Predicting Cases of Diabetes Using Machine Learning Algorithms

Joshua Qian, Tarun Hoskere

Emory University

Abstract

In this study, we used several machine learning models to predict whether or not a patient had diabetes based on several patient features. The models we used include KNN, Decision Tree, Random Forest, and Naïve Bayes classifiers, which were trained on a publicly available dataset of over 100,000 patients. The data was preprocessed by standardizing the numerical features and one hot encoding the categorical features. We then implemented feature selection by identifying features that had little correlation with the target feature (diabetes) and removed them. The models were then evaluated using several metrics: accuracy, precision, recall, F1 score, and AUC (area under curve). After identifying the random forest model as the best-performing model, we further tuned its hyperparameters and used it for our final classification task. While the dataset was imbalanced, with only 9% of patients having diabetes, the final model was able to correctly predict the presence of diabetes 97% of the time.

Introduction

The prevalence of diabetes has been steadily increasing across the world. With over 400 million cases worldwide and over 38 million cases in the United States, diabetes has a firm place as one of the world's largest health problems. It is closely linked to heart and kidney disease, amongst many other health complications. Further, the increase in consumption of sugar and other processed foods has only helped accelerate the rise in diabetes cases. Thus, the need for accurate and efficient diagnostic tools is higher than ever.

There has long been speculation as to whether or not certain factors have been linked to diabetes cases. Through our study, we aim to leverage machine learning algorithms to identify the most prevalent features with respect to diabetes and create a model to accurately classify diabetes within individuals. By exploring the impact of various features such as gender, age, body mass index, and several other health indicators, we aim to contribute to the development of more accurate diagnostic tools for diabetes.

Related Works

There have been previous works that have aimed to use machine learning techniques in the field of diabetes. The majority of work has been done on a different dataset of diabetes patients who are primarily of Pima Indian descent. That dataset is much smaller, containing only 768 samples and a slightly different set of features, but has roughly the same level of imbalance. In a study from 2023, Iparraguirre-Villanueva, Castaneda, Espinola-Linares, and Cabanillas-Carbonell performed a classification task on the Pima Indian dataset using KNN, Naive Bayes, and Decision Tree models. The models trained on that dataset yielded considerably lower scores than our models. The dataset used in our study contains many more samples and includes people of all different backgrounds which is where it differs from many previous works. We plan to build on theirs and other previous research by utilizing a larger, more inclusive dataset, comparing several models against each other, and implementing more efficient preprocessing techniques.

Methodology

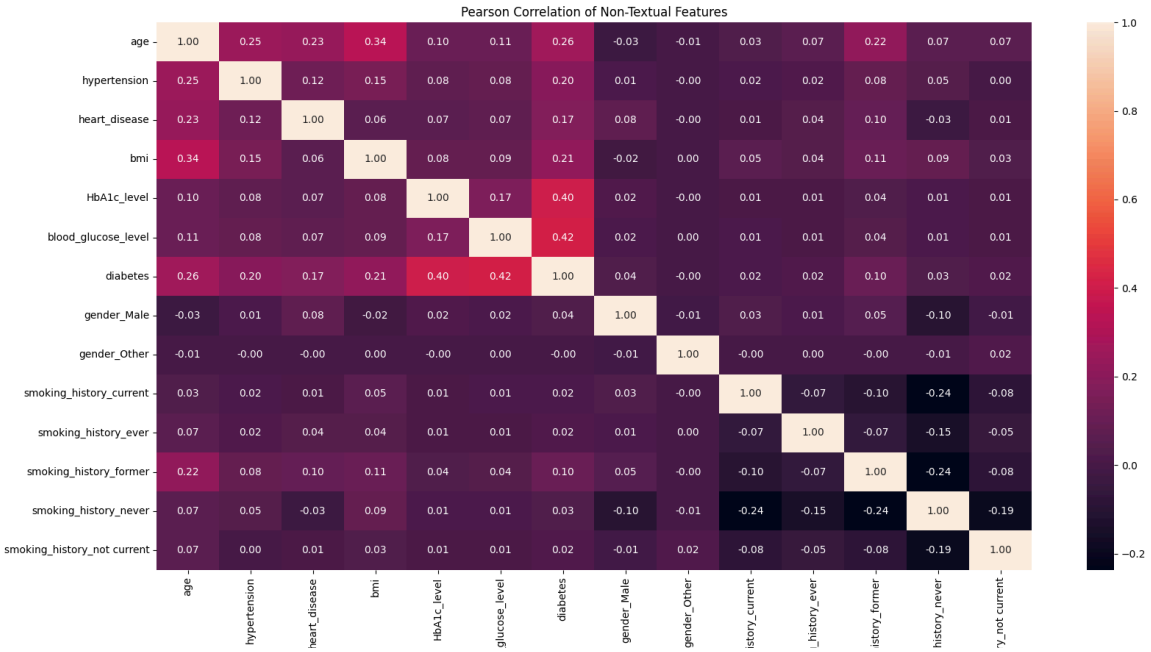
Step 1: Data Selection and Preprocessing

Our original dataset contains over 100,000 samples of patients and 9 features. These features included gender, age, body mass index, hypertension, heart disease, smoking history, HbA1c level, blood glucose level, and the target feature: diabetes. There were 5 categorical features: gender, hypertension, heart disease, smoking history, and diabetes. There were also 4 numerical features: age, body mass index, HbA1c level, and blood glucose level. To improve the performance of our models, we implemented several data preprocessing techniques:

1. One-hot encoded categorical features so the models can process them effectively
    - Gender: binary feature with values ‘Male’ or ‘Female’
    - Hypertension: binary feature with values 1, signaling the presence of hypertension or 0, signaling no hypertension
    - Heart Disease: binary feature with values 1, signaling the presence of heart disease or 0, signaling no heart disease
    - Smoking History: values ‘never’, ‘not current’, ‘current’, and ‘No Info’
    - Diabetes: binary feature with values 1, signaling the presence of diabetes or 0, signaling no diabetes
2. Standardized the numerical features. This is done to normalize the scale of the data, which can improve the convergence of some machine-learning models. The scaling was done using the formula:

$$z = \frac{x - \mu}{\sigma}$$

3. Feature Selection: We created a Pearson Correlation heatmap to examine the linear relationship between variables. We then removed the variables with little correlation to the target variable. There were no two variables with high correlation with each other, so there was no need to remove any variables to reduce redundancy.



We removed smoking history and gender as they had values < 0.1, meaning they had little linear correlation with diabetes.

4. Train/test split: we chose a train/test split of 80/20 to ensure a large enough training set and ample testing data.

Step 2: Model Selection

K-Nearest Neighbors: a classifier model that predicts the output class based on the k most similar data points in the training set. It is a very simple model but can be sensitive to outliers or reduce efficiency as datasets increase in size.

Naive Bayes: a probabilistic classifier model that assumes independence between features. It predicts the output class based on the Bayes Theorem formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

It is a very efficient model, which makes it good for large datasets. However, it assumes feature independence, which may limit its performance.

Decision Tree: a tree-structured model where nodes represent features, branches represent decision rules, and leaves represent outcome predictions. These models can handle non-linear relationships that other models may not be able to.

Random Forest: ensemble method that combines several decision tree models to create a more generalized and robust model.

Step 3: Hyperparameter Tuning

For each model, we created a parameter grid and used k-fold cross-validation with 5 folds to test the performance of each model using different parameters against each other. Accuracy was used as the scoring metric in these comparisons.

KNN Parameter Grid:

n_neighbors	3	5	7	9
weights	Uniform	Distance		
metric	Euclidean	Manhattan		

The best parameters for the KNN model were found to be n_neighbors = 9, weights = Uniform, metric = Manhattan

Decision Tree Parameter Grid:

max_depth	None	10	20	30
min_samples_split	2	5	10	
criterion	Gini	Entropy		

The best parameters for the Decision Tree model were found to be max_depth = 10, min_samples_split = 2, criterion = Entropy

Random Forest Parameter Grid:

n_estimators	50	100	200	
max_depth	None	10	20	30

min_samples_split	2	5	10	

The best parameters for the Random Forest model were found to be n_estimators = 50, max_depth = 10, min_samples_split = 10

Naive Bayes Parameter Grid:

var_smoothing	1e-09	1e-08	1e-07

The best parameters for the Naïve Bayes model were found to be var_smoothing = 1e-07

Empirical Results

After preprocessing the data and selecting the best hyperparameters for each model, we tested each model's ability to predict the correct class of diabetes (0 or 1) and recorded the accuracy, precision, recall, F1 score, and AUC to compare them.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

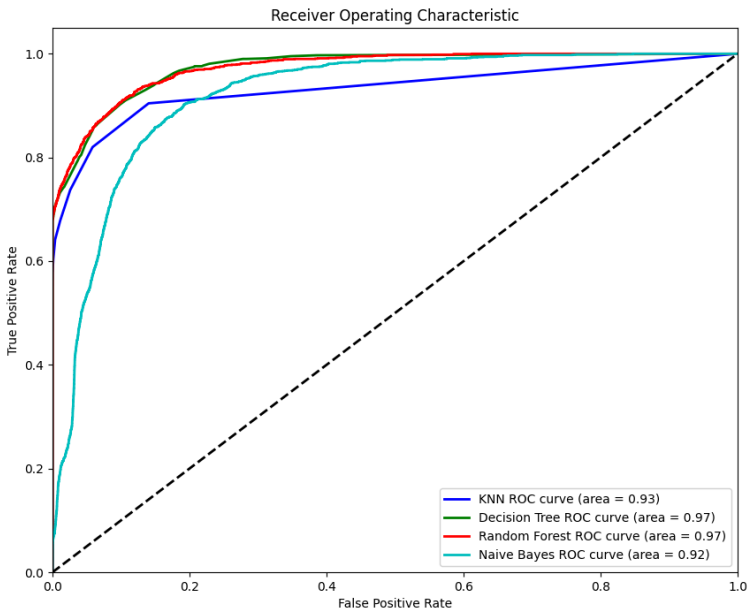
Metrics For Each Model (using best parameters):

	KNN	Decision Tree	Random Forest	Naive Bayes

Accuracy	0.9659	0.9722	0.9727	0.8648
Precision	0.9652	0.9727	0.9734	0.9300
Recall	0.9659	0.9722	0.9727	0.8468
F1 Score	0.9630	0.9699	0.9702	0.8740
AUC	0.9321	0.9729	0.9724	0.9197

*The imbalanced data may influence the high scores

We were left with the following AUC curves:



From these results, we can determine that the Random Forest model performed best. It achieved the highest accuracy, precision, recall, and F1 score. The Decision Tree model just beat the

Random Forest in AUC. The Naïve Bayes model was the worst performing. This could have been foreseen as it makes the assumption of independence between features, which was not the case here.

Discussion

All of our models performed exceptionally well, resulting in accuracy scores > 0.96, except for Naïve Bayes, which had an accuracy of 0.86. As mentioned earlier, the data was imbalanced, which contributed to the high accuracy scores, but the models we chose were also very effective in classification tasks on datasets like the one we used.

The results we obtained align with the characteristics of each model. Naive Bayes did not perform as well due to its assumption of independence between features. KNN, Decision Tree, and Random Forests performed very similarly, but the Random Forests achieved better results across the board. We attributed this to the robustness of the Random Forest model, which we identified as one of the most important features of large classification tasks. We also evaluated feature importance and found that bmi and blood glucose levels were the most important features, which is consistent with current research on diabetes.

While we did recognize the imbalance in our dataset, we did not explicitly do anything to manage it. This may have affected the results, as random forests are better for handling imbalanced datasets, as they reduce variance and avoid overfitting. If we were to expand on this study, it would be interesting to see how certain techniques, such as oversampling the minority class, might affect our results. We could also try other techniques, such as adjusting the weights of different classes so the models would pay more attention to the minority class.

Contributions

Joshua Qian: Conducted research on the diabetes dataset and other previous works. He laid out and implemented the preprocessing steps. He selected the testing metrics and assisted in writing the code for the model evaluations. Lastly, he was responsible for writing the research report.

Tarun Hoskere: Selected the models to be used in this study and set up parameter grids for each model. He planned and coded the model evaluations. He also prepared the graphics to display

model performances. Lastly, he assisted in writing the research report, proofreading the report, and contributing significantly to the hyperparameter tuning section.

GitHub Repository: https://github.com/joshuaqian11/Diabetes-ML-Classifier

Works Cited

Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes. 15 July 2023, https://doi.org/10.3390%2Fdiagnostics13142383.

Black, Jason, et al. "An Introduction to Machine Learning for Classification and Prediction." *Oxford Academic*, https://doi.org/10.1093/fampra/cmac104.

"Diabetes Statistics." *Diabetes Statistics*, National Institute of Diabetes and Digestive and Kidney Diseases, www.niddk.nih.gov/health-information/health-statistics/diabetes-statistics#:~:text=Estimated%20prevalence%20of%20diabetes%20in,8.9%25%20of%20the%20population).

Mustafa, Mohammed. "Diabetes Prediction Dataset." *Kaggle*, www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data.

Tan, Haoyuan. "Machine Learning Algorithms for Classification." *IOP Science*, iopscience.iop.org/article/10.1088/1742-6596/1994/1/012016/meta.