**FOUNDATIONS OF DATA SCIENCE**

**ASSIGNMENT 1**

Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm

SUBMITTED BY:

Pasham Raghu Priya

2018A7PS0241U

SUBMITTED TO:

Dr. Siddhaling Urolagin

**Objectives**

Risk prediction of diabetes using Machine learning techniques.

Accuracy and ROC metrics used to find best possible machine learning technique.

**Introduction**

Diabetes is harmful disease which cuts human life at early age. Now diabetes cases are increasing enormously. People having diabetes are prone to other diseases like blindness, heart attack, kidney problems and many more [[1]](about:blank). Diabetes Mellitus (DM) is metabolic disorders grouped together mainly which are caused by insulin action or insulin secretion. Insulin is secreted means increase of glucose levels i.e., hyperglycaemia. DM has become one of most ordinary endocrine disorder where more than 200 million people gets affected worldwide.

Common symptoms of diabetes are polydipsia, polyuria, polyphagia, obesity (type 2), weakness, sudden weight loss (type 1), delayed healing and many more. These common symptoms are used as attributes to predicting diabetes in this paper.

There are 6 different types of diabetes. It is hard to diagnose properly and distinguish between types of diabetes because of complexity of condition. But precise grouping will give diagnosis and can give proper treatment to people. Machine learning algorithms have power to classify and predict. There will never be one single technique that gives best performance and accuracy for all diseases. One technique shows high performance with one data set and for other data set this might not show same performance. The new study concentrates on combining different techniques from decision trees, Bayesian models random forest, neural models and many others for DM classification and prediction. In this new study different machine learning techniques are to predict diabetes mellitus at much early stage. In this paper algorithms used are Logistic Regression, Decision Tree, and Random Forest for prediction and compare those predictions to get final prediction[[2]](about:blank).

**Dataset**

The dataset is taken from [https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.Table

Description automatically generated](https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.)

**Literature survey**

In [[1]](about:blank) this paper, predictions are done using decision tree-based random forest (RF), function-based multilayer perception (MLP) and radial basis function network (RBF). These models have been used to do diagnosis of diabetes. 10-fold cross validation is used for evaluating accuracy. They found that RBF gave best accuracy in cross validation test for diabetes prediction. In [[2]](about:blank) this paper, researchers used Naïve Nets, decision stump, SVM and ensemble model for prediction. High accuracy is achieved using SVM algorithm in this paper. Performance is found using 10-fold cross validation. In [[3]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6232260/#B14) this paper, minimum redundancy maximum relevance (mRMR) and principal component analysis (PCA) models are used to reduce dimensionality. Machine learning algorithms used in this paper are random forest, decision tree and neural network. Prediction with random forest gives best performance. Method used for examining models is 5-fold cross validation. In [[4]](about:blank) this paper, prediction methods used are neural network, C5.0, Basin network and support vector machine. The best result is obtained by C5.0 which showed highest accuracy and sensitivity. In [[5]](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6) this paper, analysis of the features are performed and optimal features based on correlation values are selected. Decision tree algorithm and random forest algorithms holds best for analysis of diabetic data as they have highest specificity of 98.20% and 98%. Support vector machine and NB gives the accuracy of 77.3% and 73.48% from the existing method and therefore proposed method improves the accuracy of the classification techniques. In [[6]](https://dl.acm.org/doi/pdf/10.1145/3388218.3388231) this paper, four most popular algorithms of machine learning are used. Algorithms used are SVM, decision consecutively and neural network. Best performance is given by SVM however training SVM is difficult on large-scale because of its high time complexity O (N3). In [[7]](http://www.ijstr.org/final-print/jan2020/Prediction-Of-Diabetes-Using-Machine-Learning-Classification-Algorithms.pdf) this paper, SVM, Decision tree, KNN, Logistic Regression, Random forest are the algorithms used for the predictions of diabetes and they concluded that random forest has best performance of 75% accuracy rate. In [[8]](http://sersc.org/journals/index.php/IJAST/article/view/23972/12520) this paper, Decision Tree, SVM and Naïve Byes are used to detect early diabetes and data mining methods like Logistic regression, artificial neural network, random forest is used. Precision, Accuracy, F-measure and recall are the evaluate metric used to measure the three algorithms. Results have shown that naïve byes have the highest accuracy of 76.30% than others. In [[9]](https://academic.oup.com/jamia/article-abstract/27/9/1343/5899846?redirectedFrom=fulltext) this paper, prediction of diabetes mellitus complications is done using machine learning algorithms. Prediction is done using RNN, LSTM, and RNN gated recurrent unit. The RNN GRU model has achieved best performance for electronic record type. In [[10]](https://www.irjet.net/archives/V7/i2/IRJET-V7I2697.pdf) this paper, the algorithms used for predicting diabetes are Naïve bayes theorem, gradient boasting algorithm and SVM. SVM gives better performance.

**Methodology**



















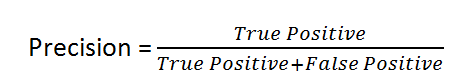


Firstly, data is pre-processed for the given dataset. Pre-processing means to fill the missing values, remove duplication and data conversion. Next, apply machine learning algorithms to training set (data after being pre-processes) for prediction. Algorithms used for prediction are Logistic regression, Random forest, and decision trees.

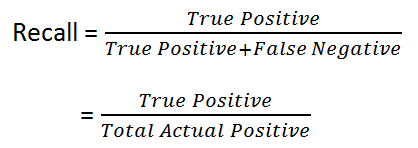
Prediction algorithms

1. Logistic Regression- Logistic Regression is machine learning algorithm used when dependent variable is categorical. The output for logistic regression is based on opportunity feature.
2. Decision Trees- Decision tree is a tree shaped diagram that is used for determining a course of action. Each branch of the tree represents a possible decision
3. Random Forest- Random forest is a versatile, easy-to-use machine learning algorithm that, in most cases, produces excellent results even without hyper-parameter tuning.

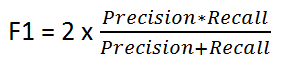
For all algorithms, confusion matrix plot and classification report are done. Classification report consists of finding precision, recall and f1-score.

Precision 

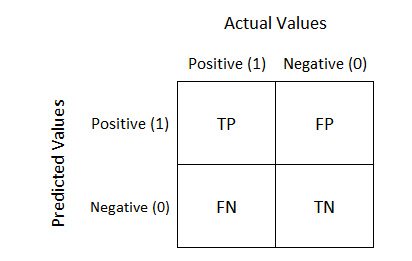
Recall



F1-score



Confusion matrix



**Experimental Setup**

The dataset is taken from UCI Machine learning repository. First, I installed required libraries using pip. Data is split twice for easy way to train. In first split, 80% of train data and 20% of test dat. In second split, train data is split further to 80% train data and 20% test data. Then ‘class’ feature from dataset is used to predict positive and negative values. Later following machine algorithms are applied,

1. Logistic regression implementation – from sklearn we imported LogisticRegression, SimpleImputer, make\_pipeline and roc\_auc\_score. Category\_encoders are imported. Pipeling and data fitting is done initially. Prediction is done to find accuracy and ROC score. Then confusion matrix plot is implemented. At the end classification report is made.

1. Decision Tree implementation - from sklearn we imported DecisionTreeClassifier, SimpleImputer, make\_pipeline and roc\_auc\_score. Category\_encoders are imported. Pipeling and data fitting is done initially. Prediction is done to find accuracy and ROC score. Then confusion matrix plot is implemented. At the end classification report is made.
2. Random Forest Model- from sklearn we imported RandomForestClassifier, SimpleImputer, make\_pipeline and roc\_auc\_score. Category\_encoders are imported. Pipeling and data fitting is done initially. Prediction is done to find accuracy and ROC score. Then confusion matrix plot is implemented. At the end classification report is made.

**Results**

| Models | Validation Accuracy | ROC Accuracy |
| --- | --- | --- |
| Decision Tree | 0.976190 | 0.971874 |
| Random Forest | 0.964285 | 0.999609 |
| Logistic Regression | 0.916666 | 0.983984 |

Confusion matrix gives us insight of errors done by classifier.

Confusion matrix plot for logistic regression

**Chart

Description automatically generated**

29 people were predicted yes, and they have disease. 48 were predicted no and they do not have disease. 3 were predicted yes, but they do not actually have disease. 4 were no, but they do have the disease.

Confusion matrix tree for decision tree**Chart, treemap chart

Description automatically generated**

31 people were predicted yes, and they have disease. 51 were predicted no and they do not have disease. 1 were predicted yes, but they do not actually have disease. 1 were no, but they do have the disease.

Confusion matrix tree for random forest

**Chart

Description automatically generated**

30 people were predicted yes, and they have disease. 51 were predicted no and they do not have disease. 2 were predicted yes, but they do not actually have disease. 1 were no, but they do have the disease.

Classification report with Logistic regression

Table

Description automatically generated

Classification report with Decision Tree

Table

Description automatically generated

Classification report with Random Forest

Table

Description automatically generated

**Conclusion**

Detecting the processing of different heart related diseases which will help humans in long run. Dataset from UCI Machine learning repository is taken for prediction. Machine learning techniques are used for processing raw data. The techniques used are Logistic Regression, decision tree and Random forest and their respective validation accuracies and ROC accuracies are 0.916666 and 0.983984, 0.976190 and 0.971874, 0.964285 and 0.999609. There is very less accuracy difference between decision tree and Random forest. At the decision tree is more helpful with prediction.

**Recommendations**

1. Comparative Analysis and Risk Prediction of Diabetes at Early Stage using Machine Learning Approach. International Journal of Future Generation Communication and Networking. 4151-4163.
2. Alehegn, Minyechil & Joshi, Rahul & Mulay, Preeti. (2018). Analysis and prediction of diabetes mellitus using machine learning algorithm. International Journal of Pure and Applied Mathematics. 118. 871-878.
3. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting Diabetes Mellitus With Machine Learning Techniques. *Front Genet*. 2018;9:515. Published 2018 Nov 6. doi:10.3389/fgene.2018.00515
4. [Prediction\_and\_Diagnosis\_of\_Diabetes\_by\_Using\_Data](about:blank)
5. 5. Sneha, N., Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. *J Big Data* **6,**13 (2019).
6. 6. Abir Al-Sideiri, Zaihisma Binti Che Cob, and Sulfeeza Bte Mohd Drus. 2019. Machine Learning Algorithms for Diabetes Prediction: A Review Paper. In Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control. Association for Computing Machinery, New York, NY, USA, 27–32. DOI:https://doi.org/10.1145/3388218.3388231
7. <http://www.ijstr.org/final-print/jan2020/Prediction-Of-Diabetes-Using-Machine-Learning-Classification-Algorithms.pdf>
8. <http://sersc.org/journals/index.php/IJAST/article/view/23972/12520>
9. <https://academic.oup.com/jamia/article-abstract/27/9/1343/5899846?redirectedFrom=fulltext>
10. <https://www.irjet.net/archives/V7/i2/IRJET-V7I2697.pdf>
11. <https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb>