DIABETES PREDICTION USING MACHINE LEARNING ALGORITHMS

LOGESHWARAN

PRUTHIEV

RENITA

SREVARDHANI

SRUTHI

1. ABSTRACT

Diabetes is a chronic disease which increases blood sugar level and becomes severe when unnoticed or unidentified. The main aim of this paper is to build a better diabetes classification model (i.e.) to report whether the patient is diabetic or not. In this paper, we have proposed six different machine learning algorithm namely Logistic Regression, Support Vector Machine, Naïve bayes, Decision Tree, k-Nearest Neighbor and Random Forest to predict diabetes. We have performed data pre-processing to remove null values, outliers. Feature engineering is performed and new features are added to the dataset. Categorical variables are converted into numeric values with the help of label encoding. Accuracy of all the six different algorithms is evaluated for both noisy as well as preprocessed data. Classification accuracy is improved with hyper parameter tuning using grid search method. In this proposed work we achieved highest accuracy of 100% on training data and 90.74% on testing data for Random Forest algorithm.

2. INTODUCTION

Diabetes is the very common word that we hear in our present day to day life. Diabetes has become a great threat to human health all over the world. Diabetes mellitus is a chronic disease which can be caused by abnormal secretion of a hormone- insulin. Pancreas secretes insulin and malfunctioning of pancreas may lead to diabetes. Diabetic patients cannot properly absorb the glucose from the food they eat. Insulin allows glucose to enter into the body cells and use it for energy. Glucose is the source of energy that tissues and organs need to function properly. When cells are resistant or not responding to insulin hormone, glucose cannot be able to enter into the cells which ultimately increases the blood sugar level and thus results in diabetes. Frequent urination, feeling thirsty, increased hunger are some warnings of high blood glucose level.

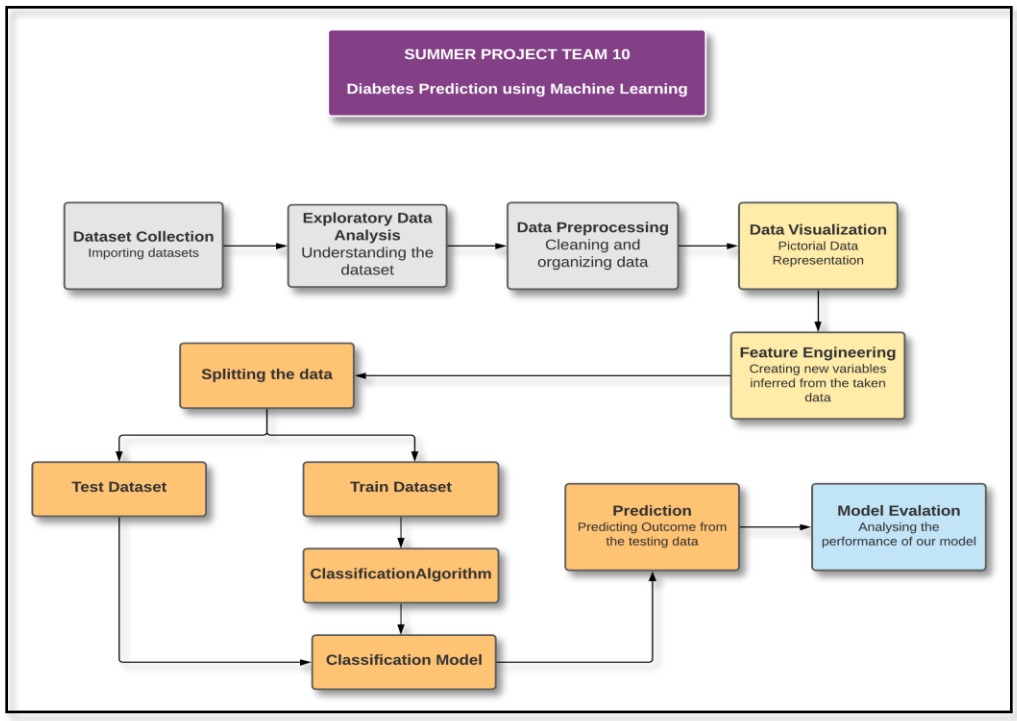
Statistical results in 2019 shows that 463 million people were living with diabetes worldwide, which will increase to 700 million by 2045 [1]. In our country, recent study by the Union Ministry of Family and Health Welfare published on January 6, 2021 that 9.3 percent senior citizens living in rural areas and 26.1 percent senior citizens living urban areas were diagnosed [2]. Though the number of cases of diabetes is similar between men and women, there

is a high prevalence among senior citizens. Statistical results demonstrates that diabetes had a arise from 4.7% to 8.5% from 1980 to 2014 among adults [3]. Diabetes may lead to life threatening complications if blood glucose level stays high for a longer duration. Complications include cardiovascular issues- coronary artery disease chest pain, stroke, high blood pressure, high cholesterol; nerve damage; kidney damage that may lead to kidney failure; eye damage- cataracts, glaucoma; foot damage; skin infections; erectile dysfunction; hearing loss; depression; dementia and also dental problems. These complications may also lead to death. Diabetes resulted in 4.2 million deaths approximately in 2019 [4]. The problem associated is, there is no long term cure, but chances of recovery is greater if early predictions are possible.

Machine Leaning techniques have become very useful in early disease predictions with the advancement in technology. In this work, six different machine learning algorithm namely Logistic Regression, Support Vector Machine, Naïve bayes, Decision Tree, k-Nearest Neighbor and Random Forest are used to predict diabetes. Pima Indians Diabetes data set was experimented in this work. Experimental performance of all the six algorithms are compared and highest accuracy of 100% on training data and 90.74% on testing data for Random Forest (RF) algorithm is achieved which shows the effectiveness of the RF in predicting the disease. The rest of the paper is structured as follows: Section-II describes about related work of various techniques of diabetes prediction, Section-III discusses the proposed work, Section-IV gives the experimental analysis and Section-V determines the conclusion of the work.

3. RELATED WORK

4. PROPOSED WORK



4.1 DATASET COLLECTION & EXPLORATORY DATA ANALYSIS

This research paper uses PIMA – Indians Diabetes dataset which is available in the UCI Machine Learning repository. The dataset contains 9 features and 768 records. Out of the 768 records 500 are Non diabetic and remaining 268 records are diabetic. EDA is the process of performing initial investigation on the dataset.

DESCRIPTION OF ATTRIBUTES:

```
Diabetes_dataset.info()
```

```
<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
```

4.2 DATA PREPROCESSING

In the first step data preprocessing is done. Preprocessing is the method by which we perform data cleaning i.e. raw dataset is converted into cleaned dataset. Missing values are replaced by the median values for improving the performance of the model. Outliers are removed with the help of Z score method. Features are scaled between 0 and 1 using Min-Max scaling technique.

4.3 DATA VISULAIZATION

Data visualization helps to understand the data and also explain the data to others. Histogram, density plot, box and whisker plot, correlation matrix plot, scatter matrix plot, pair plot have been plotted. In this paper we have found that there is a positive correlation between the features like: glucose level increases with age, insulin level increases with glucose level, skin thickness increases with insulin as well as with BMI. On plotting correlation matrix we found that glucose, BMI, Age is the most important features to identify whether the patient is diabetic or not.

4.4 FEATURE ENGINEERING

It is the process by which useful and relevant features are extracted from the raw dataset. It helps to avoid over fitting of the model. New features are created according to BMI, glucose and Insulin level. Label encoding method converts the categorical variable into numeric form which is machine readable form.

4.5 SPLITTING DATASET

PIMA Indians Diabetes dataset is divided into 70% training data and 30% testing data.

4.6 CLASSIFICATION ALGORITHM (MODEL BUILDING)

This is one of the most important phases in the machine learning where we have implemented 6 algorithms. These algorithm include Logistic Regression, Support Vector Machine, Decision Tree, Naïve Bayes, K Nearest Neighbor, Random Forest.

INPUT: Data values

OUTPUT: Prediction is made. Accuracy, confusion matrix, classification report is displayed.

PROCEDURE:

ML_algorithm_used = [LogisticRegression(), SVC(), DecisionTreeClassifier(), GuassianNB(), RandomForestClassifier(), KNeighborsClassifier()]

Initialize variable ‘i’ to 1;

While (i!=7) do

Build the model using DiabetesModel = ML_algorithm_used[i]

Implement the DiabetesModel for training data using DiabetesModel.fit()

Implement DiabetesModel classify for testing data points using DiabetesModel.predict()

Print the confusion matrix and classification report

Calculate accuracy for test data, train data and display it.

end while

4.7 HYPER-PARAMETER TUNING

The parameters which is used define the model architecture are called as hyper parameters. Hyper parameter tuning is the process of finding the ideal model architecture. Accuracy of the build model is increased with hyper- parameter tuning. Grid search is the basic method used for hyper parameter tuning.

4.8 EVALUATION

Evaluation is done by measuring the classification accuracy of the built model using the formula:

$$\text{ACCURACY} = \frac{\text{Total number of correct prediction made}}{\text{Total number of predictions made}}$$

5. EXPERIMENTAL ANALYSIS

The machine learning algorithms that we implemented for diabetes classification and their Performance Measures

5.1 Based On Accuracy:

Machine Learning Algorithm	After performing Hyper Paramter Tuning
Logistic Regression	Accuracy of Training Data : 85 Accuracy of Testing Data : 86
Support Vector Machine	Accuracy of Training Data : 92 Accuracy of Testing Data : 85
Decision Tree	Accuracy of Training Data : 90 Accuracy of Testing Data : 86
Naive Bayes	Accuracy of Testing Data : 89
Random Forest	Accuracy of Training Data : 100 Accuracy of Testing Data : 90
K Nearest Neighbour	Accuracy of Training Data : 89 Accuracy of Testing Data : 86

- Random Forest outperforms with 90% accuracy on test data and 100% accuracy on train data

5.2 Based On Confusion Matrix :

Machine Learning Algorithms	Total Instances	Correctly Classified	Incorrectly Classified
Logistic Regression	216	186	30
Support Vector Machine	216	185	31
Decision Tree	216	187	29
Naïve Bayes	216	162	54
Random Forest	216	191	25
K Nearest Neighbour	216	187	29

- Random forest has the highest, lowest number of correctly and incorrectly classified samples.

5.3 Based On Classification Report :

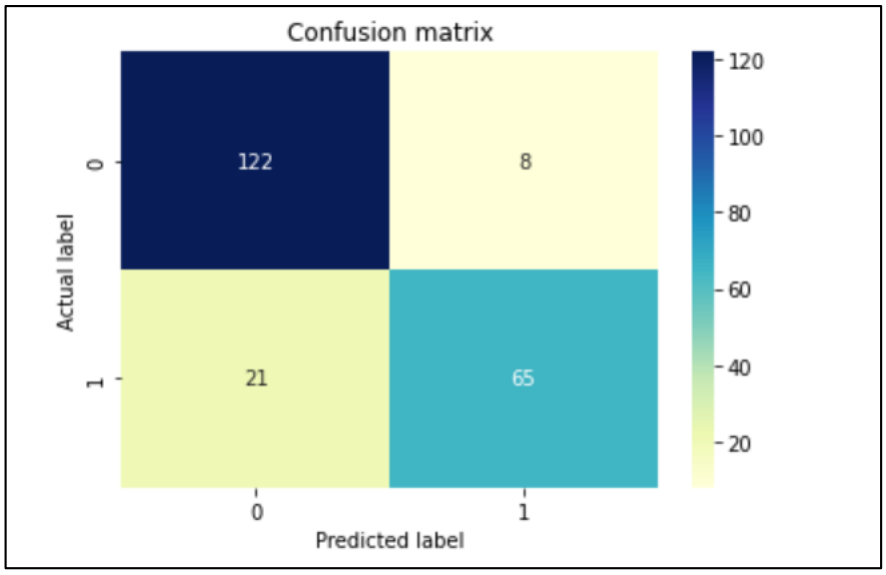
Machine Learning Algorithms	Precision	Recall	f1-score
Logistic Regression	0.86	0.86	0.86
Support Vector Machine	0.86	0.86	0.85
Decision Tree	0.87	0.87	0.86
Naïve Bayes	0.77	0.75	0.73
Random Forest	0.89	0.88	0.88
K Nearest Neighbour	0.87	0.87	0.86

- Random forest has the highest Precision , Recall and f1-score Values
- From the observations above , Random Forest would be the best algorithm for diabetes classification

6. CONCLUSION:

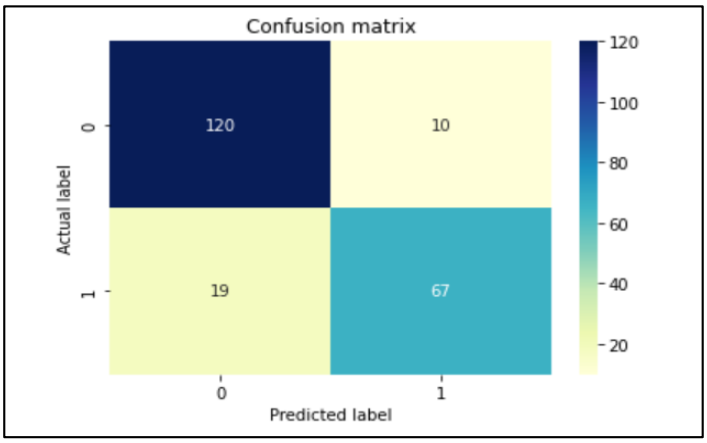
In this paper, various machine learning algorithms has been applied on the dataset. The classification has been done using many algorithms in which random forest gives highest accuracy of 89.35%. After hyperparameter tuning 88.42%. This model enhance the accuracy of diabetes prediction with this dataset compared to pre existing dataset. Further this work can be improved to find how likely non-diabetic people can have diabetes in the following years.

DECISION TREE



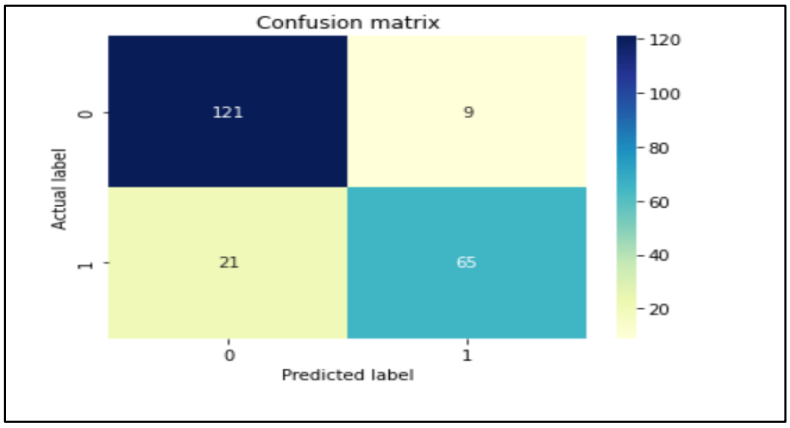
	precision	recall	f1-score	support
0	0.85	0.94	0.89	130
1	0.89	0.76	0.82	86
accuracy			0.87	216
macro avg	0.87	0.85	0.86	216
weighted avg	0.87	0.87	0.86	216

KNN



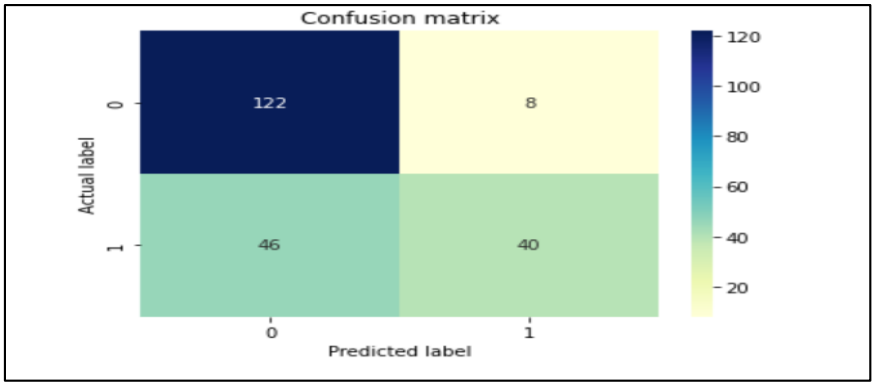
	precision	recall	f1-score	support
0	0.86	0.92	0.89	130
1	0.87	0.78	0.82	86
accuracy			0.87	216
macro avg	0.87	0.85	0.86	216
weighted avg	0.87	0.87	0.86	216

LOGISTIC REGRESSION



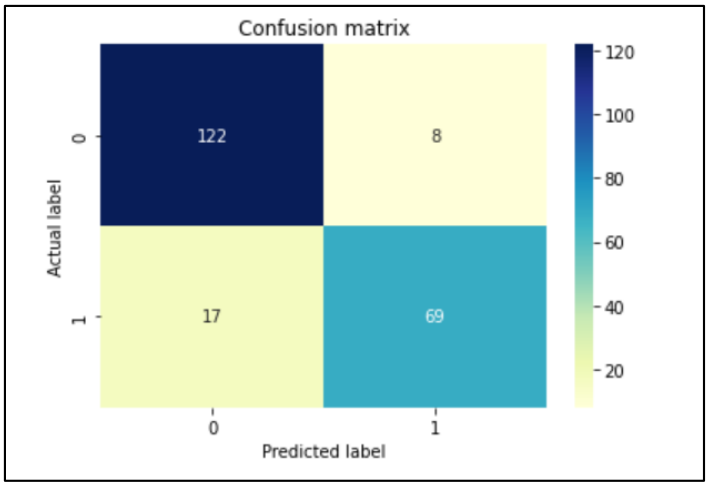
	precision	recall	f1-score	support
0	0.85	0.93	0.89	130
1	0.88	0.76	0.81	86
accuracy			0.86	216
macro avg	0.87	0.84	0.85	216
weighted avg	0.86	0.86	0.86	216

NAIVE BAYES



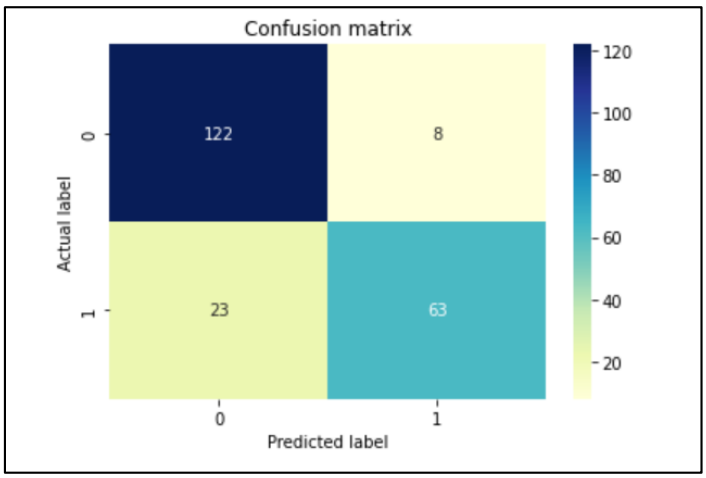
	precision	recall	f1-score	support
0	0.73	0.94	0.82	130
1	0.83	0.47	0.60	86
accuracy			0.75	216
macro avg	0.78	0.70	0.71	216
weighted avg	0.77	0.75	0.73	216

RANDOM FOREST



	precision	recall	f1-score	support
0	0.73	0.94	0.82	130
1	0.83	0.47	0.60	86
accuracy			0.75	216
macro avg	0.78	0.70	0.71	216
weighted avg	0.77	0.75	0.73	216

SVM



	precision	recall	f1-score	support
0	0.84	0.94	0.89	130
1	0.89	0.73	0.80	86
accuracy			0.86	216
macro avg	0.86	0.84	0.84	216
weighted avg	0.86	0.86	0.85	216

REFERENCES:

[1] https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

[2] https://www.downtoearth.org.in/news/health/diabetes-among-senior-citizens-more-prevalent-in-urban-india-lasi-report-75038

[3] https://ieeexplore.ieee.org/abstract/document/9076634

[4] https://en.wikipedia.org/wiki/Diabetes

https://ieeexplore.ieee.org/abstract/document/8965556

https://d1wqtxts1xzle7.cloudfront.net/54976577/IRJET-V4I1077-with-cover-page-v2.pdf?Expires=1632653395&Signature=gXunJwptESA9IF5fjGIycv4lG71EIRjLsb~-ksoq-9NvONQAHWXtTI9y9zIhoFgfJA-sB9kY5ir0rKz9~yaMKywI~sxvawi97o2BrLX3FoUHcltiY5mEhCdk5EuFseRaSHMeRRAtLzxUckI9~ti6zxzmL60NSMlSTrDIR-w5Ks~3XGFEeogRPunFXS0sUKUUVsz2V0oAbQLqU9XFYIzwkwjIjg6-fEtxy0-pjOs8VzVDlt467syENSpU3hRKbWWqsJug17aC~kRtU9r0NHTZao9CJA8jkNELU3n6

https://www.sciencedirect.com/science/article/pii/S1877050920300557

https://ieeexplore.ieee.org/abstract/document/8614871