SIGNIFICANCE OF MULTILAYER PERCEPTRON MODEL FOR EARLY DETECTION OF DIABETES OVER ML METHODS

Dr. V.Vasudha Rani[1], Dr. G.Vasavi[2], Dr. K.R.N Kiran Kumar[3]
[1]Sr.Asst Professor, GMR Institute of Technology,
Rajam, Andhra Pradesh, India.
vasudharani.v@gmrit.edu.in
[2]Assistant Professor, Dept. of CSE,
Malla Reddy Engineering College, Hyderabad, INDIA.
vasavig.cse@mrec.ac.in
[3]Professor, Dept. of CSE,
drkrnkk@gmail.com

Abstract

Diabetes is one of the chronicdiseases in the world. Millions of people are suffering with several other health issues caused by diabetes, every year. Diabetes has got three stages such as type2, type1 and insulin. Curing of diabetes disease at later stages is practically difficult. Here in this paper, we proposed a DNN model and its performance comparison with some of the machine learning models to predict the disease at an earlystage based on the current health condition of the patient. An artificial neural network (ANN) is a predictive model designed to work the same way a human brain does and works better with larger datasets. Having the concept of hidden layers, neural networks work better at predictive analytics and can make predictions with more accuracy. Novelty of this work lies in integration of feature selection method used to optimize the Multilayer Perceptron (MLP) to reduce the number of required input attributes. The results achieved using this method and several conventional machines learning approaches such as Logistic Regression, Random Forest Classifier (RFC) are compared. The proposed DNN method is proved to show better accuracy than Machine learning models for early stage detection of diabetes. This paper work is applicable to clinical support as a tool for making pre-decisions by the doctors and physicians.

Keywords: *Deep Neural networks, feature selection, diabetes, multilayer perceptron, random forest classifier, logistic regression*

1. INTRODUCTION

Early prediction of diabetes is implemented by using Deep Learning Techniques by collecting some real health records data of the people. Based on the records, Machine is trained by Deep Learning Techniques which gives effective and accurate results. The main goal of this project is to predict Diabetes early, so that the people can take precautionary actions for not getting diabetes.

Diabetes disease is identified in a patient when his/her blood glucose or blood sugar level is too high. Insulin helps in generating energy required for the body through blood glucose from the food the person has taken. Insulin, a hormone made by the pancreas. Sometimes if body couldn't make enough insulin or doesn't use insulin well more Glucose stays in the blood and doesn't reach your cells as energy. As time passes by, too much glucose in the body causes other health problems. Although diabetes can't be cured completely at least we can take precaution to prevent it before it attacks us. There are **3** stages in Diabetes i. **type 1** ii. **type 2** iii. **Insulin.** Gestational diabetes is a type of Diabetes the pregnant lady gets. People usually makes a false assumption that type1 is the initial stage of diabetes. But Type2 is the initial stage of diabetes and then type1 and later stage requires external insulin supply to the body is called insulin stage.

Deep Learning: Deep learning is the sub domain of machine learning and artificial intelligence that achieves great power and flexibility. The world is full of engineering concepts that can be trained and performed through machines is called machine learning. Make the machine to do everything a human can such as think take a decision.

ANN:An artificial neural network is a type of a computing system designed to simulate the way same the human brain analyses and processes information. Deep learning can solve even the complex problems that were not possible by humans. ANN models have self-learning capabilities to produce better results compared to existing. NN model works based on the below given formula.

$$z = f(b + x \cdot w) = f\left(b + \sum_{i=1}^{n} x_i w_i\right)$$

$$x \in d_{1\times n},\ w \in d_{n\times 1},\ b \in d_{1\times 1},\ z \in d_{1\times 1}$$

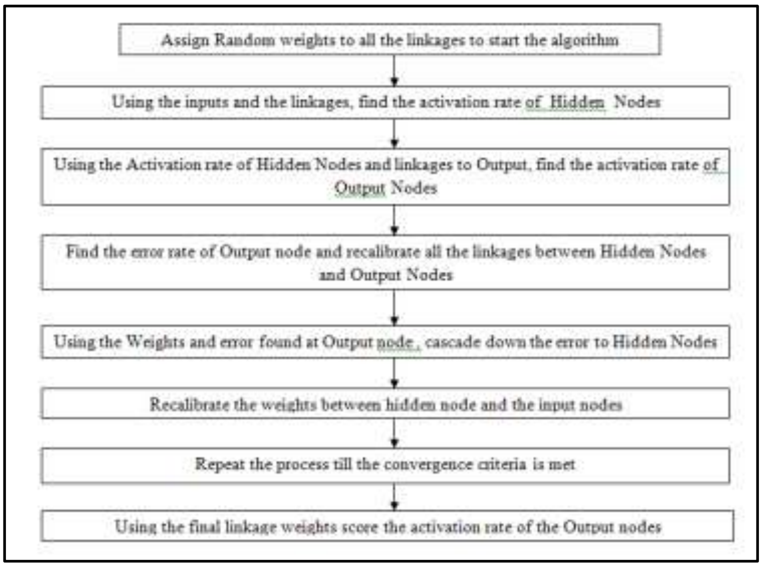

Assign Random weights to all the linkages to start the algorithm
Using the inputs and the linkages, find the activation rate of Hidden Nodes
Using the Activation rate of Hidden Nodes and linkages to Output, find the activation rate of Output Nodes
Find the error rate of Output node and recalibrate all the linkages between Hidden Nodes and Output Nodes
Using the Weights and error found at Output node, cascade down the error to Hidden Nodes
Recalibrate the weights between hidden node and the input nodes
Repeat the process till the convergence criteria is met
Using the final linkage weights score the activation rate of the Output nodes

Figure 1.1 Process iterations in an ANN model

Early detection of Diabetes is implemented by Deep Learning Techniques. Before using deep learning, feature selection technique is used to get top features from the data and using that deep learning model for predicting. The main reason for using deep learning model is predicting the diabetes effectively and getting more accuracy.Perceptron is defined as a single layered neural network that classifies the linear data. Perceptronnetwork constitutes four default components, such asInputs, Weights and Bias, Summation Functions and Activation or transformation function.

Description of the Dataset: The attributes of the dataset are:

Age: a period of human life, measured by years from birth, usually marked by a certain stage or degree of mental or physical development and involving legal responsibility and capacity

Polyuria: Polyuria is defined as the frequent passage of large volumes of urine – more than 3 liters a day compared to the normal daily urine output in adults of about 1 to 2 liters.The most common cause of polyuria in both adults and children is uncontrolled diabetes mellitus, which causes osmotic diuresis, when glucose levels are so high that glucose is excreted in the urine. Water follows the glucose concentration passively, leading to abnormally high urine output.In the absence of diabetes mellitus, the most common causes are decreased secretion of aldosterone due to adrenal cortical tumor, primary polydipsia (excessive fluid drinking).

Polydipsia: Polydipsia is the term given to excessive thirst and is one of the initial symptoms of diabetes. It is also usually accompanied by temporary or prolonged dryness of the mouth. However, if you feel thirsty all the time or your thirst is stronger than usual and continues even after you drink, it can be a sign that not all is well inside your body. Excessive thirst can be caused by high blood sugar (hyperglycaemia), and is also one of the ‘Big 3’ signs of diabetes mellitus i.e., Polyuria, Polydipsia, Polyphagia Generally, increased thirst (polydipsia) and an increased need to urinate (polyuria) will often come as a pair.

Sudden Weight Loss:Weight loss, in the context of medicine, health, or physical fitness, refers to a reduction of the total body mass, by a mean loss of fluid, body fat (adipose tissue), or lean mass (namely bone mineral deposits, muscle, tendon, and other connective tissue). Weight loss can either occur unintentionally because of malnourishment or an underlying disease, or from a conscious effort to improve an actual or perceived overweight or obese state.

Weakness: Weakness is a symptom of a number of different conditions. The causes are many and can be divided into conditions that have true or perceived muscle weakness. True muscle weakness is a primary symptom of a variety of skeletal muscle diseases.

Polyphagia: Polyphagia, also known as hyperphagia, is the medical term for excessive or extreme hunger. It's different than having an increased appetite after exercise or other physical activity. While your hunger level will return to normal after eating in those cases, polyphagia won't go away if you eat more food

Genital Thrush: Candidiasis is a fungal infection due to any type of candida(a type of yeast). When it affects the mouth, in some countries it is commonly called thrush. Signs and symptoms include white patches on the tongue or other areas of the mouth and throat. Other symptoms may include soreness and problems swallowing. When it affects the vagina, it may be referred to as a yeast infection or thrush. Signs and symptoms include genital itching, burning, and sometimes a white "cottage cheese-like" discharge from the vagina.

Visual Blurring: Blurred vision can affect your entire line of sight or just parts of your vision. This could include your peripheral vision, or how you see to the right or left of your field of vision. You can also experience blurred vision in only one eye. Causes for Visual Blurring are i) refractive errors, such as near-sightedness, far-sightedness, or astigmatism. ii) abrasions to the cornea. iii) age-related macular degeneration. iv) cataracts.

Itching: Itching is a sensation that causes the desire or reflex to scfratch. Itching has resisted many attempts to be classified as any one type of sensory experience. Itching has many similarities to pain, and while both are unpleasant sensory experiences, their behavioral response patterns are different.

Irritability: Irritability is the excitatory ability that living organisms have to respond to changes in their environment. The term is used for both the physiological reaction to stimuli and for the pathological, abnormal or excessive sensitivity to stimuli. When reflecting human emotion and behavior, it is commonly defined as the tendency to react to stimuli with negative affective states and temper outbursts, which can be aggressive. Distressing or impairing irritability is important from a mental health.

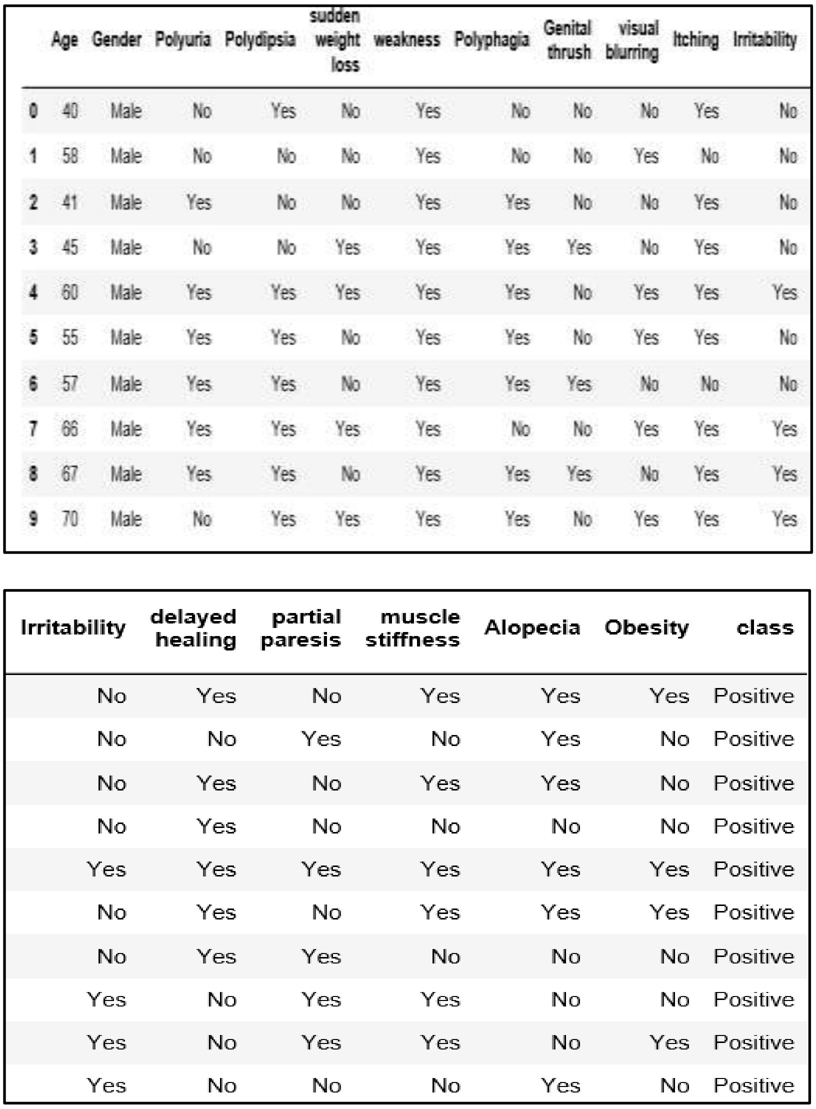
Delayed Healing: Healing is the process of the restoration of health from an unbalanced, diseased, damaged or unvitalized organism. The profession of nursing has been traditionally concerned with matters of healing, whereas historically the profession of medicine has been concerned with curing. With physical damage or disease suffered by an organism, healing involves the repair of living tissue, organs and the biological system as a whole and resumption of functioning.

Partial Paresis: Paresis involves the weakening of a muscle or group of muscles. It may also be referred to as partial or mild paralysis. Unlike paralysis, people with paresis can still move their muscles. These movements are just weaker than normal.

Muscle Stiffness: Muscle stiffness is when your muscles feel tight and you find it more difficult to move than you usually do, especially after rest. You may also have muscle pains, cramping, and discomfort. This is different from muscle rigidity and spasticity.

Alopecia: Sudden hair loss that starts with one or more circular bald patches that may overlap. Alopecia areata occurs when the immune system attacks hair follicles and may be brought on by severe stress. The main symptom is hair loss.

Obesity: Obesity is a medical condition in which excess body fat has accumulated to an extent that it may have a negative effect on health. People are generally considered obese when their body mass index (BMI), a measurement obtained by dividing a person's weight by the square of the person's height—despite known allometric inaccuracies. Obesity is correlated with various diseases and conditions, particularly cardiovascular diseases, type 2 diabetes, obstructive sleep apnea, certain types of cancer, and osteoarthritis.



	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No
1	58	Male	No	No	No	Yes	No	No	Yes	No	No
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
5	55	Male	Yes	Yes	No	Yes	Yes	No	Yes	Yes	No
6	57	Male	Yes	Yes	No	Yes	Yes	Yes	No	No	No
7	66	Male	Yes	Yes	Yes	Yes	No	No	Yes	Yes	Yes
8	67	Male	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes
9	70	Male	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes

Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
No	Yes	No	Yes	Yes	Yes	Positive
No	No	Yes	No	Yes	No	Positive
No	Yes	No	Yes	Yes	No	Positive
No	Yes	No	No	No	No	Positive
Yes	Yes	Yes	Yes	Yes	Yes	Positive
No	Yes	No	Yes	Yes	Yes	Positive
No	Yes	Yes	No	No	No	Positive
Yes	No	Yes	Yes	No	No	Positive
Yes	No	Yes	Yes	No	Yes	Positive
Yes	No	No	No	Yes	No	Positive

Figure 1.2 Patients Dataset with Sample records

II. RELATED WORK

Son Vu Truong Dao [1] proposed a novel wrapper based feature selection clustering based Swam optimization method to optimize the Multilayer Perceptron to reduce the number of required input attributes. Comparison of results among conventional machine learning algorithms with the deep neural network method.The computational results have proved that reduction in no of features will lead to higher prediction accuracy.

Muhammad Syafrudin, Jongtae Rhee [2] states that for improving the healthcare quality of people, early diabetes disease prediction plays a major role and helps individual avoid dangerous health situations before facing them without any delay.Author's proposal is a disease prediction model (DPM) that predicts an early prediction of type2 diabetes and hypertension based on individual patient's health condition data. The author could give effective results for disease prediction through Ensemble Learning Approach.

Alessandro Aliberti [3] suggested that continuous glucose monitoring systems provides an option to record blood glycaemic values at higher rate of sampling. The data collected from the glucose monitoring system can be used to train a machine learning model to analyze future patterns of glucose levels in patients for effective prevention from dangerous hyperglycaemic or hypoglycaemic states and better optimization of the diabetic treatment.

Nahla H. Barakat, Andrew P. Bradley[4] recently presented in his paper,the comparison of performance of commonly known machine-learning (ML) models such as regression models and deep learning models Long-Short-Term-Memory network (LSTM) &Temporal Convolution Network (TCN) versus a classic Autoregression with Exogenous inputs (ARX) model in the prediction of blood glucose (BG) levels using time-series data of patients with Type 1 diabetes (T1D).

III. PROPOSED WORK

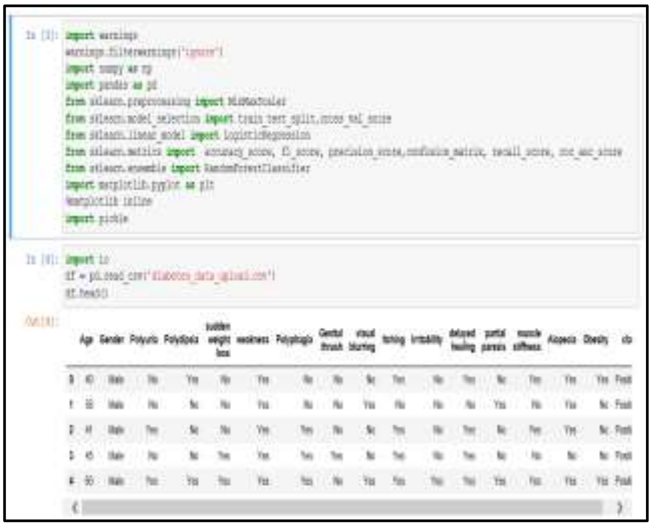
The proposed implementation of this paper includes the execution of three methodsLogistic Regression, Random Forest, Multilayer Perceptron. The process includes 1. Importing essential libraries 2.Data Preprocessing 3.Data Visualization 4.Train Test Split 5. Apply Desired Algorithm 6.Feature Selection 7.Cross Validation 8.Model Evaluation 9.Building Web Application 10.Deploying in Cloud

In our proposed system, Diabetes is predicted by Deep Learning Techniques. Before using deep learning, we have used feature selection technique to get top features from the data and using that deep learning model for predicting. The main reason for using deep learning model is predicting the diabetes effectively and getting more accuracy.

Data Pre-processing:

The process of cleaning raw data for it to be used for machine learning activities is known as data pre-processing.

Loading of Dataset:



```
In [3]: import warnings
        warnings.filterwarnings('ignore')
        import numpy as np
        import pandas as pd
        from sklearn.preprocessing import MinMaxScaler
        from sklearn.model_selection import train_test_split,cross_val_score
        from sklearn.linear_model import LogisticRegression
        from sklearn.metrics import  accuracy_score, f1_score, precision_score,confusion_matrix, recall_score, roc_auc_score
        from sklearn.ensemble import RandomForestClassifier
        import matplotlib.pyplot as plt
        %matplotlib inline
        import pickle

In [4]: import io
        df = pd.read_csv('diabetes_data_upload.csv')
        df.head()

Out[4]:
```

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	cla
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Posi
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Posi
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Posi
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Posi
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Posi

Figure 3.1 Loading of DataSet

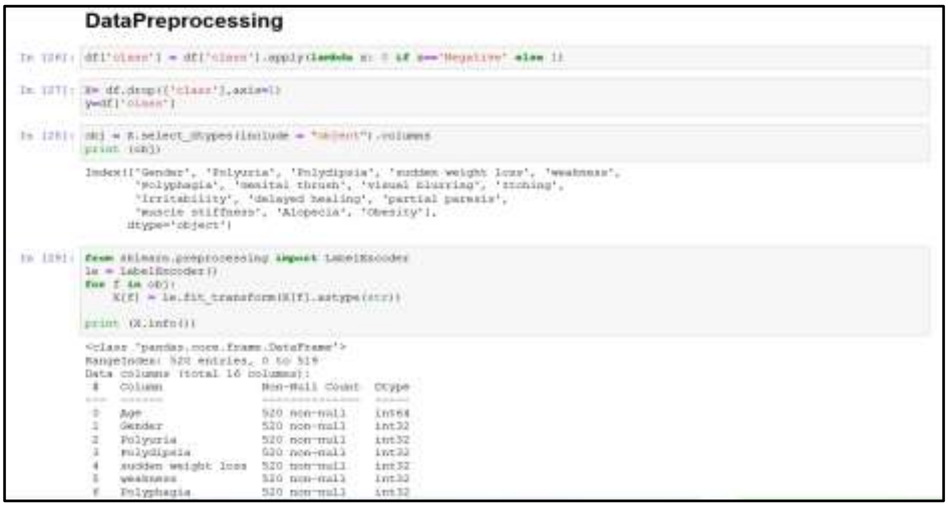
Missing Values Identification

In data preprocessing, it is essential to identify and correctly handle the missing values, failing which leads to inaccurate results and fault conclusions and interpretations from the data. There are two this problem can be handled by:

- Deleting a particular row
- Calculating the mean

Encoding of Categorical Data

Categorical data refers to such information that has specific categories of values within the dataset. The column gender is a categorical data. Machine Learning and deep learning models are primarily based on mathematical equations. Thus, it is required to convert categorical data into numerical. As seen in our dataset example, the column gender has to be converted it into numerical values. To do so LabelEncoder() class from the sci-kit learn library.fromsklearn.preprocessing has to be imported.



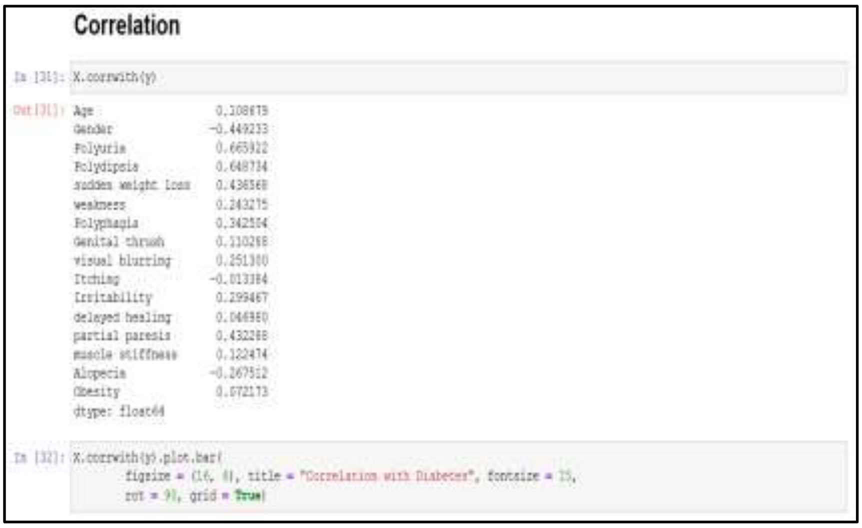
DataPreprocessing

```
In [26]: df['class'] = df['class'].apply(lambda x: 0 if x=='Negative' else 1)

In [27]: X= df.drop(['class'],axis=1)
         y=df['class']

In [28]: obj = X.select_dtypes(include = "object").columns
         print (obj)

         Index(['Gender', 'Polyuria', 'Polydipsia', 'sudden weight loss', 'weakness',
                'Polyphagia', 'Genital thrush', 'visual blurring', 'Itching',
                'Irritability', 'delayed healing', 'partial paresis',
                'muscle stiffness', 'Alopecia', 'Obesity'],
               dtype='object')

In [29]: from sklearn.preprocessing import LabelEncoder
         le = LabelEncoder()
         for f in obj:
             X[f] = le.fit_transform(X[f].astype(str))

         print (X.info())

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 520 entries, 0 to 519
         Data columns (total 16 columns):
          #   Column              Non-Null Count  Dtype
         ---  ------              --------------  -----
          0   Age                 520 non-null    int64
          1   Gender              520 non-null    int32
          2   Polyuria            520 non-null    int32
          3   Polydipsia          520 non-null    int32
          4   sudden weight loss  520 non-null    int32
          5   weakness            520 non-null    int32
          6   Polyphagia          520 non-null    int32
```

Figure 3.2: Data Preprocessing

Normalization of data: Normalization is required when various attributes are on different scale. So various attributes must be normalized to bring onto the same scale.

Correlation among the attributes data



Correlation

```
In [31]: X.corrwith(y)

Out[31]: Age                   0.108679
         Gender               -0.449233
         Polyuria              0.665922
         Polydipsia            0.648734
         sudden weight loss    0.436568
         weakness              0.243275
         Polyphagia            0.342504
         Genital thrush        0.110288
         visual blurring       0.251300
         Itching              -0.013384
         Irritability          0.299467
         delayed healing       0.046980
         partial paresis       0.432288
         muscle stiffness      0.122474
         Alopecia             -0.267512
         Obesity               0.072173
         dtype: float64

In [32]: X.corrwith(y).plot.bar(
                 figsize = (16, 8), title = "Correlation with Diabetes", fontsize = 15,
                 rot = 90, grid = True)
```

Figure 3.3: Correlation Table

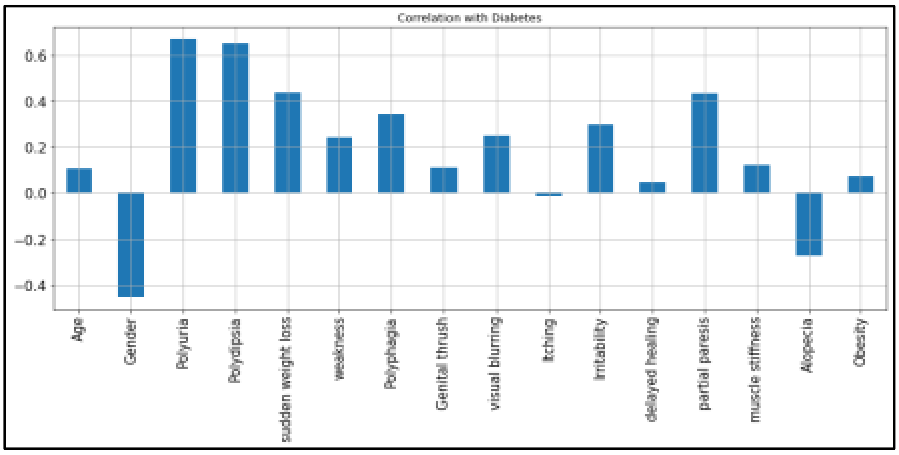
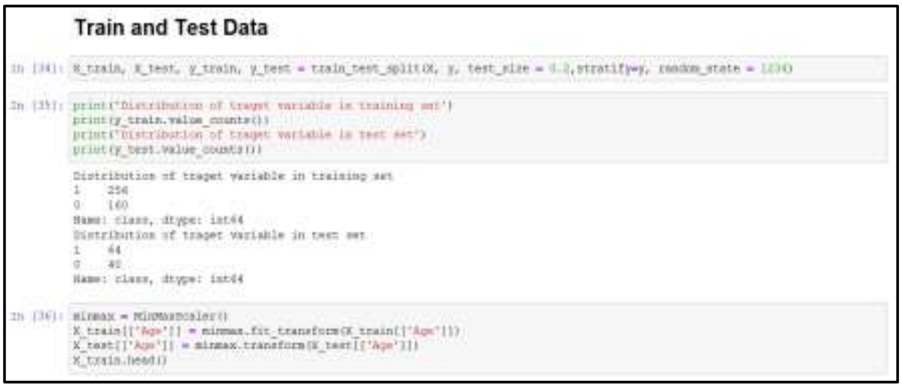


Figure 3.4: Correlation of Diabetes

Dividing into Train and Test Data



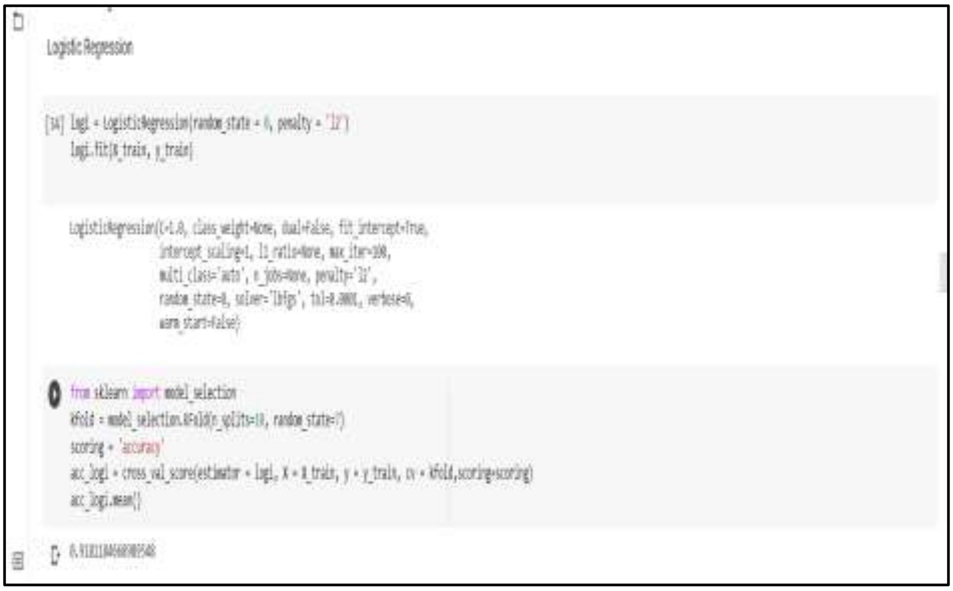
Train and Test Data

```
In [34]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,stratify=y, random_state = 1234)

In [35]: print('Distribution of traget variable in training set')
         print(y_train.value_counts())
         print('Distribution of traget variable in test set')
         print(y_test.value_counts())

         Distribution of traget variable in training set
         1    256
         0    160
         Name: class, dtype: int64
         Distribution of traget variable in test set
         1    64
         0    40
         Name: class, dtype: int64

In [36]: minmax = MinMaxScaler()
         X_train[['Age']] = minmax.fit_transform(X_train[['Age']])
         X_test[['Age']] = minmax.transform(X_test[['Age']])
         X_train.head()
```

Figure 3.5: Splitting into Train and Test Data

Feature Selection: Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.Having irrelevant features in the dataset can decrease the accuracy of the model. Benefits of feature selection are Accuracy increase, reduction in training time and reduction in over fitting. Various Feature Selection Methods are i**.** Univariate Selection ii. Feature Importance. iii. Correlation Matrix with Heatmap.

Model Building of Logistic Regression

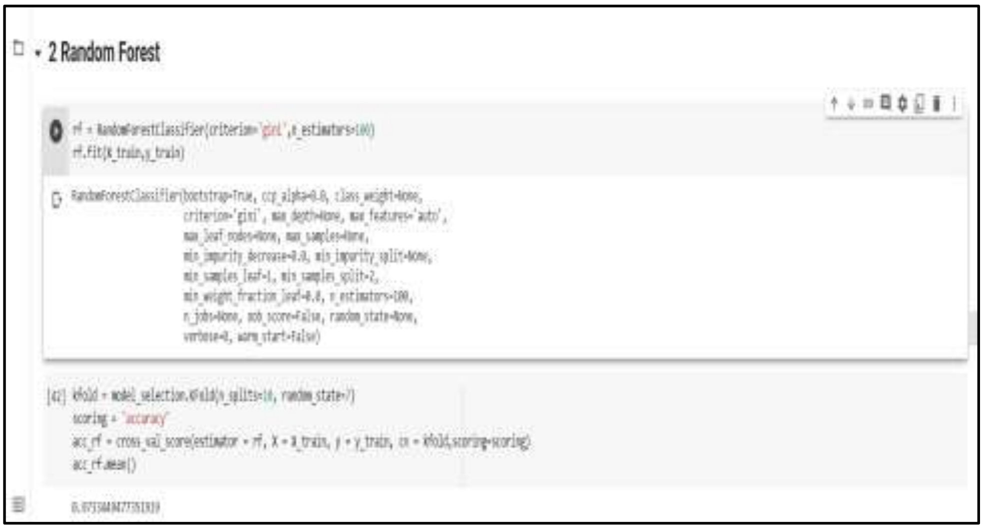


Logistic Regression

```
[34] logi = LogisticRegression(random_state = 0, penalty = 'l2')
     logi.fit(X_train, y_train)

     LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                        intercept_scaling=1, l1_ratio=None, max_iter=100,
                        multi_class='auto', n_jobs=None, penalty='l2',
                        random_state=0, solver='lbfgs', tol=0.0001, verbose=0,
                        warm_start=False)

     from sklearn import model_selection
     kfold = model_selection.KFold(n_splits=10, random_state=7)
     scoring = 'accuracy'
     acc_logi = cross_val_score(estimator = logi, X = X_train, y = y_train, cv = kfold,scoring=scoring)
     acc_logi.mean()

     0.9181184608989548
```

Figure 3.6: Model Building of Logistic Regression

Model Building of Random Forest Classifier

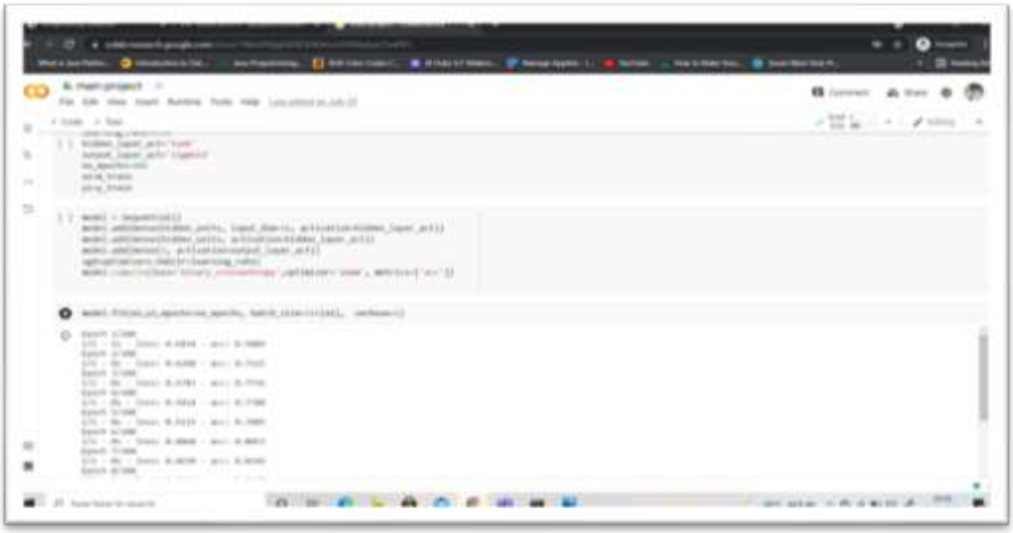


2 Random Forest

```
rf = RandomForestClassifier(criterion='gini',n_estimators=100)
rf.fit(X_train,y_train)

RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)

[42] kfold = model_selection.KFold(n_splits=10, random_state=7)
     scoring = 'accuracy'
     acc_rf = cross_val_score(estimator = rf, X = X_train, y = y_train, cv = kfold,scoring=scoring)
     acc_rf.mean()

     0.8733484773351919
```

Figure 3.7: Model Building of Random Forest

Model Building of Multilayer Perceptron



```
hidden_layer_act='tanh'
output_layer_act='sigmoid'
no_epochs=[illegible]
[illegible]_train
[illegible]_train

model = Sequential()
model.add(Dense(hidden_units, input_dim=[illegible], activation=hidden_layer_act))
model.add(Dense(hidden_units, activation=hidden_layer_act))
model.add(Dense(1, activation=output_layer_act))
[illegible]=optimizers.[illegible](learning_rate)
model.compile(loss='binary_crossentropy',optimizer='adam', metrics=['acc'])

model.fit([illegible], epochs=no_epochs, batch_size=[illegible], verbose=[illegible])

Epoch 1/100
[illegible] - loss: [illegible] - acc: [illegible]
Epoch 2/100
[illegible] - loss: [illegible] - acc: [illegible]
Epoch 3/100
[illegible] - loss: [illegible] - acc: [illegible]
Epoch 4/100
[illegible] - loss: [illegible] - acc: [illegible]
Epoch 5/100
[illegible] - loss: [illegible] - acc: [illegible]
Epoch 6/100
[illegible] - loss: [illegible] - acc: [illegible]
Epoch 7/100
[illegible] - loss: [illegible] - acc: [illegible]
Epoch 8/100
```

Figure 3.8: Model Building of Multilayer Perceptron

Confusion Matrix: There are 4 important terms in defining confusion matrix. These four parameters are used in calculation of performance metrics.

TP: the predicted result is YES and the actual result is also YES.	**TN**: The predicted result is NO and the actual result was also NO.
FP: The predicted result is YES and the actual result is NO.	**FN**: The predicted result is NO and the actual result is YES.

Figure 3.9: Formulae for Performance Metrics

Accuracy: Classification accuracy is perhaps the simplest metrics one can imagine, and is defined as the number of correct predictions divided by the total number of predictions.

Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier.

Recall: It is the number of correct positive results divided by the number of ***all*** relevant samples (all samples that should have been identified as positive).

F1 Score: F1 Score is the Harmonic Mean between precision and recall.

$Accuracy = \frac{TruePositive + TrueNegative}{TotalSample}$ **a.Precision**	$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$ **a.Precision**
$Precision = \frac{TruePositives}{TruePositives + FalseNegatives}$ **a.Precision**	$F1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}}$ **a.Precision**

Figure 3.10: Formulae for Performance Metrics

IV. EXPERIMENTS & RESULTS

Feature Selection and Correlation Matrix with Heatmap

Feature score calculation is donefor each feature, the higher the score the more important or relevant is the feature towards your output variable. Feature importance calculation is an inbuilt class with Tree Based Classifiers. Here for the taken dataset Extra Tree Classifier is used for extracting the top 10 features. Correlation defines how the features are related to each other or to the target variable. Correlation can be positive indicates that increase in one value of feature increases the value of the target variable or negative indicates that increase in one value of feature decreases the value of the target variable. Heatmap visualization makes it easy to identify the features that are most relevant to the target variable. The heatmap of correlated features is plotted using the seaborn library.

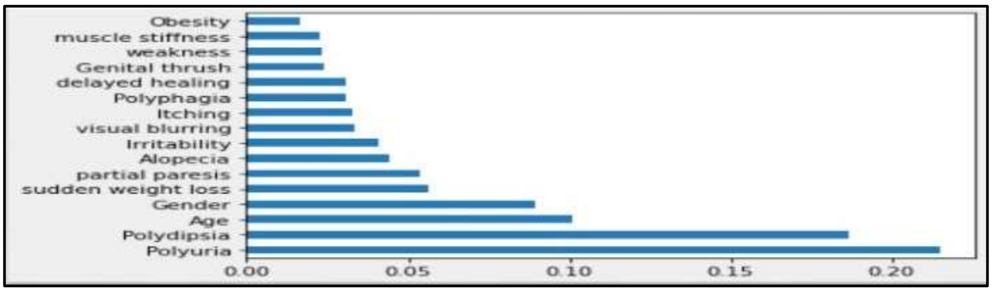
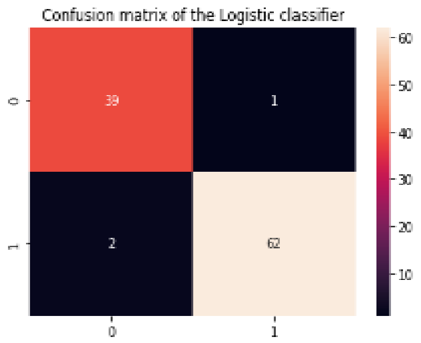
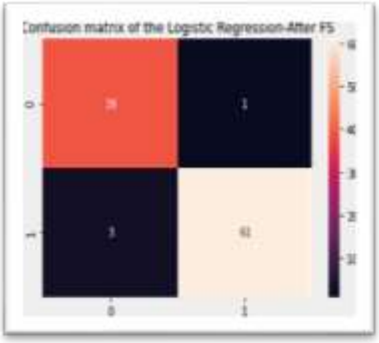
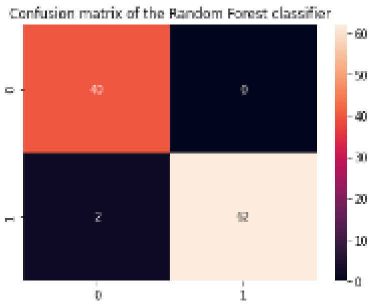
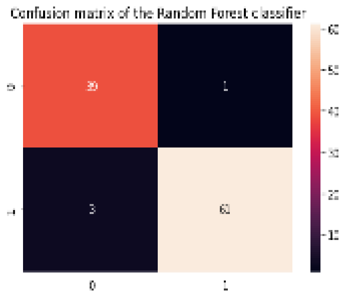
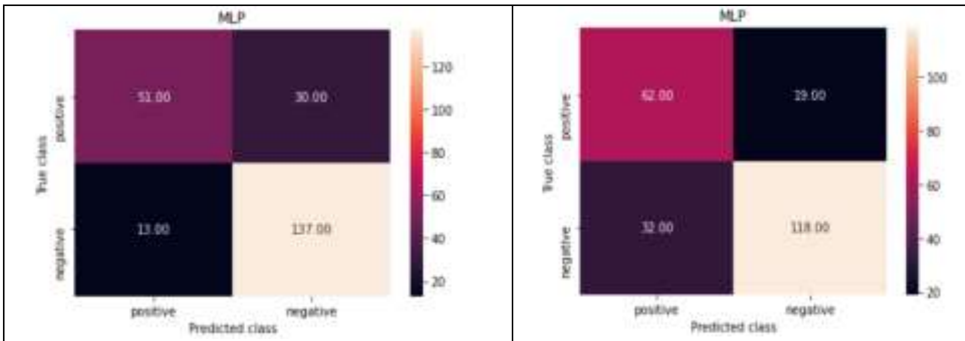


Figure 3.9 Correlation Heatmap

 **a.Logistic Regression before feature selection**	 **b.Logistic Regression after feature selection**
 **a.Random Forest before feature selection**	 **b.Random Forest before feature selection**



a.MLP with 1000 iterations

b.MLP with 300 iterations

Model Evaluation Table:

Model	Accuracy	Cross Val Accuracy	Precision	F1-score	ROC	Recall
LR	0.97115	0.91811	0.98412	0.97637	0.97187	0.96875
Random Forest	0.98076	0.97822	1.00000	0.98412	0.98437	0.96875
LR- FS	0.96153	0.89883	0.98387	0.96825	0.96406	0.95312
RF- FS	0.98076	0.96620	1.00000	0.98412	0.98437	0.96875

Model Evaluation Table for MLP:

Model	Accuracy	Precision	Recall	F1-score	Support
MLP(1000 Iterations)	0.779221	0.76	0.78	0.78	231
MLP(300 Iterations)	0.813853	0.81	0.77	0.81	231

V. Conclusion

The proposed NN based computational methods show not only much fewer attributes are required but also higher prediction and accuracy can be achieved.This work represents a Model, which is based on Neural Network. Firstly we use the best feature selection technique for identifying the top features which predicts the diabetes. Afterwards, we apply Deep Learning Model on the Dataset and get the output for it by that model. The reason behind using Deep Learning model is it predicts the result effectively and also with more accuracy. There is design of website for showing the prediction which understands to the user easily and automatically. The website consists of static home page which contains about Diabetes, why we get diabetes and also precautions to take for not to get diabetes. It also consists of meaning of attributes which are present in dataset . So that we can know that because of these symptoms, we may have chance of getting Diabetes. Finally, we want to conclude that Deep learning model predict the result correctly and also with high accuracy.

VI. References

[1] Son Vu Truong Dao. A Novel Wrapper—Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic. IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION.2020; 7869:7884.

[2] Muhammad Syafrudin, Jongtae Rhee. Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension. IEEE Access.2019; 144777:144789.

[3] Alessandro Aliberti, Enrico macii, Edoardo patti.A Multi-Patient Data-Driven Approach to Blood Glucose Prediction.IEEE Access. 2019;69311-69325.

[4] Nahla H. Barakat, Andrew P. Bradley. Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus. IEEE Transactions on information technology in biomedicine.2010; 1114-1120.

[5] JinyuXie and Qian Wang.Benchmarking Machine Learning Algorithms on Blood Glucose Prediction for Type I Diabetes in Comparison With Classical Time-Series Models.IEEE Transactions on biomedical engineering. 2020; 3101-3124.

[6] Md. kamrulhasan1 , Md. ashraful alam1 , dola das2 , eklashossain 3 , (senior member, ieee), and mahmudulhasan 2. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. *Phys Med Biol*.2013;58:97-129.

[7] Rama Naga Kiran Kumar K, Ramesh Babu I. The File System Recommendations to Reduce the Space and Time Parameters in Hadoop File Storage and Map Reduce Processing of Big Data Applications. 2020:2278:3075.