Multiple Disease Prediction using Machine Learning and Streamlit

Yashwanth Naik Bhukya, Tejesh Chintada

Computer Science and Engineering Department,
Indian Institute of Information Technology, Nagpur

Abstract- There are many existing machine learning models related to health care which mainly focuses on detecting only one disease. Therefore, this study has developed a system to forecast several diseases by using a single user interface. The proposed model can predict multiple diseases such as diabetes, heart disease and chronic kidney disease. If left untreated, these diseases pose a risk to humanity. As a result, many lives can be saved by early detection and diagnosis of these disorders. This research work attempts to implement various classification algorithms (K-Nearest Neighbor, Logistic Regression and Gaussian naive bayes.) to perform disease prediction. The characteristics needed to train and evaluate the models are extracted after preprocessing the dataset. To examine the prediction capabilities, we use accuracy scores and confusion matrices to evaluate the performance of each model. We hope to find the most effective machine learning model for predicting all three diseases through experimentation and analysis. Our findings show that machine learning algorithms have the capacity to effectively forecast all the diseases. The accuracy of each algorithm is validated and compared with each other to find the best one for prediction. Furthermore, multiple datasets (for each disease each dataset) are used to achieve utmost accuracy in the predicted results. When ranking the performance of the models, numerous other criteria, such as the F1-score, accuracy, precision, and recall, were utilized. The main goal is to create a web application capable of forecasting several diseases by using machine learning, including diabetes, heart disease and chronic kidney disease.

1. INTRODUCTION

Heart disease is a primary cause of death and morbidity around the world, offering considerable challenges to public health systems. Early and accurate identification of cardiac disease improves patient outcomes by allowing for timely intervention and therapy. With the increased availability of electronic health data and advances in machine learning techniques, there is a growing interest in using these technologies to construct accurate prediction models for heart disease.

One of the diseases that is constantly spreading and targeting even young people is diabetes and is reported to have increased to 592 million. Diabetes is a metabolic illness that causes the body to behave abnormally, with fluctuating blood glucose levels brought on by pancreatic failure that results in little to no insulin production in the patient's body The root cause of diabetes remains unknown; However, the environment and lifestyle play a significant role in disease development. Despite the fact that it is a fatal disease, treatment and medication are available to treat it. In order to understand diabetes, we need to understand how the body normally uses glucose. Our bodies break down the food we eat, especially the carbs, and convert them to sugar or glucose. Now, the pancreas is supposed to release insulin, which unlocks the cells in the body. Consequently, glucose is able to enter cells and supply the body with energy. However, this approach does not function for diabetic patients. Nowadays, machine learning algorithms are widely used in many sectors and also have shown promising results in the field of medical applications and disease detection.

This study aims to predict multiple diseases including diabetes, heart disease, and chronic kidney disease using various classification algorithms such as K-NN, Gaussian NB, and Logistic Regression. The accuracy of each algorithm is validated and compared to determine the best one for prediction. Multiple datasets are utilized to achieve the highest accuracy in the predicted results. The best-performing algorithm for each disease is chosen and integrated to build a web application where users can easily predict the required disease by entering respective attribute values.

For heart disease prediction, datasets such as the Cleveland, Hungary, Switzerland, and Long Beach V databases are used. These datasets contain 76 attributes, including the target attribute indicating the presence (1) or absence (0) of heart disease. Only a selection of 14 attributes is used in this research. The dataset for chronic kidney disease consists of 25 features and was collected over a 2-month period in India. Attributes such as red blood cell count, white blood cell count, etc., are included. The classification is binary: "ckd" (chronic kidney disease) or "notckd". This dataset contains a total of 400 records.

For diabetes prediction, datasets such as the Pima Indians Diabetes Database and a Kaggle dataset are utilized. These datasets include predictor variables such as BMI, insulin level, age, number of previous pregnancies, and more. The outcome variable "Outcome" indicates the presence (1) or absence (0) of diabetes. The combined dataset consists of 769 records and 9 columns.

Data preprocessing techniques such as Label Encoding are applied, and models are created using K-NN, Gaussian NB, and Logistic Regression algorithms. For each disease, the dataset is split into training and testing sets, and each model is trained against its training dataset. After evaluating each model against the testing dataset, the accuracy of the model is calculated and compared to select the best-performing one.

The web application's UI includes features like a sidebar for navigation and forms to input attribute values for a particular disease. The sidebar is created using the option menu method of the Streamlit library. Input value fields in the forms are generated using the text input method of Streamlit. Trained models (classifiers) are loaded into the Streamlit editor using the pickle module's load method.

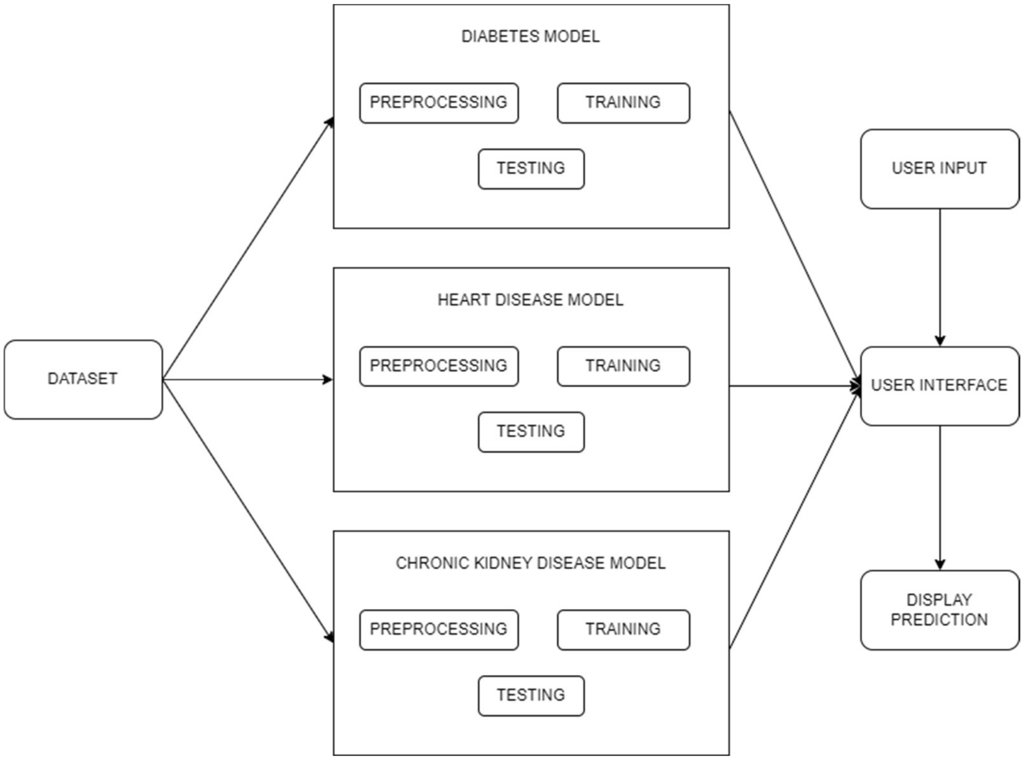
Additionally, the application includes a test result button for each disease prediction page to trigger the prediction function. The disease prediction process is carried out through the web application using the built-in prediction function of Streamlit.

2. LITERATURE REVIEW

Paper	Research Objective	Research Methodology	Key Findings
Laxmi Deepthi Gopisetti (2023 IEEE)	Multiple Disease Prediction System using Machine Learning and Streamlit. Diseases are diabetes, heart disease and chronic kidney disease.	Machine learning models using algorithms such as Naive Bayes, KNN, Random Forest, Logistic Regression and SVM. Streamlit is used for deploying user interface.	Observed more accuracy for Random forest for Diabetes, for heart disease more accuracy for SVM and for chronic kidney disease more accuracy for Random Forest.
Bhavesh Rathi (2023 IEEE)	Early prediction of diabetes disease using machine learning techniques.	In this paper KNN machine learning approach is used because it gives perfect estimation of the dataset.	Performed KNN method for different K-values such as k = 10,7,5,3 and got highest accuracy for k=5.
Abdul Hafiz (2023 IEEE)	Heart disease prediction based on machine learning technique.	The machine learning algorithms used are support vector machine (SVM), decision tree, random forest, naive bayes, logistic regression.	Among all models Logistic Regression got highest accuracy of 91.8 and second highest is for support vector of 90.16.
Srishti Mahajan (2023 IEEE)	Diabetes mellitus prediction using supervised machine learning techniques.	machine learning models used are logistic regression and random forest.	It is observed that random forest got highest accuracy of 99 and logistic regression got accuracy of 94.
Hemalatha (2023 IEEE)	Extensive review on predicting heart disease using machine learning and deep learning techniques.	Algorithms used for machine learning such as random forest and CNN. Artificial Neural Networks are used in Deep Learning to carry out complex calculations on vast volumes of data.	Observed comparative result study on different algorithms and datasets. Cleveland dataset got highest accuracy by using random forest.

3. PROPOSED METHODOLOGY

Workflow of proposed work



Step 1: Collecting datasets from various sources. For this project datasets for different diseases i.e., heart disease, diabetes, and chronic kidney is collected from Kaggle dataset

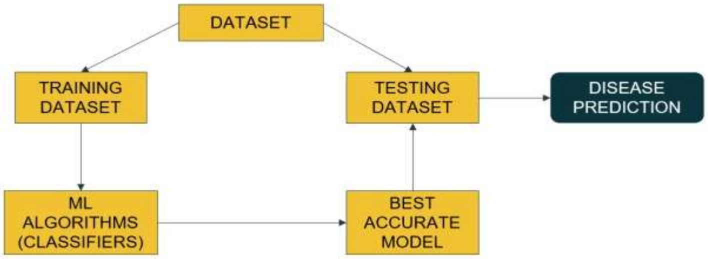
Step 2: Performing data preprocessing techniques such as Label encoding. Label encoding helped in converting the categorical data such as gender, appetite into numerical data in the form zero's and one's.

Step 3: Creating models using various machine learning algorithms such as K-NN, Gaussian NB, and Logistic Regression. For each disease various algorithms are used to create classifier models

Step 4: Training all the models against datasets. For each disease the dataset is split into two parts i.e., training set and testing set and each model is trained against its training dataset

Step 5: Evaluating the models using metrics such as accuracy score. After evaluating each model of a particular disease against testing dataset, the model's accuracy is calculated.

The below figure illustrates the utilization of various machine learning algorithms, including Naïve Bayes, K-NN, Logistic Regression, for predicting multiple diseases. This approach aims to bridge the gap between patients and healthcare providers, allowing each to pursue their objectives effectively. The accuracy of each algorithm is assessed and compared to determine the most suitable one for prediction. Integration of multiple datasets is performed to enhance the accuracy of the predicted results. To enhance user experience, a web application has been developed, enabling users to easily predict specific diseases by inputting the respective attribute values.



4. EXPERIMENTAL SETUP

A. MACHINE LEARNING MODELS

KNN: The K-nearest Neighbors algorithm is a versatile supervised learning tool used for both classification and regression tasks. It operates on the idea of grouping by placing new data into a category that matches existing data based on similarity measurement. This algorithm finds neighboring points that are similar and categorizes new data points accordingly. Due to its reliance on identifying similar neighboring points, it is commonly used as a categorization strategy.

LOGISTIC REGRESSION: Logistic Regression is a classification method that operates by determining probabilities and performs classification based on these probabilities. It uses the Sigmoid Function to calculate the probability of a row belonging to a class, which is between 0 and 1. The function takes the product of theta transpose and the parameter vector as input to compute the probability. Classification is determined based on a threshold value; if the probability is less than the threshold, one class is assigned, and if it's higher, another class is assigned.

NAÏVE BAYES: Gaussian Naive Bayes is a straightforward method for building classifiers. It assigns class labels to instances based on vectors of feature values selected from a finite set. There isn't a single technique for training these classifiers, but rather a family of algorithms that share a common principle: given the class variable, all naive Bayes classifiers assume that the value of a certain feature is independent of the value of any other feature.

B. STREAMLIT

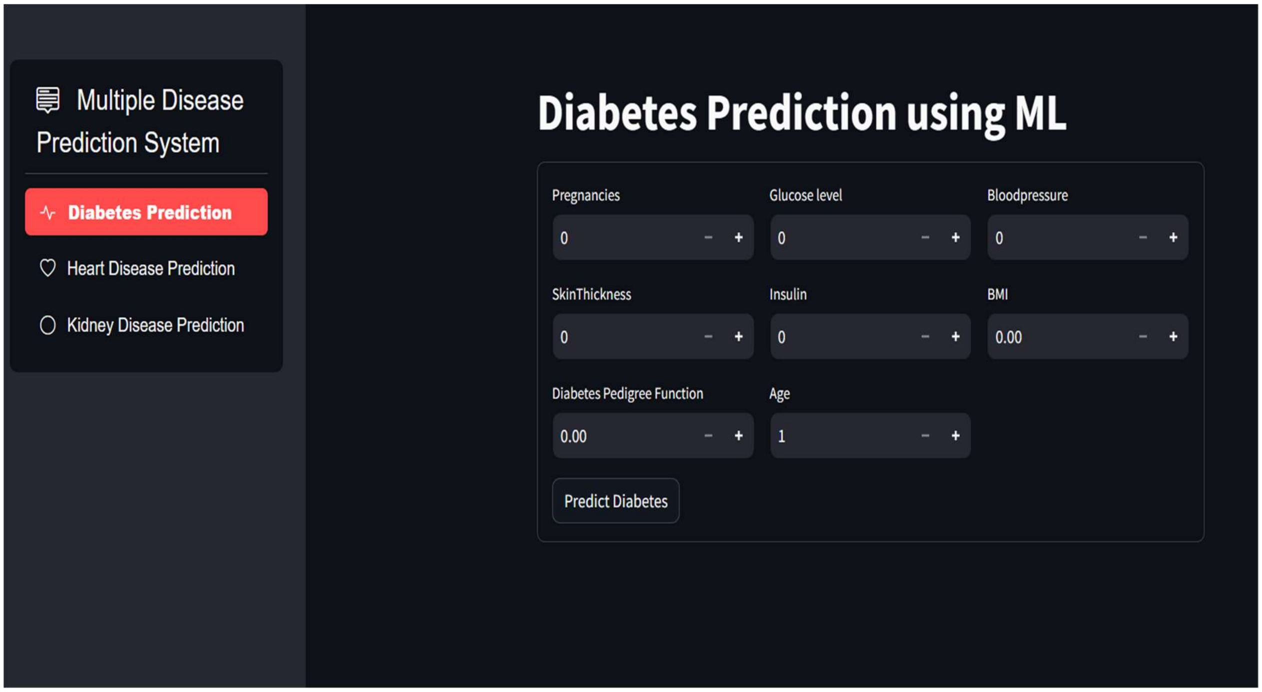
- HTML, CSS, and JavaScript are not required.

- Unlike spending days or months on building a web app, we can create impressive machine learning or data science software in just a few hours or even minutes.

- It is compatible with numerous Python libraries, such as Pandas, Matplotlib, Seaborn, Plotly, Keras, PyTorch, and SymPy (latex).

- With minimal code, remarkable online applications can be developed.

- Data caching simplifies and speeds up computation pipelines.



Diabetes Prediction using ML

Pregnancies: 0

Glucose level: 0

Bloodpressure: 0

SkinThickness: 0

Insulin: 0

BMI: 0.00

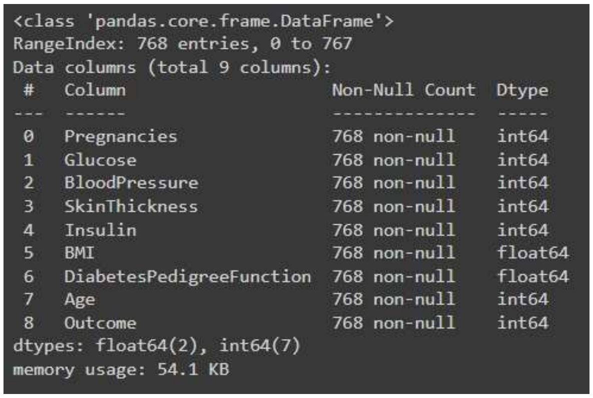
Diabetes Pedigree Function: 0.00

Age: 1

C. DATASET PREPARATION

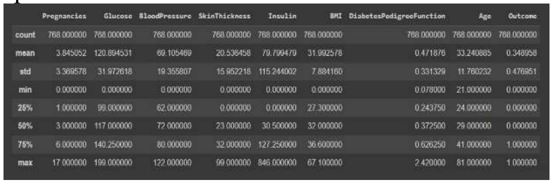
DIABETES

After studying the dataset, we got to know that there are no missing values, there are no categorical values, no need to normalize and we have handled the outliers by filling the mean values in it, the below image consists the details after the changes.



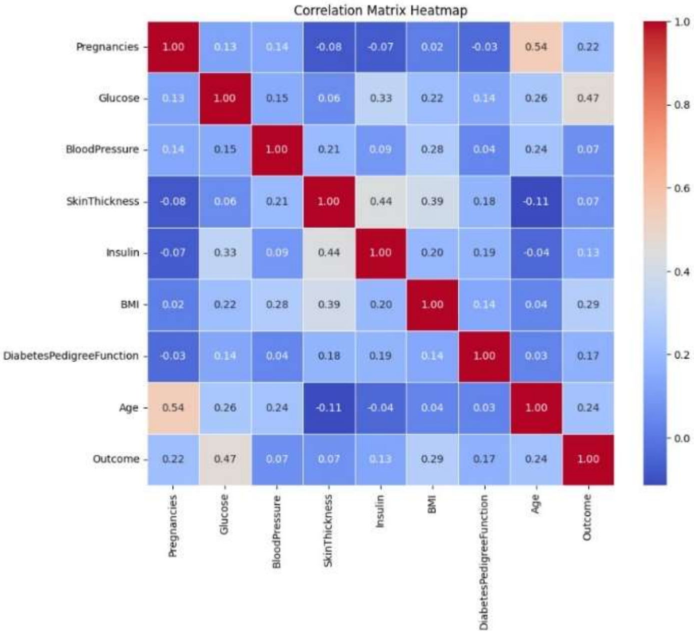
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Feature selection approaches are used to find the most useful information. These methods choose a subset of pertinent characteristics in an effort to decrease dimensionality improve model performance.



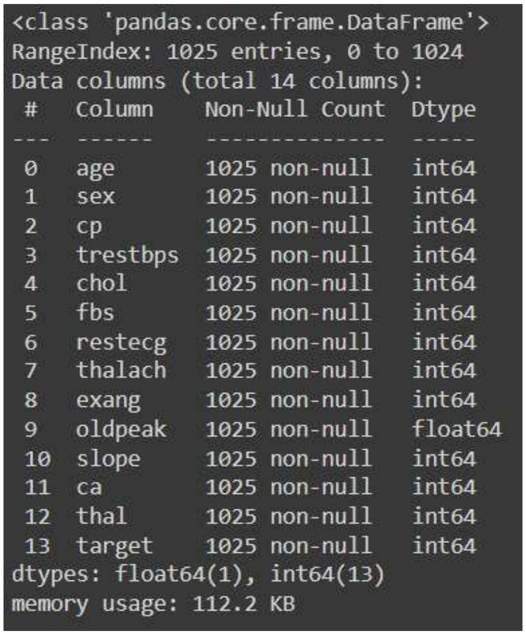
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	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

To rank and choose the most important features, one might use well-liked techniques like correlation analysis or recursive feature elimination (RFE). By observing correlation matrix we can drop some unwanted features the more the value is closer to zero the less the classification depends on it, we can remove the features whose value is in between -0.05 and +0.05.



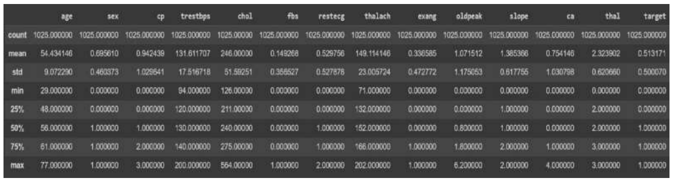
HEART DISEASE

After studying the dataset, we got to know that there are no missing values, there are no categorical values, no need to normalize and we have handled the outliers by filling the mean values in it, the below image consists the details after the changes.



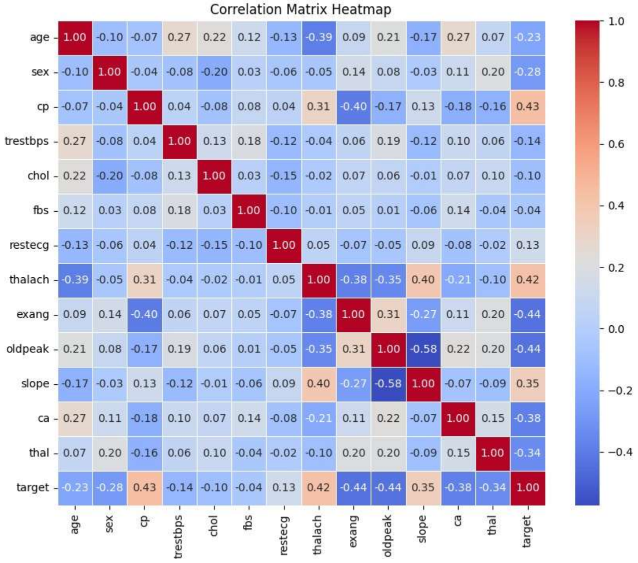
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        1025 non-null   int64
 1   sex        1025 non-null   int64
 2   cp         1025 non-null   int64
 3   trestbps   1025 non-null   int64
 4   chol       1025 non-null   int64
 5   fbs        1025 non-null   int64
 6   restecg    1025 non-null   int64
 7   thalach    1025 non-null   int64
 8   exang      1025 non-null   int64
 9   oldpeak    1025 non-null   float64
 10  slope      1025 non-null   int64
 11  ca         1025 non-null   int64
 12  thal       1025 non-null   int64
 13  target     1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

Feature selection approaches are used to find the most useful information. These methods choose a subset of pertinent characteristics in an effort to decrease dimensionality improve model performance.



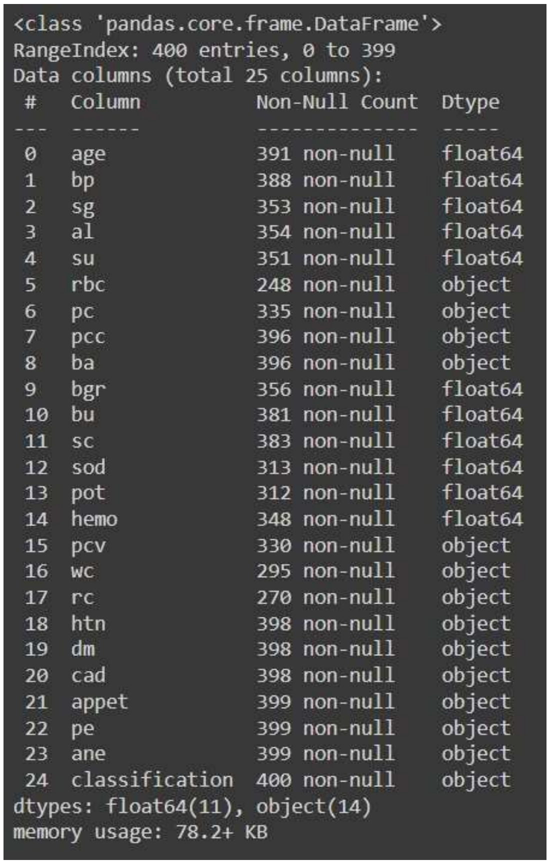
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	1025.000000	1025.000000	1025.000000	1025.000000	1025.00000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000	1025.000000
mean	54.434146	0.695610	0.942439	131.611707	246.00000	0.149268	0.529756	149.114146	0.336585	1.071512	1.385366	0.754146	2.323902	0.513171
std	9.072290	0.460373	1.029641	17.516718	51.59251	0.356527	0.527878	23.005724	0.472772	1.175053	0.617755	1.030798	0.620660	0.500070
min	29.000000	0.000000	0.000000	94.000000	126.00000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	48.000000	0.000000	0.000000	120.000000	211.00000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	56.000000	1.000000	1.000000	130.000000	240.00000	0.000000	1.000000	152.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	275.00000	0.000000	1.000000	166.000000	1.000000	1.800000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.00000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Correlation matrix using heatmap



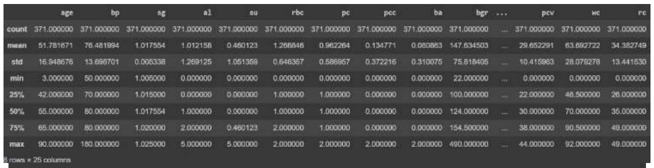
KIDNEY DISEASE

After studying the dataset, we got to know that there are no missing values, but here we have few categorical values like rbc, pc, pcc, ba and so on, so firstly we had tokenized the categorial values into numerical values and then normalized the data and handled the outliers by filling the mean values in it, the below image consists the details before the changes.



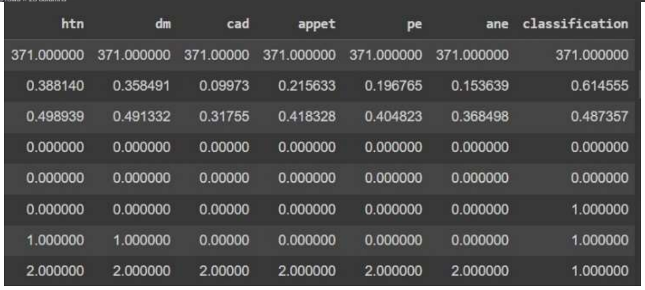
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 25 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             391 non-null    float64
 1   bp              388 non-null    float64
 2   sg              353 non-null    float64
 3   al              354 non-null    float64
 4   su              351 non-null    float64
 5   rbc             248 non-null    object
 6   pc              335 non-null    object
 7   pcc             396 non-null    object
 8   ba              396 non-null    object
 9   bgr             356 non-null    float64
 10  bu              381 non-null    float64
 11  sc              383 non-null    float64
 12  sod             313 non-null    float64
 13  pot             312 non-null    float64
 14  hemo            348 non-null    float64
 15  pcv             330 non-null    object
 16  wc              295 non-null    object
 17  rc              270 non-null    object
 18  htn             398 non-null    object
 19  dm              398 non-null    object
 20  cad             398 non-null    object
 21  appet           399 non-null    object
 22  pe              399 non-null    object
 23  ane             399 non-null    object
 24  classification  400 non-null    object
dtypes: float64(11), object(14)
memory usage: 78.2+ KB
```

The below image displays the changes made in the respective categories and tokenized all the categorial features into numerical



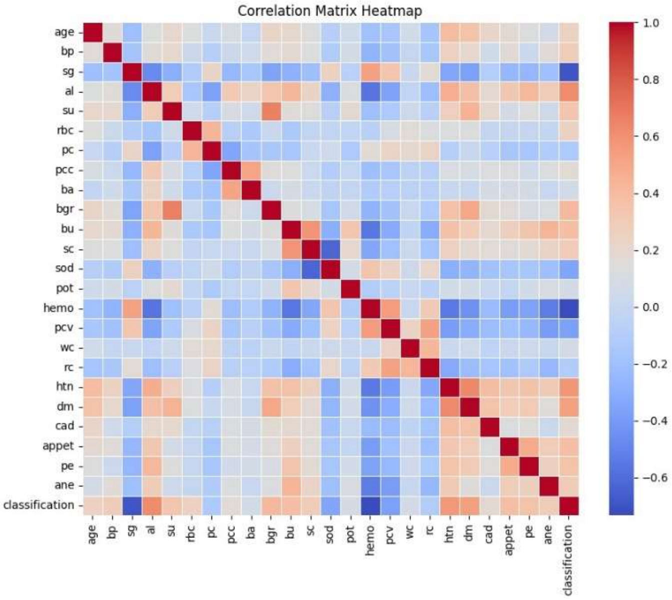
	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	pcv	wc	rc
count	371.000000	371.000000	371.000000	371.000000	371.000000	371.000000	371.000000	371.000000	371.000000	371.000000	...	371.000000	371.000000	371.000000
mean	51.781671	76.481994	1.017554	1.012158	0.460123	1.266846	0.962264	0.134771	0.080863	147.634503	...	29.652291	63.692722	34.382749
std	16.948676	13.698701	0.005338	1.269125	1.051369	0.646367	0.586957	0.372216	0.310075	75.818405	...	10.415963	28.079278	13.441530
min	3.000000	50.000000	1.005000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	22.000000	...	0.000000	0.000000	0.000000
25%	42.000000	70.000000	1.015000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	100.000000	...	22.000000	46.500000	26.000000
50%	55.000000	80.000000	1.017554	1.000000	0.000000	1.000000	1.000000	0.000000	0.000000	124.000000	...	30.000000	70.000000	35.000000
75%	65.000000	80.000000	1.020000	2.000000	0.460123	2.000000	1.000000	0.000000	0.000000	154.500000	...	38.000000	90.500000	49.000000
max	90.000000	180.000000	1.025000	5.000000	5.000000	2.000000	2.000000	2.000000	2.000000	490.000000	...	44.000000	92.000000	49.000000

8 rows × 25 columns



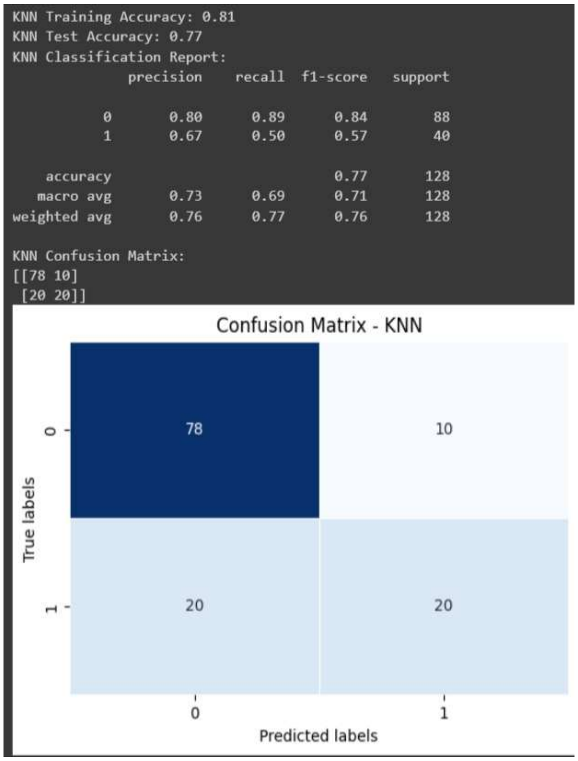
htn	dm	cad	appet	pe	ane	classification
371.000000	371.000000	371.00000	371.000000	371.000000	371.000000	371.000000
0.388140	0.358491	0.09973	0.215633	0.196765	0.153639	0.614555
0.498939	0.491332	0.31755	0.418328	0.404823	0.368498	0.487357
0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	1.000000
1.000000	1.000000	0.00000	0.000000	0.000000	0.000000	1.000000
2.000000	2.000000	2.00000	2.000000	2.000000	2.000000	1.000000

Feature selection approaches are used to find the most useful information. These methods choose a subset of pertinent characteristics in an effort to decrease dimensionality improve model performance. Correlation matrix using heatmap



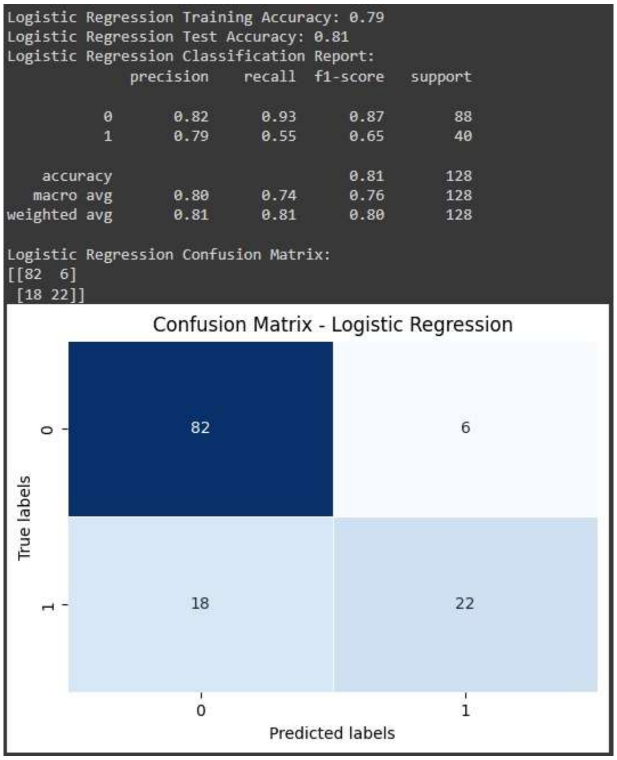
5. RESULTS

DIABETES



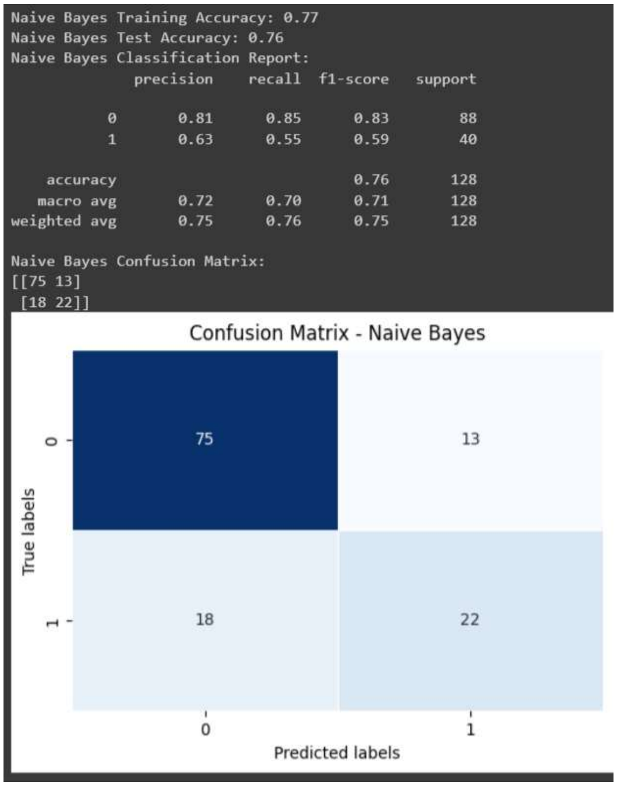
```
KNN Training Accuracy: 0.81
KNN Test Accuracy: 0.77
KNN Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.89      0.84        88
           1       0.67      0.50      0.57        40

    accuracy                           0.77       128
   macro avg       0.73      0.69      0.71       128
weighted avg       0.76      0.77      0.76       128

KNN Confusion Matrix:
[[78 10]
 [20 20]]
```

Accuracy using KNN: 77%



```
Logistic Regression Training Accuracy: 0.79
Logistic Regression Test Accuracy: 0.81
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.93      0.87        88
           1       0.79      0.55      0.65        40

    accuracy                           0.81       128
   macro avg       0.80      0.74      0.76       128
weighted avg       0.81      0.81      0.80       128

Logistic Regression Confusion Matrix:
[[82  6]
 [18 22]]
```

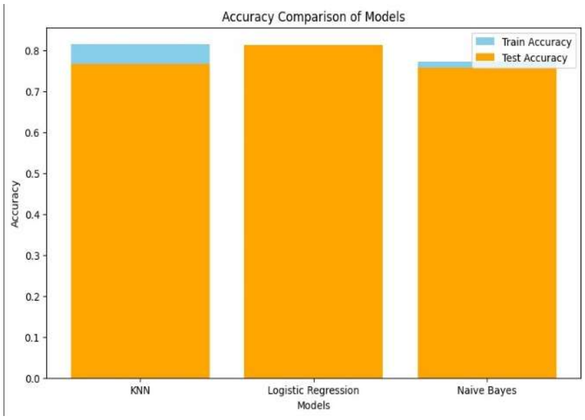
Accuracy using Logistic Regression: 81%



```
Naive Bayes Training Accuracy: 0.77
Naive Bayes Test Accuracy: 0.76
Naive Bayes Classification Report:
              precision    recall  f1-score   support

           0       0.81      0.85      0.83        88
           1       0.63      0.55      0.59        40

    accuracy                           0.76       128
   macro avg       0.72      0.70      0.71       128
weighted avg       0.75      0.76      0.75       128

Naive Bayes Confusion Matrix:
[[75 13]
 [18 22]]
```

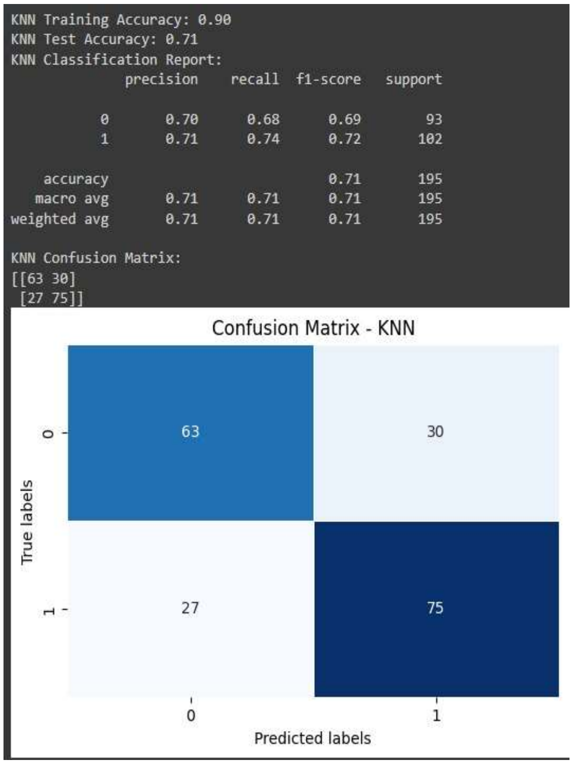
Accuracy using Naïve Bayes: 76%

Algorithm	Accuracy
KNN	77
Logistic Regression	81
Naïve Bayes	76



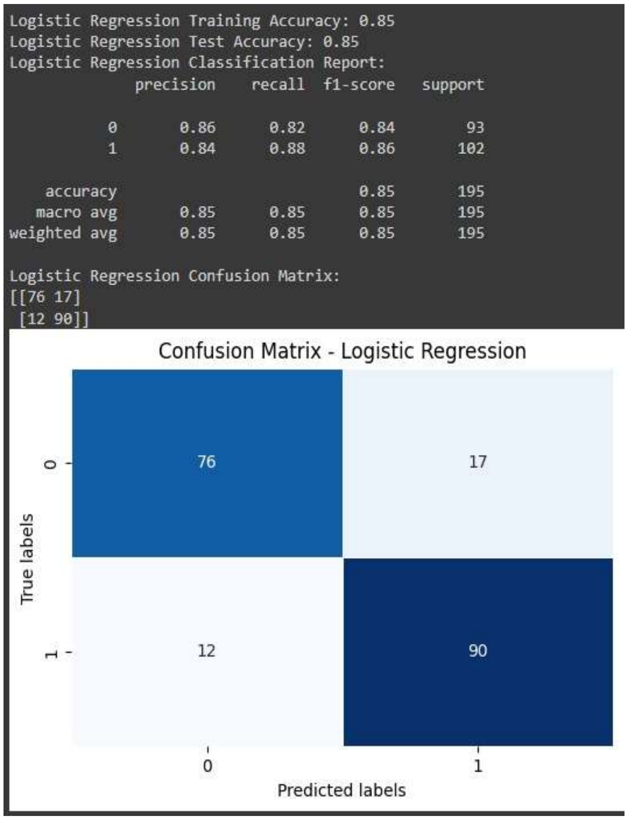
Comparison between the training and testing accuracy of each model.

HEART DISEASE



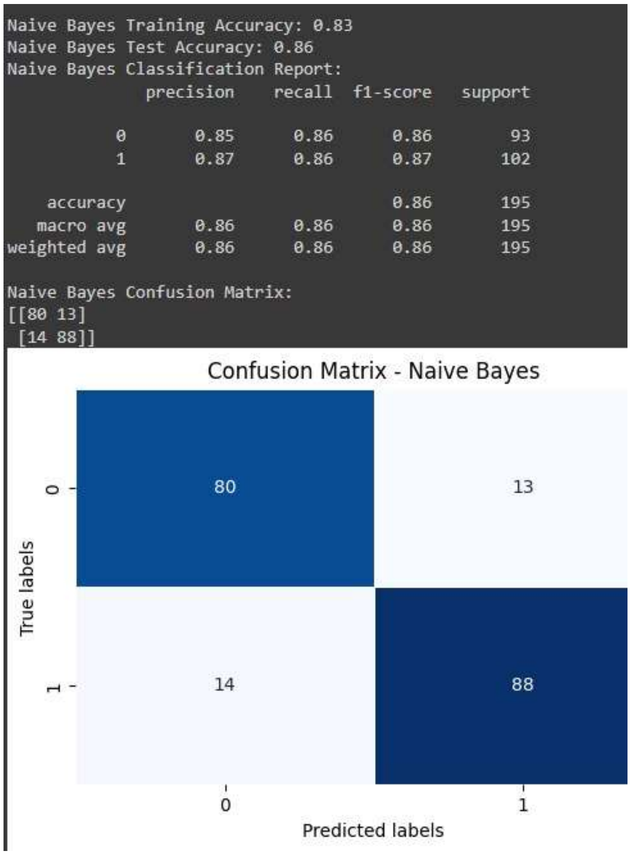
```
KNN Training Accuracy: 0.90
KNN Test Accuracy: 0.71
KNN Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.68      0.69        93
           1       0.71      0.74      0.72       102

    accuracy                           0.71       195
   macro avg       0.71      0.71      0.71       195
weighted avg       0.71      0.71      0.71       195

KNN Confusion Matrix:
[[63 30]
 [27 75]]
```

Accuracy using KNN: 71%



```
Logistic Regression Training Accuracy: 0.85
Logistic Regression Test Accuracy: 0.85
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.82      0.84        93
           1       0.84      0.88      0.86       102

    accuracy                           0.85       195
   macro avg       0.85      0.85      0.85       195
weighted avg       0.85      0.85      0.85       195

Logistic Regression Confusion Matrix:
[[76 17]
 [12 90]]
```

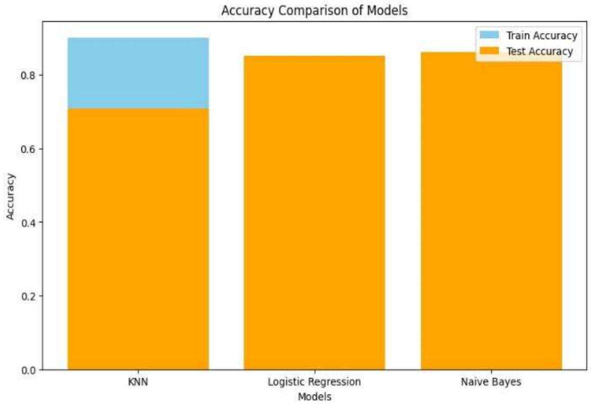
Accuracy using Logistic Regression: 85%



```
Naive Bayes Training Accuracy: 0.83
Naive Bayes Test Accuracy: 0.86
Naive Bayes Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.86      0.86        93
           1       0.87      0.86      0.87       102

    accuracy                           0.86       195
   macro avg       0.86      0.86      0.86       195
weighted avg       0.86      0.86      0.86       195

Naive Bayes Confusion Matrix:
[[80 13]
 [14 88]]
```

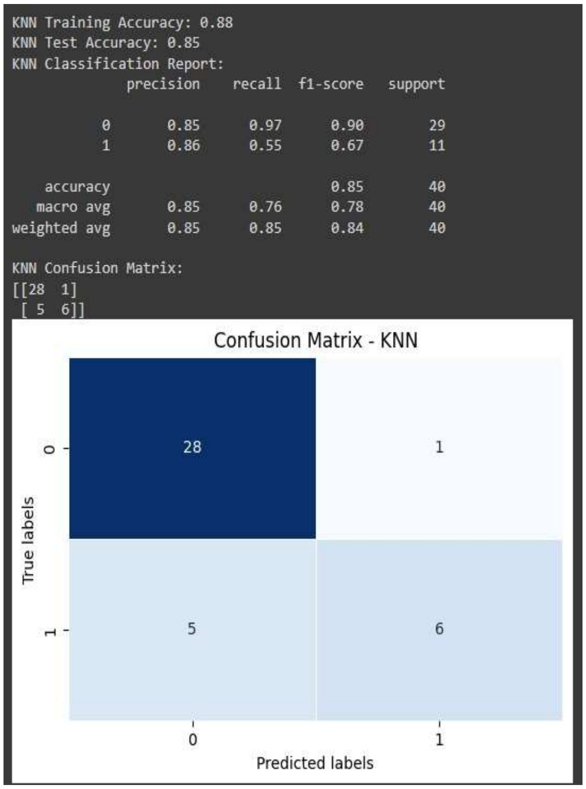
Accuracy using Naïve Bayes: 86%

Algorithm	Accuracy
KNN	71
Logistic Regression	85
Naïve Bayes	86



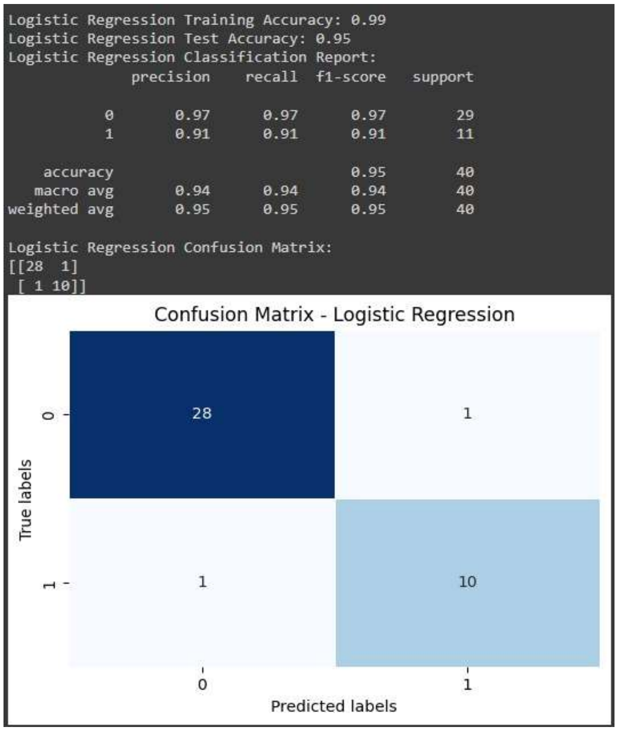
Comparison between the training and testing accuracy of each model.

KIDNEY DISEASE



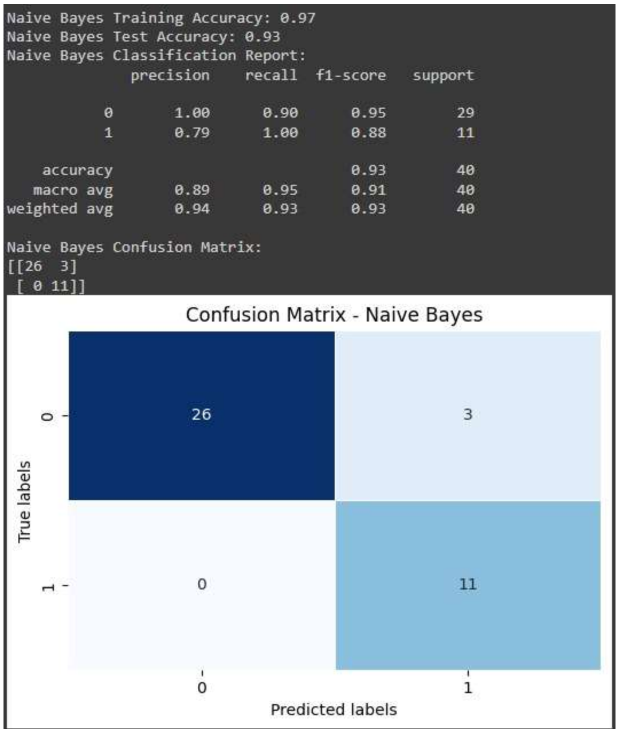
```
KNN Training Accuracy: 0.88
KNN Test Accuracy: 0.85
KNN Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.97      0.90        29
           1       0.86      0.55      0.67        11

    accuracy                           0.85        40
   macro avg       0.85      0.76      0.78        40
weighted avg       0.85      0.85      0.84        40

KNN Confusion Matrix:
[[28  1]
 [ 5  6]]
```

Accuracy using KNN: 85%



```
Logistic Regression Training Accuracy: 0.99
Logistic Regression Test Accuracy: 0.95
Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97        29
           1       0.91      0.91      0.91        11

    accuracy                           0.95        40
   macro avg       0.94      0.94      0.94        40
weighted avg       0.95      0.95      0.95        40

Logistic Regression Confusion Matrix:
[[28  1]
 [ 1 10]]
```

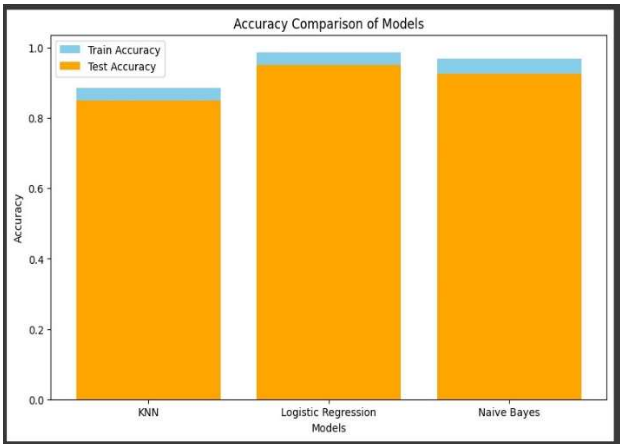
Accuracy using Logistic Regression: 95%



```
Naive Bayes Training Accuracy: 0.97
Naive Bayes Test Accuracy: 0.93
Naive Bayes Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.90      0.95        29
           1       0.79      1.00      0.88        11

    accuracy                           0.93        40
   macro avg       0.89      0.95      0.91        40
weighted avg       0.94      0.93      0.93        40

Naive Bayes Confusion Matrix:
[[26  3]
 [ 0 11]]
```

Accuracy using Naïve Bayes: 93%

Algorithm	Accuracy
KNN	85
Logistic Regression	95
Naïve Bayes	93



Comparison between the training and testing accuracy of each model.

6. RESULT AND ANALYSIS DISCUSSION

In this work, we looked into the use of various machine learning models for predicting multiple diseases. The study made use of a few publicly accessible disease datasets that included clinical and demographic details such age, sex, blood pressure, cholesterol levels, and the existence or absence of heart disease. To guarantee data quality and compatibility, each dataset underwent preparation procedures that included addressing missing values, encoding categorical variables, and normalizing numerical features and handling outliers. The most useful features for disease predictions were then found using feature selection approaches. On the pre-processed datasets, a number of machine learning methods were trained and assessed, including Logistic Regression, K-Nearest Neighbors, and Naive Bayes. The models' propensity for prediction was evaluated using performance metrics like accuracy, precision, recall, and F1-score. The Logistic Regression outperformed the other evaluated models, displaying the greatest accuracy of 81% in diabetes and 95% in chronic kidney disease, Naïve Bayes outperformed the other evaluated models, displaying the greatest accuracy of 86% in heart disease. The model's higher performance is attributable to its capacity to recognize crucial decision limits and capture non-linear relationships within the data. The findings of this study demonstrate how machine learning methods may be used to precisely forecast multiple diseases. The results show the efficiency of employing machine learning models for early identification and diagnosis of diseases offering useful insights for medical practitioners and researchers in the field. To improve prediction accuracy and robustness, future study may involve further tuning the various model and investigating ensemble approaches.

7. CONCLUSION AND FUTURE SCOPE

In this work, we investigated the use of various machine learning models for predicting diabetes, heart disease, and chronic kidney disease. Using publicly accessible datasets on these diseases, we evaluated various algorithms and discovered that different algorithms give the best results for different diseases, depending on the dataset and the disease type. Our results highlight the promise of machine learning methods for precise disease prediction based on clinical and demographic characteristics.

It's crucial to recognize our model's limits, though. The reliance on a particular dataset, which might not accurately reflect the diverse population and differences in disease features, is one drawback. It is important to look into and evaluate the generalizability of the model to various populations and environments using larger and more varied datasets. Additionally, the selected features from the dataset are the main focus of our model. Although these characteristics have been demonstrated to be significant in the prediction of the diseases, our analysis did not account for all important clinical, genetic, or lifestyle factors. To enhance the model's predictive performance and comprehensiveness, future studies should include more pertinent features.

Furthermore, it is critical to use caution when interpreting the model's predictions. In the field of healthcare, it is crucial to be able to articulate the underlying causes of disease prediction. Gaining the confidence and approval of medical experts can be achieved by creating models with greater interpretability, which can aid in understanding the justification for the predictions.

Despite these drawbacks, the results of this study offer important new perspectives on how machine learning might be used to predict various diseases. In light of the above-mentioned constraints, additional study and model improvement will result in greater accuracy, generalizability, and interpretability, ultimately enhancing clinical decision-making and patient care in the context of managing diseases.

Future Work: Although this study's findings are encouraging, there are still a number of opportunities for further investigation and development. First, by combining the strengths of various models, ensemble approaches like Random Forest or XGBoost may be useful for improving prediction performance. Furthermore, using sophisticated feature engineering methods, including feature extraction or domain-specific feature selection algorithms, may improve the models' capacity for prediction. Additionally, examining the interpretability of the models might offer important insights into the underlying elements influencing the prediction of various diseases. In order to improve clinical decision-making and patient care, techniques like feature importance analysis and decision rule extraction can help in understanding the main aspects and patterns driving the predictions. The generalizability and robustness of the created predictive models can be improved by verifying the model performance on a variety of larger datasets. Building more complete and dependable models for the prediction of diseases can be aided by including data from various sources and demographics. Overall, the results of this study provide new opportunities for the application of machine learning in human health, opening the door to enhanced risk assessment, early detection, and individualized treatment plans for people with diabetes, heart disease, and chronic kidney disease.

8. REFERANCES

- https://ieeexplore.ieee.org/document/10060903
- https://ieeexplore.ieee.org/document/10150023
- https://ieeexplore.ieee.org/document/10064682
- https://ieeexplore.ieee.org/document/9040562

Datasets

- https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
- https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
- https://www.kaggle.com/datasets/mansoordaku/ckdisease