selection.

The following Learning Curves were obtained:

Fig 19: Mendeley’s Diabetes Dataset Learning Curves

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

Fig 20: Pima Indians Diabetes Dataset Learning Curves

|  |  |
| --- | --- |
|  |  |
|  |  |

Fig 21: Heart Failure Dataset Learning Curves

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

Fig 22: Heart Disease Dataset Learning Curves

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

**Discussion and conclusion**

This project delineates the seven models used in the study and makes use of them to accurately predict development of diabetes, heart failure and heart diseases’ presence. The plotted graphs and charts depict that the models were adequately trained and gave good accuracy. SVM, logistic regression, neural networks and KNN outperformed other models used.   
The evaluation of the first dataset shows signs of overfitting as the accuracy is very high, that is, instead of having the model learning from the dataset and being able to generalise it, the model seems to have learnt the dataset and its outcomes.

However, in the other datasets, the models have a good accuracy overall and keep varying from between the 80s and 90s percentage range.

Future research projects will be conducted in a similar manner but with implementation of combinations of models using ensemble methods, hyperparameter tuning, clustering and feature selection to work upon the prediction performance in this domain in addition to rectifying possibilities of overfitting of data.