,

Diabetes Prediction using different Machine Learning and Ensemble Classifiers

Era Sharma *General , M.Tech*
IIIT Delhi
New Delhi , India
era19121@iiitd.ac.in

Gaurav Lodhi *General , M.Tech*
IIIT Delhi
New Delhi , India
gaurav19063@iiitd.ac.in

Mansi Sharma *General , M.Tech*
IIIT Delhi
New Delhi , India
mansi19092@iiitd.ac.in

Abstract

Diabetes is a metabolic disease that causes high blood sugar . The risk factors and complication due to diabetes can be significantly reduced if the early prediction is possible.The accurate prediction of diabetes is different because of the outliers and missing data. In this paper, We will use framework where Missing values are filled , Standardisation of data , Feature Selection , K-Fold Cross Validation and different machine learning classifier along with deep learning model were employed to predict the diabetes. Along with it , different Ensembling methods were used to improve the performance of the prediction . The learning of data using models makes even more sense when parameter are tuned accurately which will be done by grid search technique.As an improvement ensemble models are used and tested as both individually and combined.

1 Paper Summary

1.1 Aim

Diabetes is a known disease now a days . Its a metabolic disease The insulin hormone produced by the pancreas moves sugar from blood to cells .The lack of that hormone due to malfunctioning of the pancreas forms diabetes which can result in coma, renal and retinal failure, pathological destruction of pancreatic beta cells, cardiovascular dysfunction, cerebral vascular dysfunction, peripheral vascular diseases, sexual dysfunction, joint failure,

weight loss, ulcer, and pathogenic effects on immunity. Researches have shown that diabetes patients has increased from 4.7 % to 8.5 % in 1980 to 2014 and is still growing.

There is no long term cure for diabetes, but it can be controlled and prevented if its detected at an early stage

Hence our aim is to predict accuractely whether a person is suffering from diabetes or not using PIMA dataset. We took the dataset from PIMA specifically , PIMA Indians Diabetes (PID) dataset of 768 female diabetic patients from the Pima Indian population near Phoenix. This dataset consists of 268 diabetic patients (positive) and 500 non-diabetic patients (negative) with eight different medical features attributes.

1.2 Methodology

1. Firstly , **processing** of the data is done on PIMA dataset which includes checking and filling null values , outlier detection , Standardistation of the data and Feature extraction .

   - Filling NULL values : To fill the missing values , mean value of the feature is assigned.
   - Outlier detection : Machine learning algorithms are sensitive to the range and distribution of attribute values and outliers can mislead the training process resulting in longer training times, less accurate models and ultimately poorer results.So removing outliers is important.

     According to the study, Outlier are removed with the help of:

     $$P(x) = (x \qquad if(Q11.5IQRxQ3 + 1.5IQR)$$
     $$reject \qquad otherwise$$

     Where, X :–> Instances of Multidimensional Features vector

     Q1,Q2,IQR :–> first quartile, third quartile, and interquartile range of the attributes respectively .
   - Standardistation :Standardisation of the data is important if the dataset has range of values. It will transform our data in a manner that it has mean as 0 and standard deviation as 1. Standardization is useful for data which has negative values. For this we made use of sklearn, Standardscaler library.
   - Feature Extraction : The feature extraction is an important step in the machine learning due to the curse of dimensionality . As the features are more , data can become more sparser and can overfit.

     For this we made use of three extraction techniques which are :

     (a) Principal Component Analysis
     (b) Select-K best
     (c) Correlation-based relation

2. **Cross Validation** Cross validation is the common techniques to estimate the error and for the selection of the correct models . In this project, we made use of 5 fold

cross validation i.e. 4 out of 5 parts of the training data will be used for training and 1 part will be used for testing and will repeated 5 times. To make use of this, We used cross-validate library from sklearn model-selection.

3. **Machine Learning models** We made used of different machine learning models and Deep learning such as :

   - Decision Tree
   - K-NN
   - MLP
   - CNN
   - Random Forest
   - Naive Bayes
   - Xg-boost
   - Adaboost
   - Bagging

   Ensembling of classifiers is done by ensembling 2,3,4,5,6 models with each other.these models were the models as in table 1.

   - Adaboost + Xgboost
   - KNN + Xgboost + Decision Tree
   - Adaboost + KNN + Xgboost + Decision Tree
   - KNN + Xgboost + Decision Tree + Random forest + Naive bayes
   - KNN + Xgboost + Decision Tree + Random forest + Naive bayes + Adaboost

4. **Grid Search** The parameter are the important factor of the machine learning models . The correct hyper parameters will train the data accurately hence predictions will be more accurate . For different machine learning used , the different hypermeter were tunned with the help of GridsearchCV library of the sklearn and test data will predicted on the best parameter of the above models.

1.3 Results

We tested each model with 30 % data on the optimal parameters of the models and calculated the accuracy as shown in table 1 .

Ensembling models results are shown in table 2 .

1.4 conclusion

After looking at the performance of each model we can conclude that adaboost and naive bayes model performed better with accuracy 76% followed by Xg boosting with accuracy of 75% than all other models in the paper also they concluded that ensemble methods boosting performed better than all other methods

TABLE 1: Classification Report (%) of different models

Models	Precision	Recall	F1-score	Accuracy
KNN	65.5	65	65.5	69
Decision tree	69	70.5	69	71
Xg boosting	73	73	73	75
Multi layer perceptron	71	70.5	71	74
Random Forest	72.5	73	73	75
Adaboosting	72.5	72.5	72.5	75
Naive Bayes	72	69.5	70.5	74
Bagging	71	71.5	71	74

TABLE 2: classification Report (%) of different models

Models	Precision	Recall	F1-score	Accuracy
ab+xb	86.1	78.3	82.01	89.3
knn+xg+dt	84.4	77.8	80.9	88.5
ad+knn+xg+dt	87.3	76.3	81.43	89.1
knn+xb+dt+rf+nb	83.6	76.8	80.05	88.0
knn+xb+dt+rf+nb+ad	81.3	77	81	89

2 New Analysis

2.1 Objective

Our objective in this entire project was not only to improve the accuracy but also do the exploratory data analysis to see the impact of different factors on the outcome. We used different techniques to find these factors as well as to improve accuracy. One more objective we have was to find reproducibility of the proposed solution.

2.2 Improvement over Previous Work

There are some improvements we did in our approach.

2.2.1 Exploratory Data analysis:

- We used sns heatmap to visualize the correlation between the features and to find the impact of one feature on the outcome. we found some interesting observations from this analysis
- Glucose and age is heigh correlated to the outcome.
- Age is heighly correlated to the pregnencies so one feature can represent both the features.
- Blood pressure having very less correlation with the outcome. We can eliminate this feature for predicting the outcome.

2.2.2 Model improvement:

- Model accuracy in the selected paper was heigh and was the state of art model. But still we were able to improved the accuracy from 88% to 90% by using Keras Supported sequential model.
- We also used accuracy weighted ensemble methods to improve stability of accuracy across the models. Previously used models were heighly oriented towards the roc auc score but this model is more stable for accuracy measure.

2.3 Conclusion

This was the state of the art heighly accurate model in the field of diabetes detection. We made this model more accurate by using our accuracy weighted approach. We also implemented a model by using Keras supported sequential model which is giving us significant improvement than the previous models.

2.4 Relevance

Used novel approaches are heighly related to the previous research work. We worked in accuracy improvement as well as explore the possibilities in the dataset to find more useful insights.

3 Instructions to Implement Code

We are using jupyter notebook to build the project.There are some requirements for implementation of the functional code these are following:

Required libraries:

- Sklearn
- pandas
- numpy
- matplotlib
- seaborn
- matplotlib

There are following some steps to run the code successfully:
Input: It takes dataset.csv file as input dataset.

- set the path variable for the dataset.csv location.
- Run the jupyter notebook cells sequentially.

Results:

Output: It produces output for different ensemble models for different 5-folds. This output contains different metric scores(eg. precision, recall, f1-score) for each model. In the last it produces output average of overall ensemble methods.

References

[1] Data: https://www.kaggle.com/uciml/pima-indians-diabetes-database

[2] A. Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. Reza-Albarrán, and K. L. Ramaiya, "Diabetes in developing countries," J. Diabetes, vol. 11, no. 7, pp. 522–539, Mar. 2019.

[3] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm based feature selection and MOE fuzzy classification algorithm on pima indians diabetes dataset," in Proc. Int. Conf. Comput. Netw. Informat. (ICCNI), Oct. 2017, pp. 1–5.

[4] The Emerging Risk Factors Collaboration, "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: A collaborative meta-analysis of 102 prospective studies," Lancet, vol. 375, pp. 2215–2222, Jun. 2010.

[5] N. H. Cho, J. E. Shaw, S. Karuranga, Y. Huang, J. D. da Rocha Fernandes, A. W. Ohlrogge, and B. Malanda, "IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," Diabetes Res. Clin. Pract., vol. 138, pp. 271–281, Apr. 2018