Predicting Diabetes Using Machine Learning Models

Govind Buttar

8/22/22

Abstract

According to the CDC, in 2019, 37.1 million people aged 18 or older were diagnosed with diabetes, with 8.5 million not aware or did not report that they met the laboratory criteria for diabetes[1]. Developing a tool that is widely available for people to access can cut into these numbers, as it would result in more interest in seeking preventative care and establishing diagnoses.

My research question concerns how AI can improve diabetes detection by giving a prediction of the risk that a patient is in for developing diabetes. This encourages people to check with their doctors and seek preventative care. Most people are concerned with getting preventative care, but are unwilling to consult their doctor because of the costs with care for a condition they may never develop. Several models such as LR, SVC, and RF were experimented with to find the model with the lowest false negative return, or the highest recall value. This will make the model more reliable and will allow people to use it as a tool and discuss its predictions with their doctors. Our final model achieved 0.67 recall for people at high risk for developing diabetes, and 0.88 for people at low risk for developing diabetes. The model achieved a high recall for people at low risk for being diabetic, but achieved poor recall for people at high risk. This model can be used as a screening tool available to everyone, and get people more interested in preventative care if the model predicts they are at high risk for developing diabetes.

1. Introduction

Machine Learning is a subset of Artificial Intelligence, usually utilized as a way to solve problems from a variety of fields. Machine Learning is ideal for this problem since there are already thousands of electronic records that can be deidentified and used to train machine learning models. The trained model can then be used to automate diagnoses using the data.
By automating diagnoses, we can allow patients to understand the risk of developing these diseases. While people are concerned about preventative care for problems like cancer, they are hesitant to approach their doctor because of the cost for their care. They are concerned about losing money but are not aware that the care will save them from developing these types of problems. A solution that will give people an estimate if they are at high risk can save many lives.

Our work concerns developing a supervised classification model that receives numerical input from the patient to allow them to analyze their risk of developing diabetes, which is the output of our model. This would encourage them to see their doctor and get them interested in preventative care.

2. Background

The concept of predictive models in healthcare is not a novel concept, as many others have attempted to determine how to derive efficient models for the industry. In the research article *Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis* by Gopi Batteneni, models such as RF and NN can detect CD's with higher detection at different diagnoses[2]. However, the model also highlighted the troubles with working on a high dimensional dataset, such as accuracy problems. It also highlighted how much of this advanced technology can be used to assist doctors and nurses, but is not readily available to patients as a pre-screening technique and get them more invested in preventative care.

3. Dataset

Dataset Selection and Characteristics

We obtained our dataset from Kaggle.com [3]. Our dataset contains 768 patients and eight variables (Number of Pregnancies, Plasma Glucose Concentration, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age) to consider. All the patients in this dataset are female and over the age of 21.

Outcome Variable

Our outcome variable is whether the patient has been diagnosed with diabetes or not. The patient was diagnosed with diabetes if they had a glucose concentration above 300 mg/dl two hours after ingesting 75 grams of a carbohydrate solution.

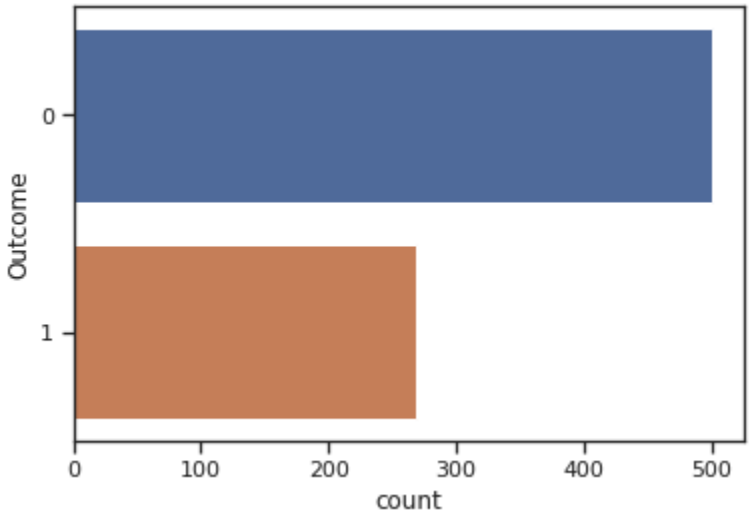

Figure 1: Outcome Countplot

Data Feature Extraction
Initially, PCA was used to obtain the most important features to include in our processed dataset. However, it was determined that the model would lose important information if features were excluded from the dataset.

The dataset was then split into training and testing datasets for our model. The training testing split was 80:20, meaning that the training dataset has 80% of the total data

4. Methodology / Models

Preprocessing our dataset
Since our dataset did not contain any null values, it was not necessary to fill any in. It was also not necessary to drop any of our model inputs since valuable information would be lost since our dataset only contains eight variables.

Model Development
Multiple models (LR, Decision Tree, Ridge Classification, RF, SVC) were experimented with to narrow down the two models that would provide the highest recall. Each model was trained on its default hyperparameters (SVC C= 1, gamma = scale, RF max_features = sqrt, n_estimators = 100, etc.), the top two models were chosen for hyperparameter tuning due to their higher performance. After printing out the recall values of all the models, the two models with the highest recall were chosen for Hyperparameter Tuning. However, the Random Forest model had a considerably higher recall value, so it was decided that the Random Forest model would be used for our final model.

Hyperparameter Tuning
Building off our initial Random Forest model, multiple different parameters were computed using sklearn's GridSearchCV function to find the best combination of the hyperparameters that were inputted.

Evaluating Feature Importance
By analyzing our model, the features in order of highest importance are: Glucose, BMI, Age, DiabetesPedigreeFunction, Pregnancies, BloodPressure, Insulin, and SkinThickness. The importance of these features were measured and output using SHAP.

5. Results and Discussion

Table 1: Models' Performance

Model	Accuracy	Precision	Recall	F1-score
LR	0.78	0.76	0.56	0.65
Decision Tree	0.69	0.58	0.53	0.55
Ridge Classifier	0.78	0.76	0.56	0.65
RF	0.81	0.76	0.67	0.71
SVC	0.81	0.82	0.58	0.68

Looking at Table 1, the RF model is the highest performing model in terms of both f1_score and recall, the two scores that are most helpful for preventing false negatives. After performing hyperparameter tuning on this model, we can compare the two to see which one is more effective.

Results of Tuned Model vs Untuned Model

Table 2: Untuned RF vs Tuned RF

Model	Accuracy	Precision	Recall	F1-score
Tuned RF	0.82	0.8	0.65	0.72
Untuned RF	0.8	0.76	0.67	0.71

Our tuned model, when trained, ended up with a worse recall and fl_score compared to its untuned counterpart.

Model Reasonability

Using SHAP, we can observe what features are most important to our model and check to see if its decisions have a basis given current knowledge.

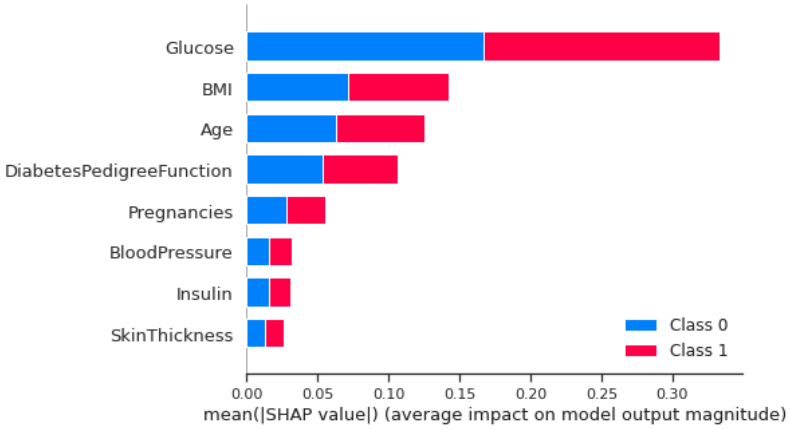

Figure 2: Feature importance of highest performance model

By observing Figure 3, we can conclude that the model's most important factor does make sense in the context of what the research paper defined what diabetes meant to them. A glucose concentration higher than 200 mg/dl two hours after ingesting a 75gm of a carbohydrate solution was classified as diabetic. Most of the other factors make sense because overweight and elderly people are more prone to being diabetic. The importance of insulin also makes sense in this context because our dataset specifies that the research was specifically for non-insulin dependent diabetes mellitus.

Limitations

One thing to take into consideration is the amount of hyperparameters that were chosen. For our research, we focused on two hyperparameters. If we had chosen more to work with, our model would be more tuned and would be performing way better. There was little minimizing of data imbalances, which would help correct the model from inherently skewing towards diagnosing a client with no diabetes or diabetes. With little time to create and train the model, there was little focus on exploration into other models, such as Deep Learning models, or expanding the size of the dataset.

6. Conclusions

This study explored the solutions to developing Machine Learning tools available for the public to become more concerned and interested in preventative care. We developed Logistic Regression, Tree, Ridge Classifiers, Support Vector, Random Forest, models and found that the Random Forest model performed the best. Starting to make tools for the public to use may be considered to be a driving factor to misuse and fatality, but teaching the public how to properly use these tools will reduce the magnitude of these problems. In the future, obtaining more data to explore is crucial to developing a better model as well. In the future, a focus on exploring other models and the usage of more hyperparameters who drastically increase model performance and reliability. This model can be implemented as a tool for doctors to easily perform preliminary screenings as a quick check-in tool for patients. This model can also be used by clients who are interested in their long term health, in which case they can bring the model's diagnosis with them to their doctor, which would allow more people to invest in their health and avoid the negative impacts of health neglect.

Acknowledgments

A thank you to Inspirit AI and my mentor Jacklyn Luu. Without their guidance and help, this paper would not have been possible.

References

1. "Prevalence of Both Diagnosed and Undiagnosed Diabetes." *Centers for Disease Control and Prevention*, 29 Dec. 2021, www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html. Accessed 20 Aug. 2022.

2. Battineni, G.; Sagaro, G.G.; Chinatalapudi, N.; Amenta, F. Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis. *J. Pers. Med.* **2020**, *10*, 21. https://doi.org/10.3390/jpm10020021

3. https://www.kaggle.com/datasets/shantanudhakadd/diabetes-dataset-for-beginners

4. "Artificial Intelligence in Health Care." *The Washington Post*, 13 Jan. 2021, www.washingtonpost.com/washington-post-live/2021/01/13/artificial-intelligence-health-care/. Accessed 31 July 2022.

5. Oslon, Parmy. "Too Much AI May Not Be Good for Your Health or the NHS."*TheWashington Post*, 12 May 2022, www.washingtonpost.com/business/too-much-ai-may-not-be-good-for-your-health-or-the-nhs/2022/05/11/af033e64-d186-11ec-886b-df76183d233f_story.html. Accessed 31 July 2022.

6. Chen, RC., Dewi, C., Huang, SW. *et al.* Selecting critical features for data classification based on machine learning methods. *J Big Data* 7, 52 (2020). https://doi.org/10.1186/s40537-020-00327-4