REPORT
2022 AUG-SEP

DISCRIPTION –

Diabetes is a type of chronic disease which is more among the people of all age groups. Predicting this diesease at an aerly stage can help a person to take the necessary precautions and change his/her lifestyle accordingly to either prevent the occurance of this disease or control the disease(for the people who already have the disease).

Tasks –

Build a model which can give higher accuracy of predicting the disease.

SEPTEMBER 11

EXPOSYS DATA LABS INTERNSHIP
Authored by: NAIMISH RAJBHAR

Early-stage Diabetes Prediction using various Machine Learning Techniques and Aggregation of Algorithms used.

Report by- Naimish Rajbhar Data Science Department **(Undergraduate) NIET**

Abstract –

Diabetes is an enduring sickness set off by extended sugar levels in human blood and can influence different organs whenever left untreated. It adds to coronary illness, kidney issues, harmed nerves, harmed veins, and visual deficiency. Opportune illness expectation can save valuable lives and empower medical services consultants to deal with the circumstances. Most diabetic patients have close to zero familiarity with the risk factors they face before conclusion. These days, clinics convey fundamental data frameworks, which create tremendous measures of information that can't be changed over into legitimate/helpful data and can't be utilized to help decision making for clinical purposes. There are different mechanized methods accessible for the previous expectation of illness. Gathering learning is an information examination method that joins different strategies into a solitary ideal prescient framework to assess inclination and variety, what's more, to further develop expectations. Diabetes information, which included fifteen factors, were accumulated using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. The prescient models utilized in this study incorporates Logistic Regression, Decision Tree, Neural Network, XGB Classifier, Random Forest, Voting Classifier to look at the aggregate all models on soft Voting.

Keywords: data mining; diabetes dataset; prediction; Logistic Regression; Decision Tree; ANN; XGB; Voting classifier.

Contents

---

1 INTRODUCTION

What is Diabetes?

Diabetes Also called: Diabetes mellitus

- A group of diseases that result in too much sugar in the blood (high blood glucose).

MOST COMMON TYPES

- Type 2 diabetes A chronic condition that affects the way the body processes blood sugar (glucose).
- Type 1 diabetes A chronic condition in which the pancreas produces little or no insulin.
- ✓ Prediabetes A condition in which blood sugar is high, but not high enough to be type 2 diabetes.
- ✓ Gestational diabetes A form of high blood sugar affecting pregnant women.

Because of rising expectations for everyday comforts, diabetes has become more common in individuals' day to day existences. Diabetes, normally alluded to as diabetes mellitus, is an ongoing condition gotten on by an ascent blood glucose level. Various physical and compound tests can be utilized to recognize this condition. Diabetes that is left untreated and undetected can hurt fundamental organs including the eyes, heart, kidneys, feet, and nerves, as well as cause passing. Diabetes is a persistent condition that can wreck worldwide wellbeing. The World Health Organization (WHO) has directed ongoing investigations that uncover an increment in the number and mortality of diabetic patients worldwide. The WHO guesses that by 2030, diabetes will rank as the seventh driving reason for death. As showed by information from the International Diabetes Federation (IDF), there are at present 537 million diabetics around the world, and this figure is supposed to be 643 million by. The main technique for forestalling diabetes intricacies is to recognize and treat the illness early. The early location of diabetes is significant because its confusions increment after some time.

Diabetes recently affects around 346 million people.
Also, the mayor cause for:

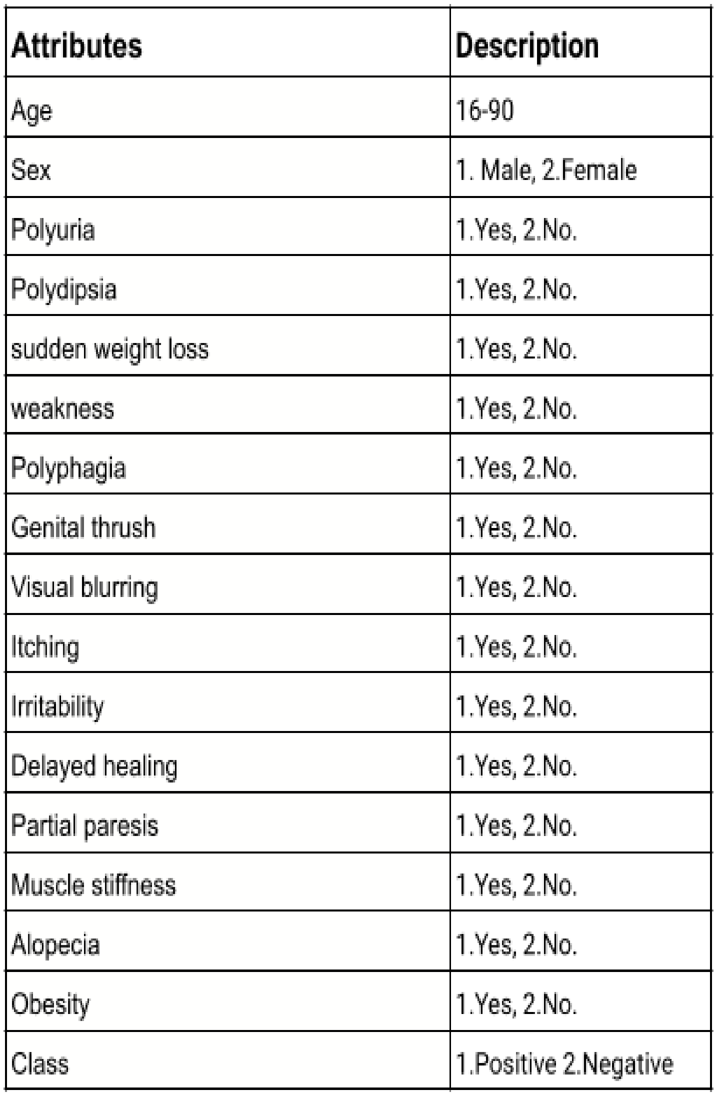
Heart stroke Kidney failure
Lower-limb amputation
Blindness

One-third go undetected in early stage.
Early detection and treatment - substantial health benefits, (avoiding or minimizing the mentioned complications). MECO 2015, Budva, June 2015, Mentenegro.

2. DATA PREPROCESSING AND METHODOLOGY

Data pre-processing is a crucial stage in data mining when dealing with incomplete, noisy, or inconsistent data that transforms the data into a usable and optimal form. To continually formulate data in a coherent and correct form, data preparation covers different activities such as data cleaning, data discretization, data integration, data reduction, data transformation, and so on. For this case study, diabetes data with 17 attributes were collected from the UCI repository which contains different datasets. The dataset utilized here comprises 17 attributes reflecting patient and hospital outcomes. It has been used to assess the accuracy of the prediction by applying ensemble techniques and is made up of clinical treatment data that were gathered by direct surveys from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh, and were validated by the doctors. Some data mining techniques find discrete characteristics easier to deal with. Discrete attributes, often known as nominal attributes, are those that characterize a category. Ordinal characteristics are those qualities that characterize a category and have significance in the order of the categories. Discretization is the process of turning a realvalued attribute into an ordinal attribute or bin. A discretize filter was applied here because the input values are real, and it could be useful to assemble them into bins [34]. In this study, 520 instances are used, with 15 attributes including a class attribute used to predict the positive and negative rate of chances of having diabetes or not.

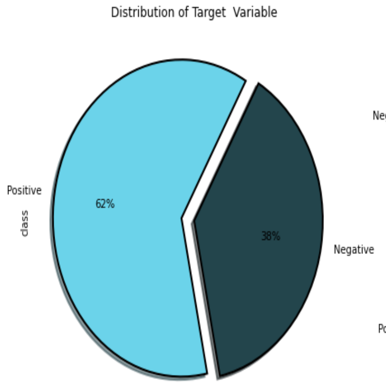
The relevant attributes are evaluated in this research using the Feature selection technique. Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve. For diabetic data, the Feature selection technique is applied to select important attributes desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model. Statistical-based feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable. These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables. As such, it can be challenging for a machine learning practitioner to select an appropriate statistical measure for a dataset when performing filter-based feature selection.

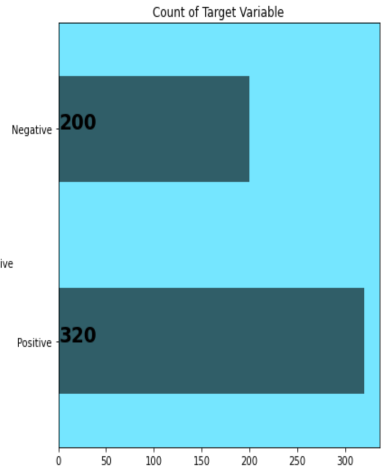


Attributes	Description
Age	16-90
Sex	1. Male, 2.Female
Polyuria	1.Yes, 2.No.
Polydipsia	1.Yes, 2.No.
sudden weight loss	1.Yes, 2.No.
weakness	1.Yes, 2.No.
Polyphagia	1.Yes, 2.No.
Genital thrush	1.Yes, 2.No.
Visual blurring	1.Yes, 2.No.
Itching	1.Yes, 2.No.
Irritability	1.Yes, 2.No.
Delayed healing	1.Yes, 2.No.
Partial paresis	1.Yes, 2.No.
Muscle stiffness	1.Yes, 2.No.
Alopecia	1.Yes, 2.No.
Obesity	1.Yes, 2.No.
Class	1.Positive 2.Negative

a - Logistic Regression (LR)

In linear regression, a threshold is decided for classifying, whereas in binary logistic regression it uses a sigmoid function (equation 2) for defining the thresholds for classification . For Y (output) tending to infinity, it is classified as 1, i.e., "Diabetes", else 0 i.e. "No Diabetes".

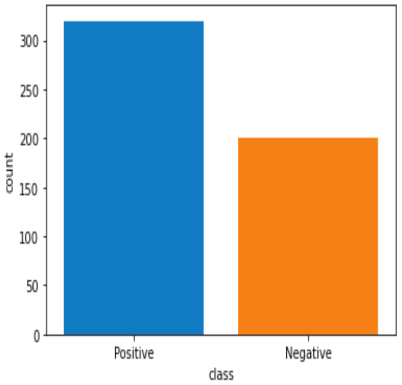




$$Y = 1 / (1 + e^{-z})$$

b – Decision Tree

Decision Tree (DT) Decision trees classify data by creating a top-down tree by dividing the dataset into smaller sub datasets. ID3 along with entropy and information Gain is used recursively for building the decision tree, the root node of the tree signifies the classification.



c - ANN (Artificial Neural Network)

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

d – XGB Classifier

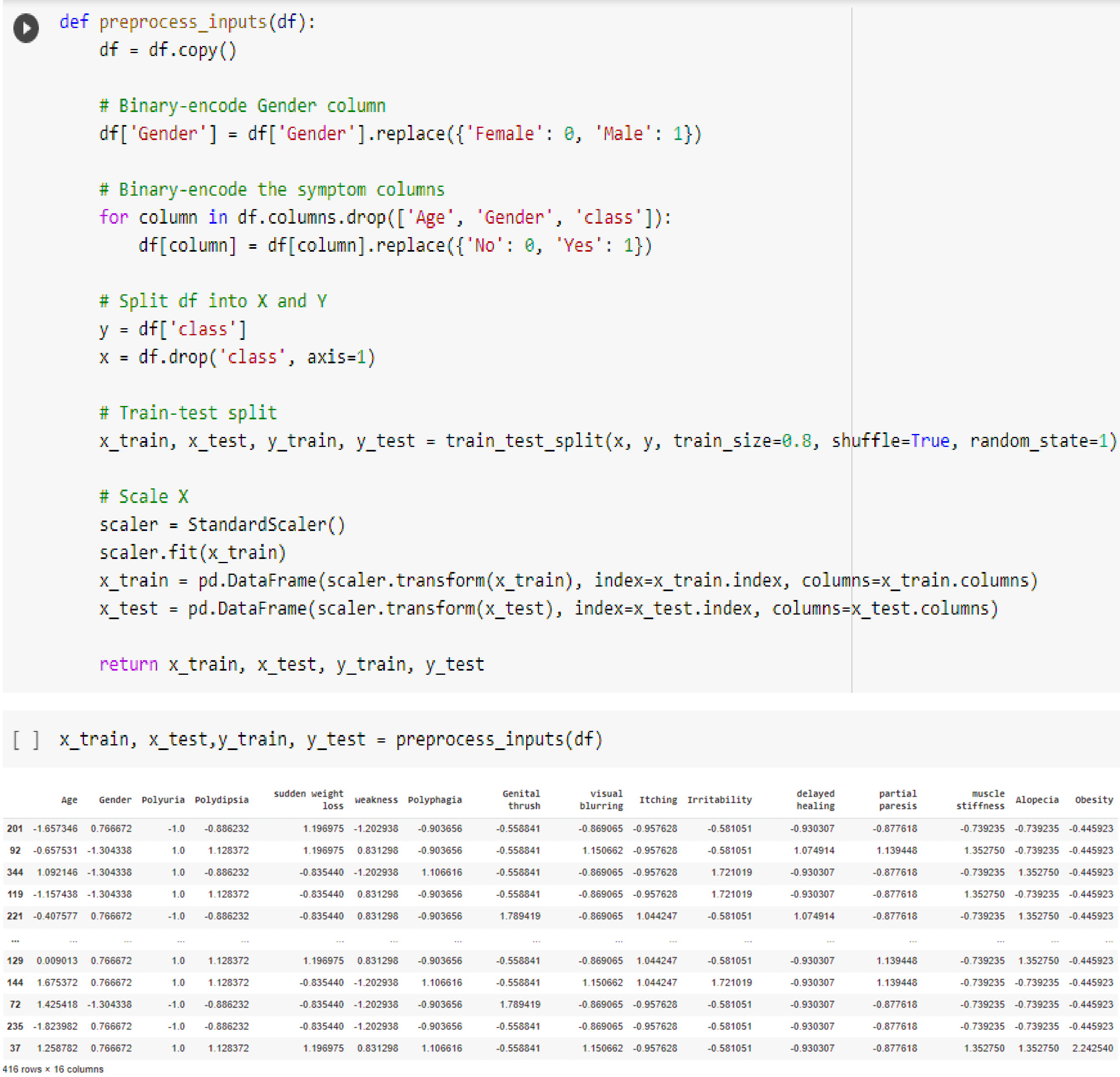
XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

e - Random Classifier

Random forest classifies a dataset by creating several decision trees. It helps correct the over-fitting problem of decision trees. It selects the class by calculating the mode of the trees. It is a very efficient classifier.

3 Splitting data into training and testing Set.

Split the data set into two pieces — a training set and a testing set. This consists of random sampling without replacement about 75 percent of the rows (you can vary this) and putting them into your training set. The remaining 25 percent is put into your test set. Note that the colors in "Features" and "Target" indicate where their data will go ("X_train," "X_test," "y_train," "y_test") for a particular train test split.



```
def preprocess_inputs(df):
    df = df.copy()

    # Binary-encode Gender column
    df['Gender'] = df['Gender'].replace({'Female': 0, 'Male': 1})

    # Binary-encode the symptom columns
    for column in df.columns.drop(['Age', 'Gender', 'class']):
        df[column] = df[column].replace({'No': 0, 'Yes': 1})

    # Split df into X and Y
    y = df['class']
    x = df.drop('class', axis=1)

    # Train-test split
    x_train, x_test, y_train, y_test = train_test_split(x, y, train_size=0.8, shuffle=True, random_state=1)

    # Scale X
    scaler = StandardScaler()
    scaler.fit(x_train)
    x_train = pd.DataFrame(scaler.transform(x_train), index=x_train.index, columns=x_train.columns)
    x_test = pd.DataFrame(scaler.transform(x_test), index=x_test.index, columns=x_test.columns)

    return x_train, x_test, y_train, y_test
```

```
[ ]  x_train, x_test,y_train, y_test = preprocess_inputs(df)
```

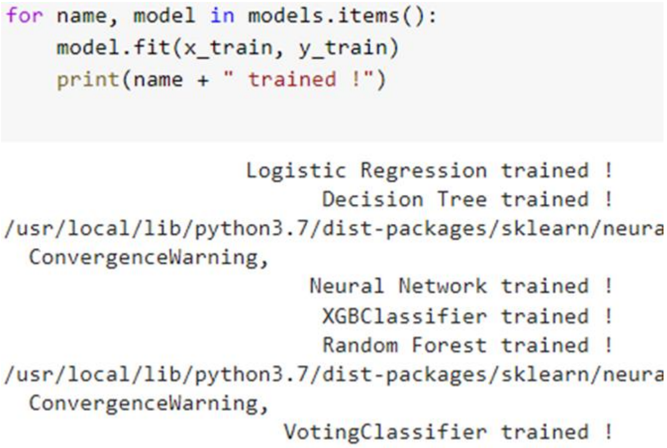
	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity
201	-1.657346	0.766672	-1.0	-0.886232	1.196975	-1.202938	-0.903656	-0.558841	-0.869065	-0.957628	-0.581051	-0.930307	-0.877618	-0.739235	-0.739235	-0.445923
92	-0.657531	-1.304338	1.0	1.128372	1.196975	0.831298	-0.903656	-0.558841	1.150662	-0.957628	-0.581051	1.074914	1.139448	1.352750	-0.739235	-0.445923
344	1.092146	-1.304338	1.0	-0.886232	-0.835440	-1.202938	1.106616	-0.558841	-0.869065	-0.957628	1.721019	-0.930307	-0.877618	-0.739235	1.352750	-0.445923
119	-1.157438	-1.304338	1.0	1.128372	-0.835440	0.831298	-0.903656	-0.558841	-0.869065	-0.957628	1.721019	-0.930307	-0.877618	1.352750	-0.739235	-0.445923
221	-0.407577	0.766672	-1.0	-0.886232	-0.835440	0.831298	-0.903656	1.789419	-0.869065	1.044247	-0.581051	1.074914	-0.877618	-0.739235	1.352750	-0.445923
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
129	0.009013	0.766672	1.0	1.128372	1.196975	0.831298	-0.903656	-0.558841	-0.869065	1.044247	-0.581051	-0.930307	1.139448	-0.739235	1.352750	-0.445923
144	1.675372	0.766672	1.0	1.128372	-0.835440	-1.202938	1.106616	-0.558841	1.150662	1.044247	1.721019	-0.930307	1.139448	-0.739235	-0.739235	-0.445923
72	1.425418	-1.304338	-1.0	-0.886232	-0.835440	-1.202938	-0.903656	1.789419	-0.869065	-0.957628	-0.581051	-0.930307	-0.877618	-0.739235	-0.739235	-0.445923
235	-1.823982	0.766672	-1.0	-0.886232	-0.835440	-1.202938	-0.903656	-0.558841	-0.869065	-0.957628	-0.581051	-0.930307	-0.877618	-0.739235	-0.739235	-0.445923
37	1.258782	0.766672	1.0	1.128372	1.196975	0.831298	1.106616	-0.558841	1.150662	-0.957628	-0.581051	-0.930307	-0.877618	1.352750	1.352750	2.242540

416 rows × 16 columns

X_train

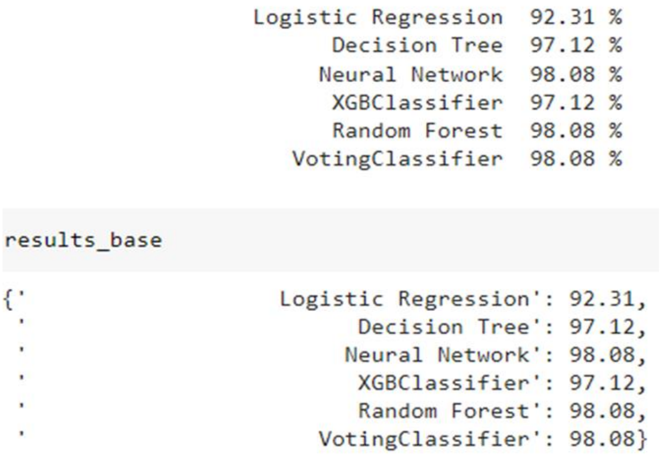
4 Training of Datasets

Training data (or a training dataset) is the initial data used to train machine learning models. Training datasets are fed to machine learning algorithms to teach them how to make predictions or perform a desired task.



```
for name, model in models.items():
    model.fit(x_train, y_train)
    print(name + " trained !")
```

```
                     Logistic Regression trained !
                           Decision Tree trained !
/usr/local/lib/python3.7/dist-packages/sklearn/neura
  ConvergenceWarning,
                          Neural Network trained !
                           XGBClassifier trained !
                           Random Forest trained !
/usr/local/lib/python3.7/dist-packages/sklearn/neura
  ConvergenceWarning,
                         VotingClassifier trained !
```

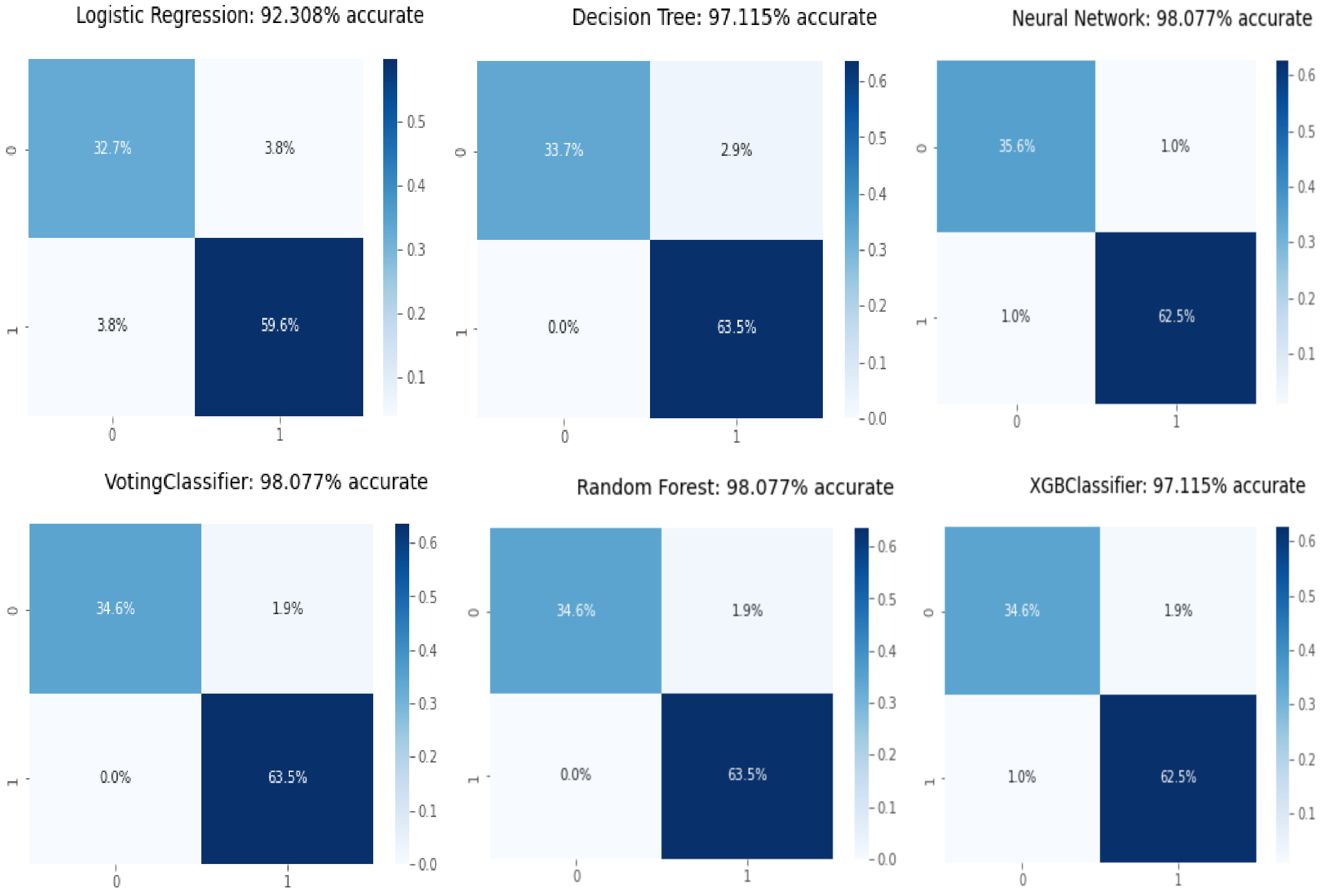


```
Logistic Regression  92.31 %
      Decision Tree  97.12 %
     Neural Network  98.08 %
      XGBClassifier  97.12 %
      Random Forest  98.08 %
   VotingClassifier  98.08 %
```

```
results_base
```

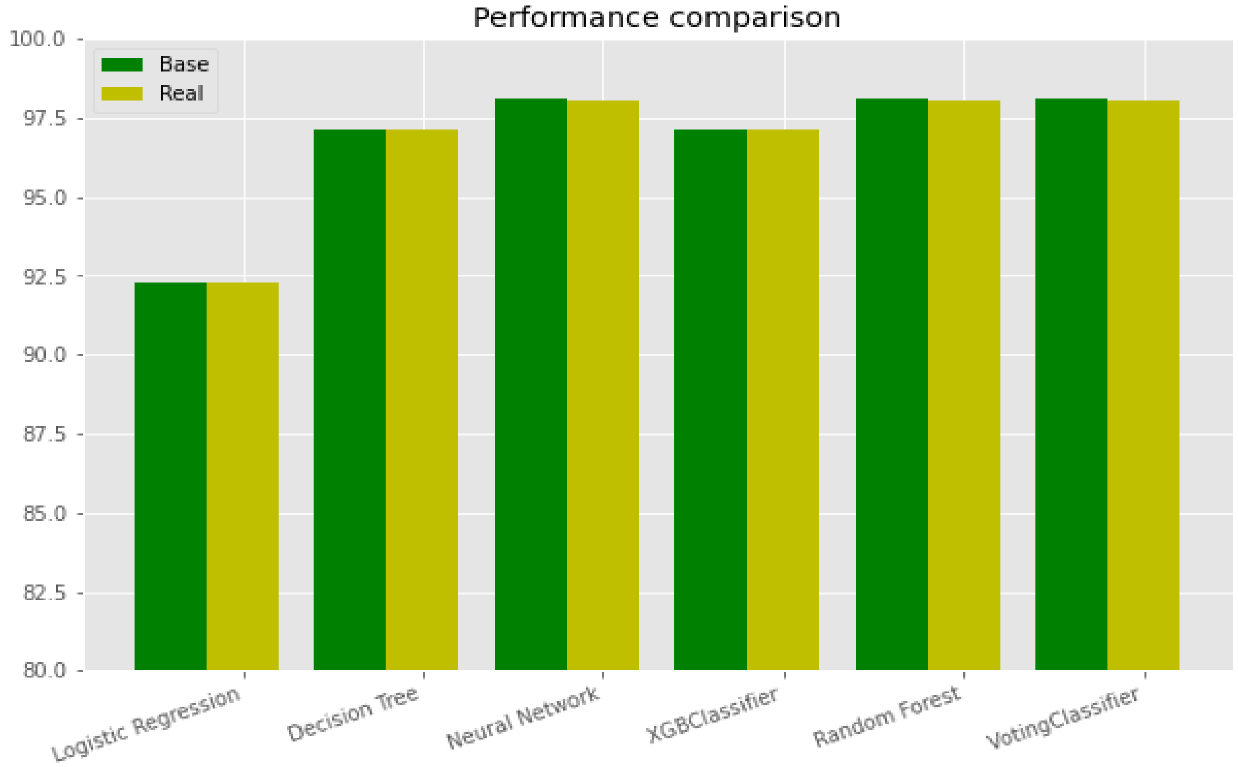
```
{'               Logistic Regression': 92.31,
 '                     Decision Tree': 97.12,
 '                    Neural Network': 98.08,
 '                     XGBClassifier': 97.12,
 '                     Random Forest': 98.08,
 '                  VotingClassifier': 98.08}
```

5 Visualization of Prediction Accuracy of different model



6 Performance Comparison

Performance of an ML model is just "how good" it does at a particular task, but the definition of "good" can take many forms. A "good" model could be one that predicts well, one that trains quickly, one that finds a robust solution, or any combination of the above.



7 Results

The best models for predicting diabetes in this dataset are:-

- Decision Tree
- Neural Network
- Random Forest.

The best accuracy we can get in Decision Tree, Neural Network and Random Forest, all with 98.07% accuracy with feature selection using voting classifier.

8 Conclusion

Proposing a system that can distinguish the patients effectively. It can assist Specialist to determine the outcomes easily by entering symptoms as my best selected models can predict accurately around 98 percent which is a good score. Reduces other clinical complications.