DIABETES PREDICTION USING MACHINE LEARNING TECHNIQUES

Final year project Review-1 Phase - 2
Batch – (2019-2023)

Group Members : Abhay Kumar, Sungjemkaba
Reg_no : 2019105194, 2019105219

17-Feb-2023

Department of Computer Science and Engineering

National Institute of Technology Nagaland

Contents

- Aim and objective
- Glimpse of previous review
  - Research gap
  - Data Pre- Processing
  - Balancing the dataset
- Proposed model
- Accuracy gain
- Comparative analysis
- Conclusion
- Reference

Aim & objective

- Our primary aim is to predict the diabetes using ensemble learning techniques.
- Measuring the accuracy of the model.
- Balancing the dataset is our secondary objective.

Glimpse of previous review

- Total no of hypotheses in PIMA diabetes dataset is 768, out of which 268 is predicting it to be 1(having diabetes) rest is 0(not having diabetes), so the dataset is imbalance that's why balancing of dataset is required.

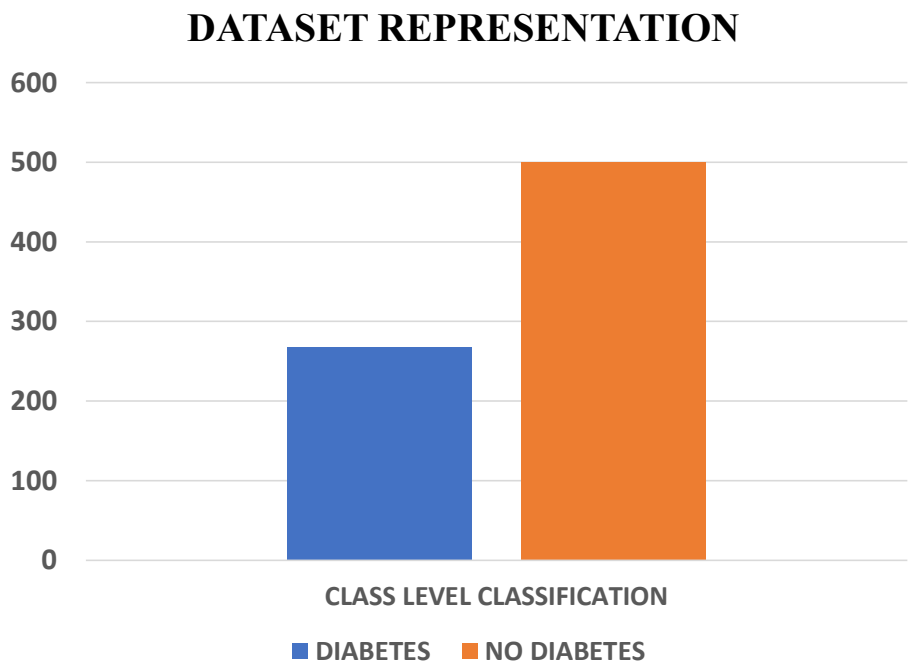


Fig No : 1

Research gap

- Using the KNN , Random forest classifier and 5-fold cross validation method ,by removing the noise they get an accuracy of 73.82%.
- Using the Naive Bayesian classifier and partition with the help of 10 fold cross validation The performance was evaluated using the measures of the accuracy, the precision, recall, and the F-measure. The highest accuracy was obtained by the Naive Bayes, which reached 74.30% .

Data pre-processing

- Out of 768 hypothesis column of some attribute contain 0 value.
- That particular attribute value has been replaced with the mean of that particular column.

Table representing the null value attribute

Total no of rows	768
Pregnancies	111
Glucose	5
Blood pressure	35
Skin thickness	227
Insulin	374
BMI	11
Diabetes pedigree function	0
Age	0

Table No : 1

Balancing the dataset

- Dataset will be balanced using oversampling, under sampling.
- In oversampling we have only 268 samples of 1 class level value , so the dataset have been randomly selected and made it to 500. Now the dataset is balanced.
- In under sampling we have only 268 samples of 1 class level value so only 268 samples have been selected of class level 0 value in order to balance the dataset.

Correlation

Table Representing Correlation between different attribute and class level value.

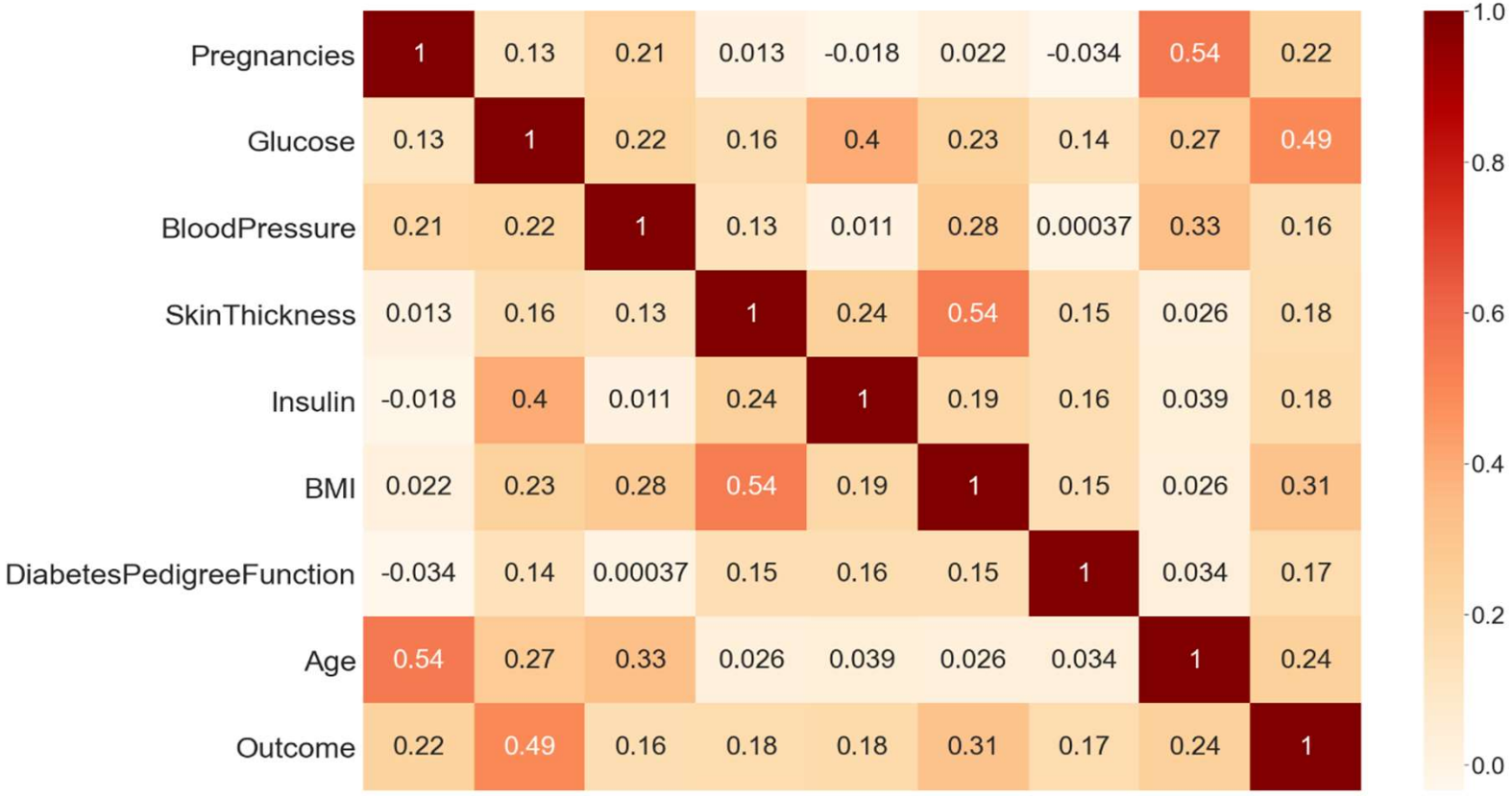


Fig No : 2

Correlation contd.

- A statistical measure called correlation shows how much two or more variables fluctuate in connection to one another.
- When two variables rise or decrease simultaneously, there is a positive correlation, when there is a negative correlation, one variable increases as the other falls.
- On the basis of that we will decide which attribute has the higher effect on the output data.
- The correlation between Blood Pressure and output is minimum i.e (0.16) and that of output and glucose is maximum i.e (0.49).

Proposed Model

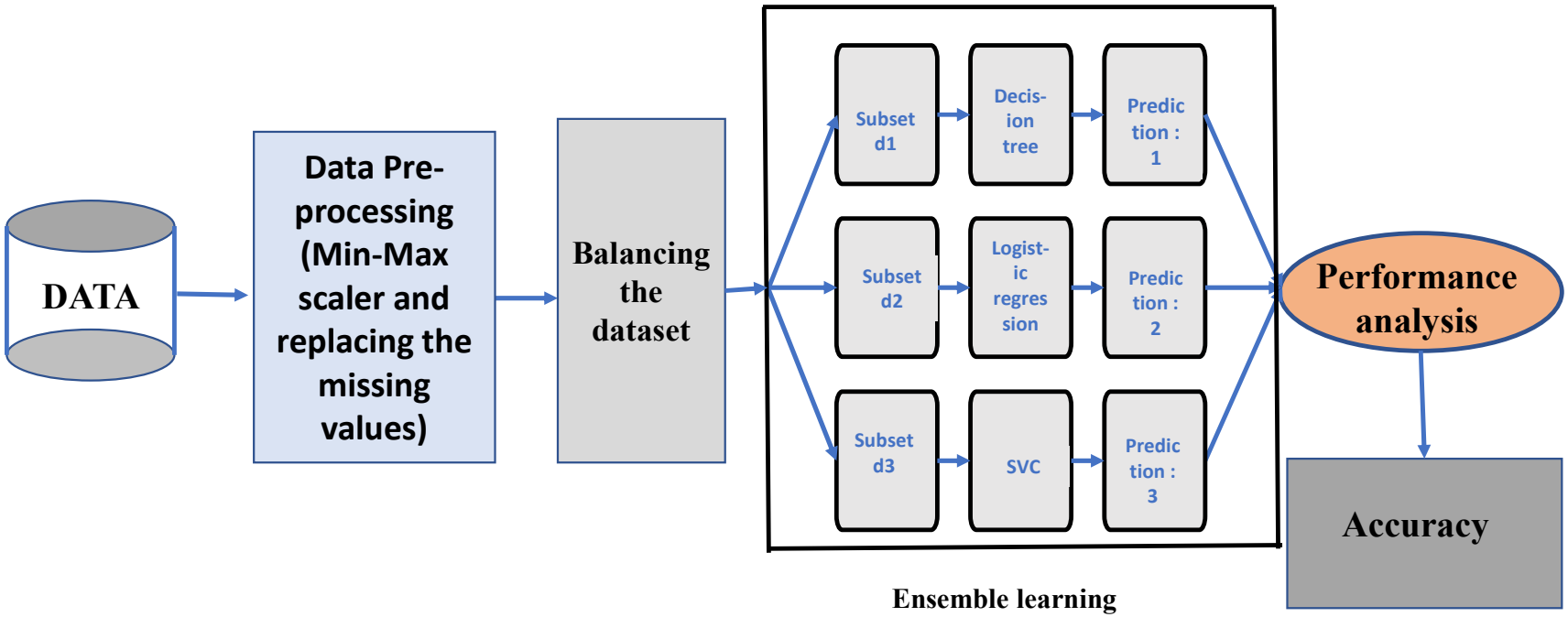


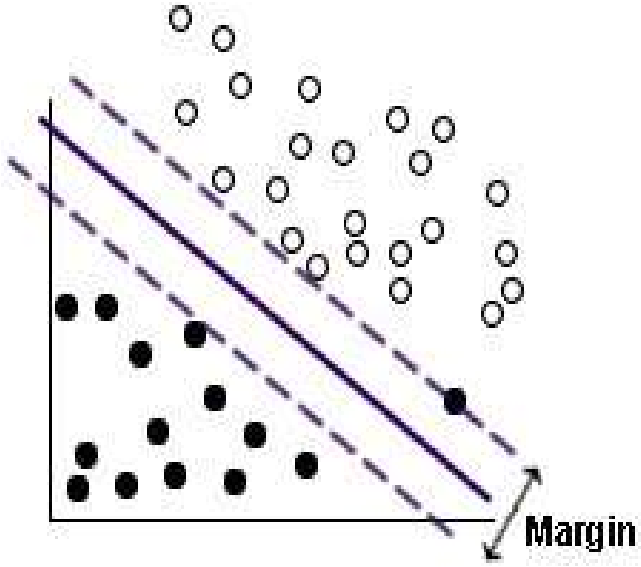
Fig No : 3 **Representation of the work flow of the model**

Why ensemble learning ?

- There are two main reasons to use an ensemble over a single model
  - Performance
  - Robustness
- Performance: An ensemble learning can make better predictions and achieve better performance than any single contributing model.
- Robustness : An ensemble reduces the spread or dispersion of the predictions and model performance.
- It reduces the risk of overfitting. By reducing the variance minimize modelling method bias(it's the incorrect assumption in the model).
- It is of two type:
  - Bagging (it learn from the different model independently in parallel and combine them to determine the model avg and it also reduces the variance).
  - Boosting (in this model learner learn sequentially to increase the model predictions);

SVC

- Support Vector Machine (SVC) is a relatively simple Supervised Machine Learning Algorithm used for classification and regression.
- SVM finds a hyper-plane that creates a boundary between the types of data. In 2-dimensional space, this hyper-plane is nothing but a line.
- In this we use kernel which take data as input and transform it into the required form of processing data.
- The kernel used here is (rbf) radial basis function.
- Transformation in N-dimensional space, where N is the number of features/attributes in the data.



https://www.ibm.com/docs/en/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/images/svm_improved.jpg

Fig No : 4 **Diagram representing SVM classification**

Decision Tree

- The most effective and well-liked technique for categorization and prediction is the decision tree. A decision tree is a type of tree structure that resembles a flowchart, where each internal node represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label.
- Decision trees classify instances by ordering them from the root of the tree to a leaf node, which then indicates the instance's categorization.
- Decision tree draw the hyper rectangles in input space to solve the problem.

Logistic regression

- Logistic regression is basically a supervised classification algorithm.
- Logistic regression estimates the probability of an event occurring.
- The decision for the value of the threshold value is majorly affected by the values of precision(tp/tp+fp) and recall (tp/tp+fn).

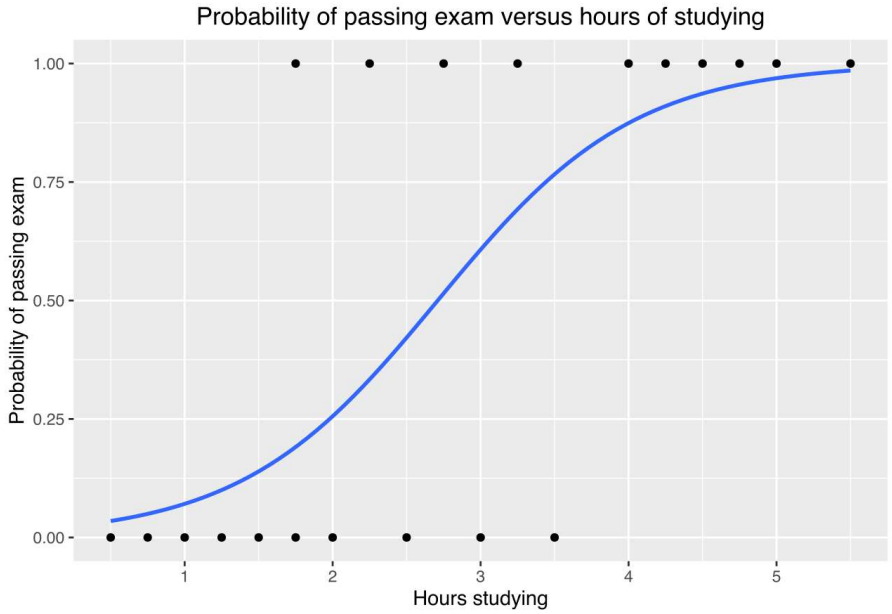


Fig No : 5 **Graph representing the logistic regression**

Accuracy gain

- Using the SVC the accuracy gain on the training data set is 79% and on testing data is 80%.
- Using the Decision tree classifier the accuracy gain on training dataset is 98% and on testing data set is 68%.
- Using the logistic regression the accuracy gain on training dataset is 73% and that's on testing data set is 80%.
- Finally using the voting classifier we get 82% on testing dataset as well as 88% on training dataset.

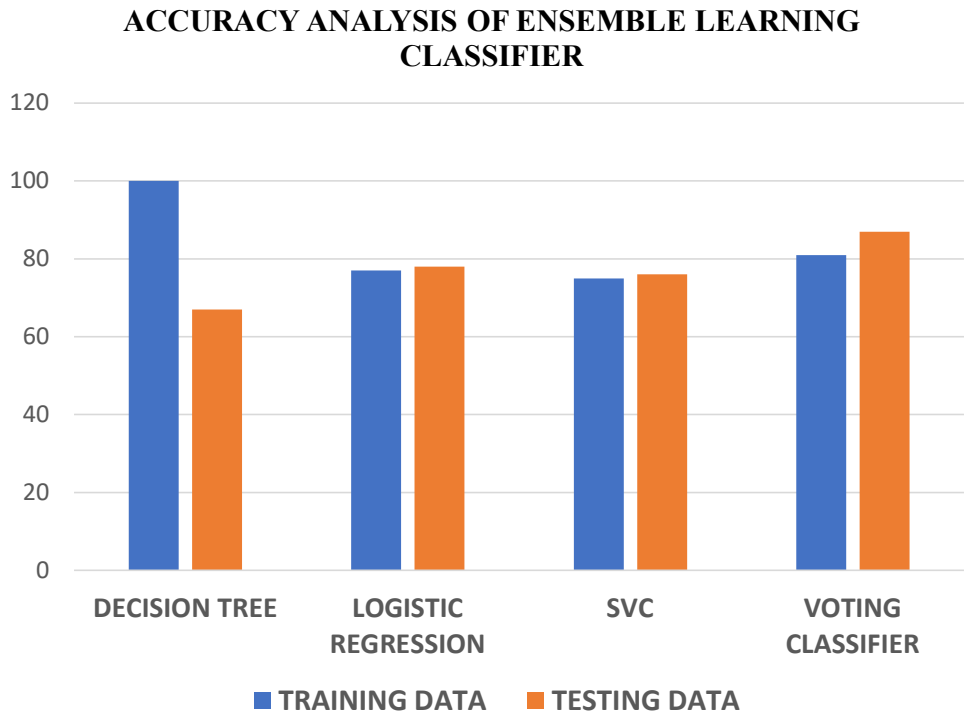


Fig No : 6

Comparative analysis

- Using the KNN and random forest earlier the accuracy gain was 73.82%.
- Using the ensemble learning classification algorithm we get an accuracy of 88 % which is better than using the individual model.

Conclusion

- Balancing the PIMA Indian diabetes dataset plays a very important role in increasing the accuracy.
- Using the ensemble learning approach we get higher accuracy for our diabetes prediction model.

References

[1] Shubham Joshi, Ali Rizwan , Basant Tiwari; A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques; Journal of Healthcare engineering; vol. 2022, page (1-10), 2022.

[2] Neha prerna tigga , shruti garg; Prediction of Type 2 Diabetes using Machine Learning Classification Methods; Jounal of procedia computer science; vol. 167 ,page (706-716),2020.

[3] V. Jackins, S. Vimal, M. Kaliappan ,M.Y. Lee ;AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes; Journal of Supercomputing; vol. 77,page (5198–5219), 2021.

[4] Ionnis kavakiotis, Olga Tsave,Athanasios Salifoglou; Machine Learning and Data Mining Methods in Diabetes Research; computational and structural biotechnology journal;vol. 15 , page (104-116),2020.

[5] Aishwarya Majumdar,Dr. V. vaidehi; Diabetes prediction using machine learning algorithm; International conference on recent trends in advanced computing; vol.165,page (192-199),2019.

Thank You