PREDICTION OF DIABETES USING CLASSIFICATION MODELS BASED ON PIMA INDIANS STUDY

Apoorv Saraogee
Northwestern University, Practical Machine Learning
May 28, 2022

Abstract

Diabetes is a chronic health condition that is a significant contributor to the worldwide healthcare burden. The healthcare burden can be drastically reduced with precise early detection and preventative measures. Predictive analytics (tree-based methods) using medical measurements including glucose level and others offer a useful method to diagnose diabetes early. This study analyzes a dataset of female individuals of Pima Indian heritage near Phoenix, Arizona with a high incidence of diabetes. To predict the diagnosis of diabetes using these measurements, we compare traditional logistic regression classification models with and without interactions and tree-based classification methods such as random forest. The areas under the receiver operating characteristic curve (ROC AUC) is analyzed for each five-fold cross-validated model. The best ROC AUC of 0.84 is found using random forest model which naturally takes interactions into account.

1. Introduction

Diabetes, having excess glucose, is a leading chronic illness that affects up to 470 million people in the world in 2019. It is important for treatment to occur at the appropriate time to prevent deterioration of the organs of the body and therefore essential to diagnose diabetes or a risk of developing diabetes as early as possible (Saxena et al. 2022). As diabetes is a complex medical condition, various factors such as insulin levels, weight and age contribute to the diagnosis. Machine learning classification methods offer a robust way to predict diabetes from the vast amounts of healthcare information data (AlJarullah 2011). This study's central research topic is to analyze a diabetes dataset of 798 women of Pima Indians heritage using classification models with 8 explanatory variables – pregnancies, glucose, blood pressure, skin thickness, insulin, body mass index (BMI), pedigree and age. Diabetes is diagnosed if plasma glucose is at least 200 mg/dl at

any survey examination for the patients (Smith et al. 1988). Research questions include whether tree-structured models work better than a logistic regression model (scored using the correlation coefficient ROC AUC). Other questions will explore interactions between glucose and age, odds ratios of the explanatory variables and models with a binarized glucose variable (high/low).

2. Literature Review

More complicated connectionist models such as neural networks were used in the paper that first described this dataset by filtering for females (Smith et al. 1988). The literature includes a review with many examples of machine learning classification being used on this dataset including comparison studies between different models such as Naïve Bayes, connectionist models such as neural networks, decision tree, random forest and support vector (Larabi-Marie-Sainte et al. 2019). Notable studies compared different models with others such as neural networks (Hasan et al. 2020; Saxena et al. 2022; Chang et al. 2022; Naz and Ahuja 2020; Zou et al. 2018). Solely decision trees were used for classification in the textbook and other literature (Izenman 2008; AlJarullah 2011; Dudkina et al. 2021). This investigation explores random forest networks which were found to have the highest accuracy at 79.8% when compared with connectionist models such as multi-layer perceptions (Saxena et al. 2022). However, this investigation uses a more simple logistic regression model for predictions for comparisons and focused studies on interactions between glucose and age (Christensen 1997).

3. Methods

This research will be conducted on the Pima Indians diabetes dataset in Jupyter notebook using kernels for both R for visualization as well as Python3 for analysis with models in the sklearn

package. Real diabetes data is used to train the classification models and each model is evaluated using their ROC AUC scores. The key objectives include analyzing the odds ratios for the explanatory variables, evaluating the interactions between Glucose and Age and comparing logistic regression models with tree-structured methods of classification such as random forest with relative importance. Glucose will be binarized into a high/low variable for a logistic regression model and calculate an odds ratio. Other variables will also be binarized to calculate odds ratios using contingency tables. The probability of diabetes with age will be compared for high/low glucose groups using a fixed set of explanatory variables using the average values across the dataset to evaluate interactions between these two variables as in (Christensen 1997). A decision tree will be constructed to look for interactions with a max_depth=4 for ease of interpretation and min_samples_leaf=10 as in (Izenman 2008). A logistic regression model with the original set of variables and with an added Glucose*Age interaction variable be compared at five-fold cross validation with a random forest classifier with min_samples_leaf=10.

3. Results

The ROC AUC scores for the five-fold cross-validated models were 0.73, 0.67, 0.75 and 0.84 for the logistic regression with binarized Glucose variables, logistic regression with the original explanatory variables, logistic regression with the original variables and an added Glucose*Age interaction variable and a random forest classification model respectively. The most important variables were Glucose, Insulin and Pregnancies from the odds ratios calculated using the two-way contingency tables with 9.0, 3.8 and 2.7 respectively. This odds ratio of 9.0 was much higher than the odds ratio of 1.6 for Glucose from the logistic regression model with the binarized Glucose variable although both show uneven odds. Glucose, Pedigree and Pregnancies are the

most important in the binarized Glucose logistic regression model as they have the highest model coefficients of 0.48, 0.14 and 0.11 respectively as shown in Table 1. The probabilities are calculated using the logistic regression model for high/low glucose groups with a fixed set of explanatory variables (mean) as shown in Figure 4. We see a steady increase in probability with age and a marked difference between high and low glucose groups with varying slopes by about 15% indicating an interaction. This is further confirmed with the decision tree shown in Figure 5 where Age is shown in the same branch node as Glucose. Using a tree-structured classification method such as random forest yielded Glucose, BMI and Age as having highest feature importance with 0.35, 0.18 and 0.15 respectively as shown in Table 2. From these results, I learned that tree-structured classification methods work better than traditional methods by accounting for interactions. This is clearly seen in our dataset after the interaction variable is added to the logistic regression model with the increase in the AUC score from 0.67 to 0.75.

4. Conclusions

Within this study, the random forest model using the original set of explanatory variables is the best because it had the highest ROC AUC score of 0.84. This is quite good as it is above 0.75-0.8. However, a key disadvantage is the larger difference in the average scores from the test and training sets between the tree-structured models and traditional models (0.1 compared to 0.02) suggesting greater overfitting in the tree-structured models. Glucose, BMI and Age are likely the most important features in the diagnosis of diabetes as shown by the relative importance scores in the random forest model as well as the raw correlation scores. More data is needed in order to obtain a more generalizable model of these measurements to diagnose diabetes in the general population. This should include comparing with other diabetes datasets (Zou et al. 2018)

5. Appendices

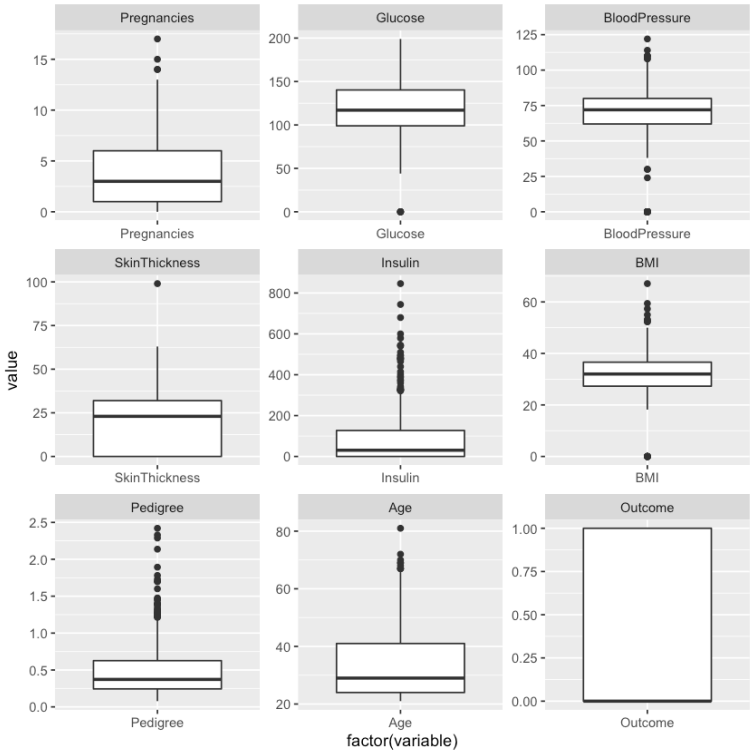

Figure 1. Box plots of explanatory variables of diabetes dataset, using R

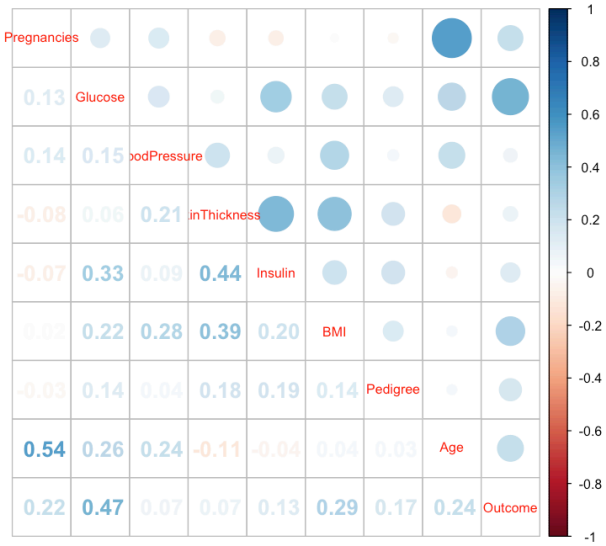

Figure 2. Correlation plot of explanatory variables for diabetes dataset, using R

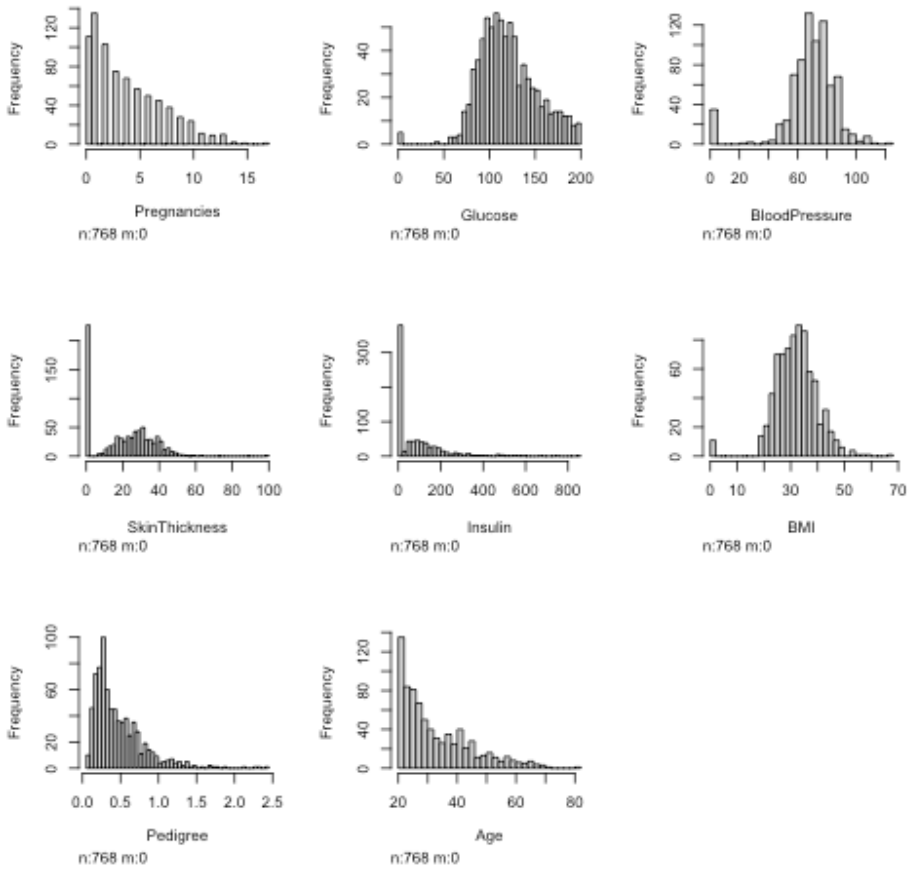

Figure 3. Histograms of explanatory variables, using R

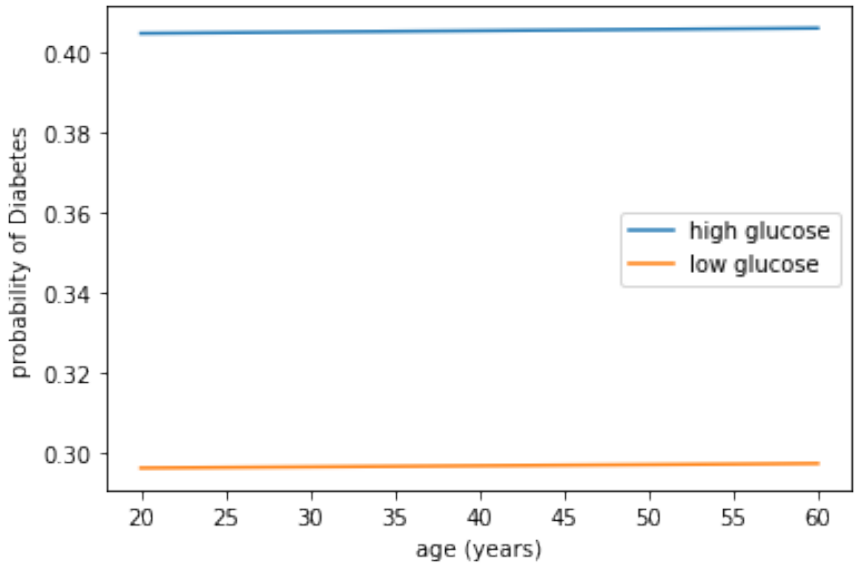

Figure 4. Probability of a diagnosis of diabetes calculated for high and low glucose groups based on logistic regression model prediction coefficients for a fixed set of explanatory variables (mean) compared to age.

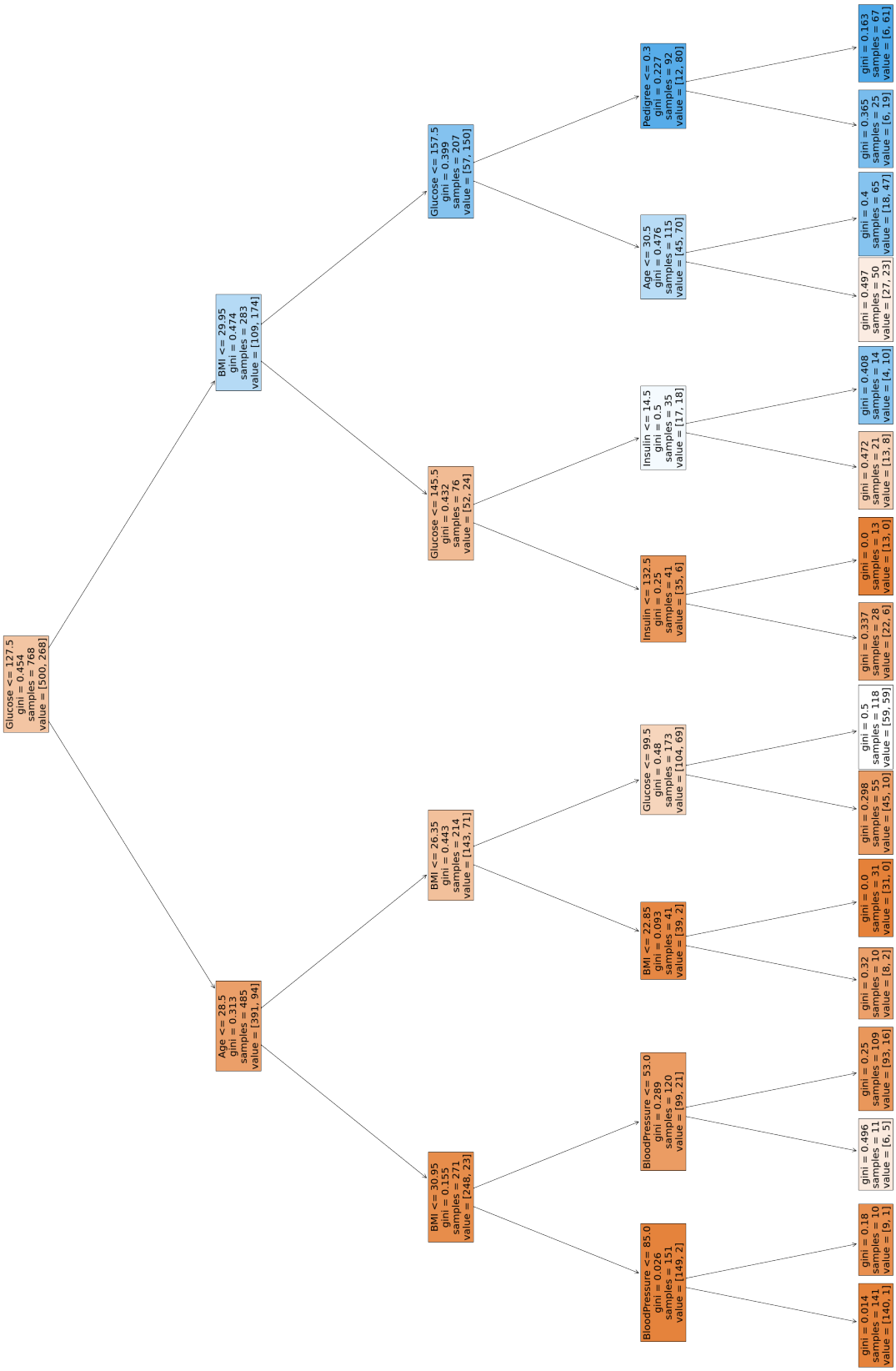

Figure 5. Decision tree with max_depth=4 and min_samples_leaf=10 where Age is seen in the same branch on the left from the parent node of Glucose <127.5 indicating interaction.

Table 1: Logistic regression model coefficients for logistic regression model with binarized Glucose variable

	High/Low Glucose
Pregnancies	0.113083
BloodPressure	-0.024462
SkinThickness	-0.002657
Insulin	0.001627
BMI	0.021940
Pedigree	0.143040
Age	0.000134
Glucose	0.478957

Table 2: Relative feature importance scores calculated for the random forest model in Python3

	Random Forest
Pregnancies	0.083568
Glucose	0.351063
BloodPressure	0.035996
SkinThickness	0.041866
Insulin	0.065490
BMI	0.183520
Pedigree	0.090326
Age	0.148171

Supporting Files

- Assignment3.html
- Assignment3.py
- Assignment3.ipynb
- pima.csv

References

AlJarullah, Asma A. 2011. “Decision Tree Discovery for the Diagnosis of Type II Diabetes.” *2011 International Conference on Innovations in Information Technology, IIT 2011*, 303–7. https://doi.org/10.1109/INNOVATIONS.2011.5893838.

Chang, Victor, Jozeene Bailey, Qianwen Ariel Xu, and Zhili Sun. 2022. “Pima Indians Diabetes Mellitus Classification Based on Machine Learning (ML) Algorithms.” *Neural Computing & Applications*, March, 1. https://doi.org/10.1007/S00521-022-07049-Z.

Christensen, Ronald. 1997. *Log-Linear Models and Logistic Regression*. 2nd ed. Springer.

Dudkina, Tetiana, Ievgen Meniailov, Kseniia Bazilevych, Serhii Krivtsov, and Anton Tkachenko. 2021. “Classification and Prediction of Diabetes Disease Using Decision Tree Method.” *IT&AS*. http://ceur-ws.org/Vol-2824/preface.pdf.

Hasan, Md Kamrul, Md Ashraful Alam, Dola Das, Eklas Hossain, and Mahmudul Hasan. 2020. “Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers.” *IEEE Access* 8: 76516–31. https://doi.org/10.1109/ACCESS.2020.2989857.

Izenman, Alan J. 2008. “Modern Multivariate Statistical Techniques,” Springer Texts in Statistics, . https://doi.org/10.1007/978-0-387-78189-1.

Larabi-Marie-Sainte, Souad, Linah Aburahmah, Rana Almohaini, and Tanzila Saba. 2019. “Current Techniques for Diabetes Prediction: Review and Case Study.” *Applied Sciences (Switzerland)* 9 (21). https://doi.org/10.3390/APP9214604.

Naz, Huma, and Sachin Ahuja. 2020. “Deep Learning Approach for Diabetes Prediction Using PIMA Indian Dataset.” *Journal of Diabetes and Metabolic Disorders* 19 (1): 391. https://doi.org/10.1007/S40200-020-00520-5.

Saxena, Roshi, Sanjay Kumar Sharma, Manali Gupta, and G. C. Sampada. 2022. “A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods.” Edited by Arpit Bhardwaj. *Computational Intelligence and Neuroscience* 2022 (April): 1–11. https://doi.org/10.1155/2022/3820360.

Smith, Jack W, J E Everhart, W C Dicksont, W C Knowler, and R S Johannes. 1988. “Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus.” https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/pdf/procascamc00018-0276.pdf.

Zou, Quan, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang. 2018. “Predicting Diabetes Mellitus With Machine Learning Techniques.” *Frontiers in Genetics* 9 (November): 515. https://doi.org/10.3389/FGENE.2018.00515/BIBTEX.