Diabetes Prediction Using Machine Learning

Professor Bingin Zhao
Stat 4770
Python For Data Science
April 12, 2024

Nikhita Pabbaraju

ABSTRACT

This study investigates the predictive performance of three machine learning models for diagnosing diabetes using the Pima Indians Diabetes dataset. With diabetes becoming increasingly prevalent globally, early and accurate diagnosis is essential for effective management and prevention of associated complications. Leveraging exploratory data analysis (EDA) techniques and machine learning algorithms including logistic regression, random forest, and k-nearest neighbors, this study aims to identify significant predictors of diabetes onset and evaluate model performance using comprehensive metrics such as accuracy, precision, sensitivity, specificity, F1 score, and area under the ROC curve (AUC). The findings reveal strong correlations between certain variables like glucose and BMI with the diabetes outcome, underscoring their importance in predictive modeling. Among the models tested, the random forest classifier demonstrates promising results in terms of accuracy and discriminatory power, with an accuracy of 73.59%. Future research could explore feature engineering methods and advanced modeling techniques to further enhance prediction accuracy and expand the application of machine learning in personalized diabetes management and the prediction of other metabolic disorders. This study contributes to the broader efforts of public health initiatives aimed at combating the diabetes epidemic and improving patient outcomes.

INTRODUCTION

Understanding and accurately diagnosing diabetes is critically important due to the increasing prevalence of the disease worldwide. Diabetes, particularly Type 2, is a major health concern that affects millions globally. Early and precise diagnosis is essential for effective management and prevention. Consequently, this study aims to enhance our understanding of the predictive factors associated with diabetes onset but also contributes to the broader efforts of public health initiatives aimed at predicting and curbing the diabetes epidemic.

The Pima Indian Diabetes Dataset, from the National Institute of Diabetes and Digestive and Kidney Diseases, contains information on 768 Pima Indian women over 21 years old from Phoenix, Arizona. In Pima Indians, the prevalence of Type 2 diabetes is over 10%, making diabetes a significant concern in health for this demographic. This dataset contains 8 diabetes predictor variables which collectively contribute to the binary target variable, which indicates the presence or absence of diabetes.

The objective of analyzing the dataset is to determine the presence of diabetes in patients based on specific measurements in the data. This project uses exploratory data analysis (EDA) and machine learning (ML) techniques to find correlations and predict diabetes onset. Specifically, logistic regression, random forest, and k-nearest neighbors (KNN) classifiers are utilized. The performance of these models is assessed using a comprehensive suite of metrics, including accuracy, precision, sensitivity, specificity, F-score, and area under the ROC curve (AUC).

DATA

The PIMA Indians Diabetes Database is publicly available on Kaggle and has been widely utilized in diabetes-related research. It includes medical measurements of 768 female Pima Indian individuals, a group that has a notably high prevalence of Type 2 diabetes. This dataset is structured with eight independent variables and one binary target variable indicating the presence or absence of diabetes.

- Number of Pregnancies
- Glucose concentration
- Diastolic blood pressure (in mmHg)
- Skin thickness (triceps skin fold thickness in mm):
- BMI (in kg/m2)
- Age
- Diabetes Pedigree Function (function that predicts diagnosis from ancestral history)
- Insulin (in mu U/ml).
- Outcome: Expresses the final result of diagnosis (1 corresponds to being diagnosed and 0 being not diagnosed)

The dataset required several preprocessing steps before analysis could begin. Initially, an examination for missing or null values was conducted. Certain variables, like "Glucose", "Blood Pressure", "Skinfold Thickness", "Insulin", and "BMI", had zeros in places where they were biologically implausible. These zeros were treated as missing values. Missing values were imputed using the mean of the non-missing data.

The dataset was split into training and testing sets to validate the performance of the predictive models. 70% of the data was used for training the models, while the remaining 30% served as the test set to evaluate model accuracy and other performance metrics.

PROBLEMS AND METHODS

This project addresses the overarching problem of predicting diabetes onset using clinical health data. The key problems explored include:

1. Understanding Risk Factors: Determining relationships between clinical variables that are significant predictors of diabetes.
2. Model Development and Selection: Developing accurate and robust machine learning models that can predict the onset of diabetes based on risk factors.
3. Model Evaluation and Comparison: Assessing and comparing the performance of different models to determine the most effective approach.

To address these problems, the following methods were used:

1. EDA: Initial data analysis to understand the distribution of variables and identify potential correlations using histograms and correlation matrices.
2. Machine Learning Models: Three primary models were selected for this study:
    a. Logistic Regression: A statistical model that estimates the probability of a binary outcome.

   b. Random Forest Classifier: A method that makes predictions by combining the results of a series of regression decision trees
   c. K-Nearest Neighbors (KNN): A method that classifies cases based on their similarity to other instances.
3. Model Evaluation: Models were evaluated using accuracy, precision, sensitivity, specificity, F-score, and the area under the ROC curve (AUC). These metrics provide a comprehensive view of model performance, considering both the ability to predict diabetes correctly and the rate of false positives and negatives.

These methods are appropriate as they provide a thorough understanding of the dataset and support the development of predictive models evaluated by comprehensive metrics. Logistic Regression offers a baseline for performance comparison, while Random Forest and KNN allow exploring more complex relationships in the data.

Alternative approaches considered included using more advanced machine learning models such as Support Vector Machines (SVM) and neural networks. However, due to the relatively small size of the dataset and the complexity of these models, simpler models were chosen to avoid overfitting and ensure model interpretability.

RESULTS AND DISCUSSION

EDA

In the EDA phase of this project, this dataset was first examined through the analysis of a correlation matrix. By visualizing the correlation matrix using a heatmap, I gained insights into which features were most strongly correlated with the outcome variable, identifying potential predictors of diabetes. The last row shows that glucose levels and BMI have the highest two correlations with the diabetes outcome than the other variables. Additionally, notable correlations were observed between skin thickness and BMI, glucose and insulin, as well as pregnancies and age, highlighting interactions that could influence the modeling approach.

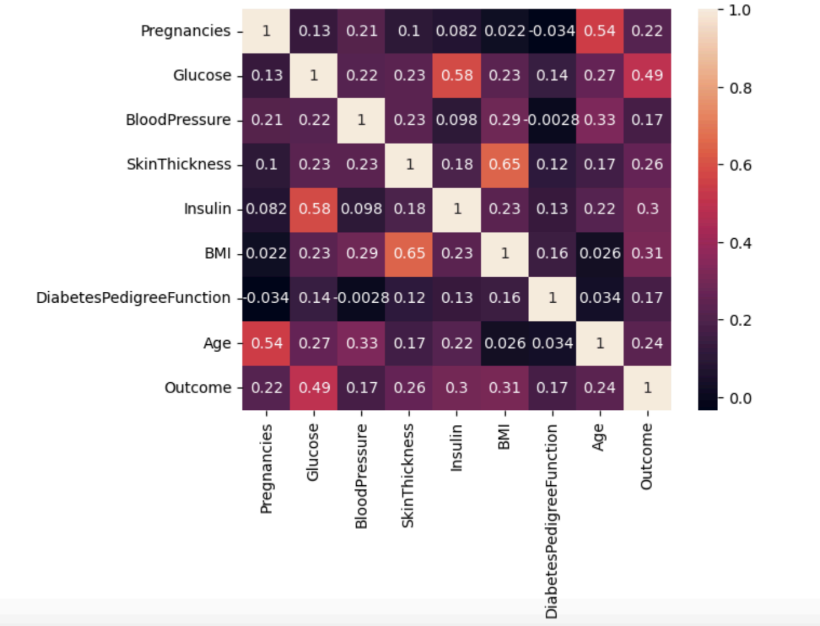
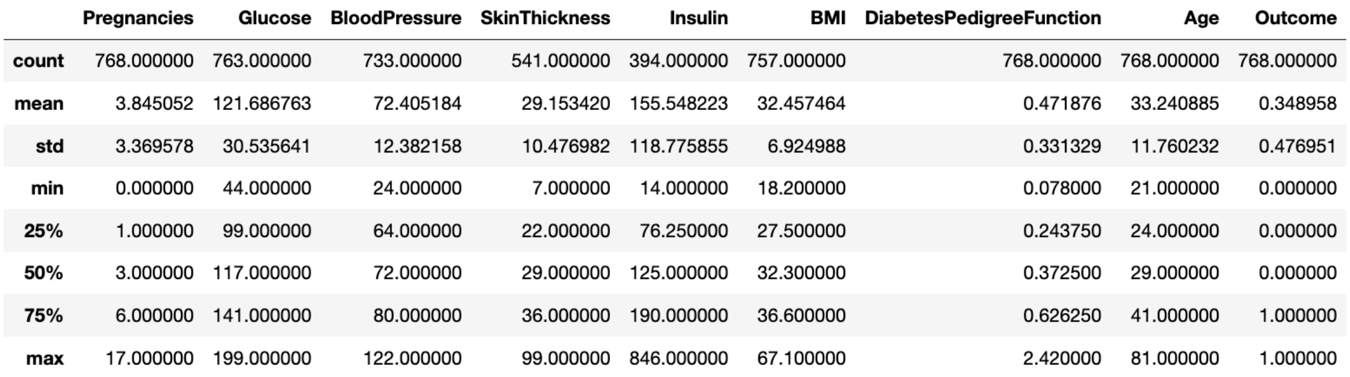


Figure 1. Heatmap of variables



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	763.000000	733.000000	541.000000	394.000000	757.000000	768.000000	768.000000	768.000000
mean	3.845052	121.686763	72.405184	29.153420	155.548223	32.457464	0.471876	33.240885	0.348958
std	3.369578	30.535641	12.382158	10.476982	118.775855	6.924988	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	64.000000	22.000000	76.250000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	29.000000	125.000000	32.300000	0.372500	29.000000	0.000000
75%	6.000000	141.000000	80.000000	36.000000	190.000000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Table 1. Descriptive statistics of the 9 variables.

Additionally, I wanted to look at the distribution of the outcome variables to note whether the outcome was balanced or imbalanced. I found that the outcome was imbalanced, with 500 with no diabetes diagnosis and 268 with a diabetes diagnosis.

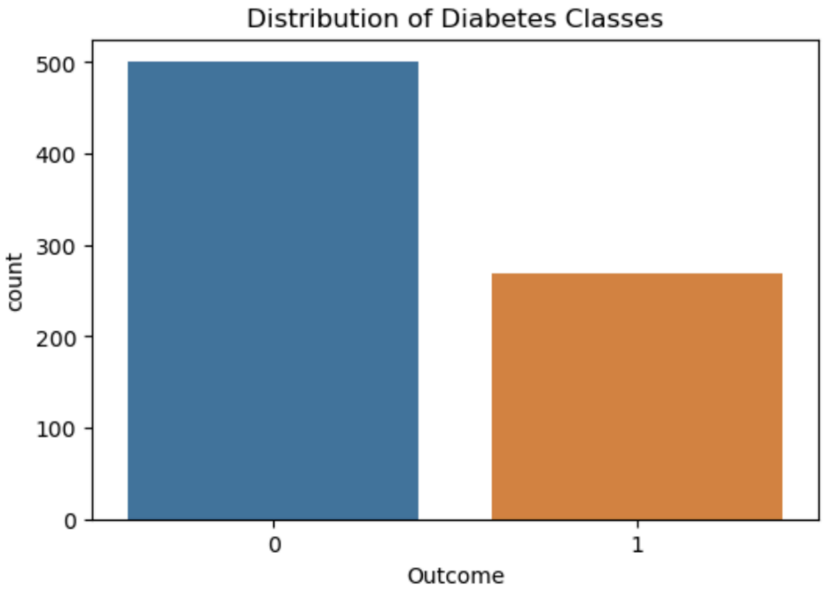
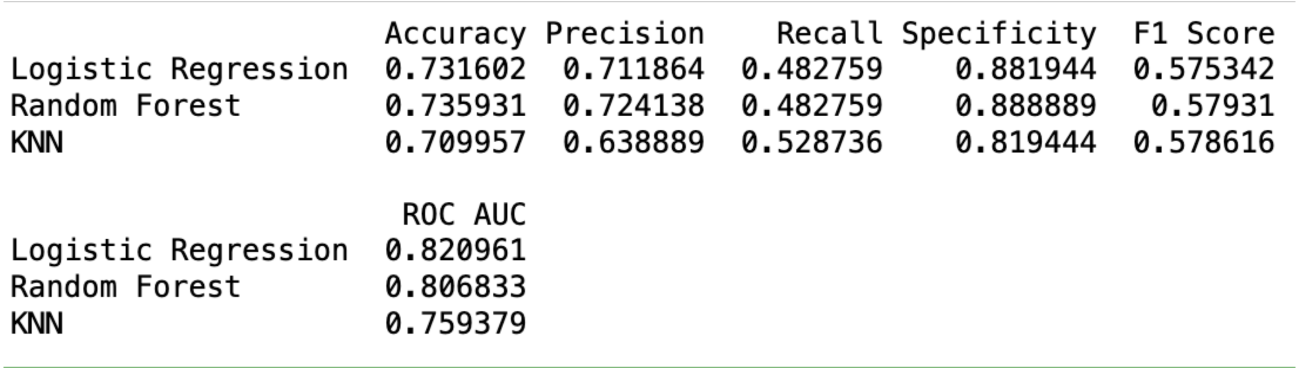


Figure 2. Distribution of Outcome Variable

Machine Learning Models


```
                     Accuracy Precision    Recall Specificity  F1 Score
Logistic Regression  0.731602  0.711864  0.482759     0.881944  0.575342
Random Forest        0.735931  0.724138  0.482759     0.888889   0.57931
KNN                  0.709957  0.638889  0.528736     0.819444  0.578616

                      ROC AUC
Logistic Regression  0.820961
Random Forest        0.806833
KNN                  0.759379
```

Figure 3. Accuracy, Precision, Recall (Sensitivity), Specificity, F1 score, and ROC (Area Under Curve) for Logistic Regression, Random Forest, and K Nearest Neighbors models

Logistic Regression

The Logistic Regression model achieved an accuracy of approximately 73.16%, with a precision of 71.19% and a recall (sensitivity) of 48.28%. This model showed a high specificity (88.19%) and a solid ROC AUC of 0.821. The balance between precision and recall is modest, reflecting in the F1 score of 0.575, suggesting that while the model is specific in identifying negatives, it's somewhat conservative in predicting positives.

Random Forest

Random Forest performed similarly in accuracy (73.59%) and slightly better in precision (72.41%) compared to Logistic Regression. It matched the recall of Logistic Regression at

48.28% but achieved a slightly higher specificity (88.89%) and a slightly higher F1 score (0.579). However, its ROC AUC of 0.807 was slightly lower than that of the Logistic Regression model.

K-Nearest Neighbors (KNN)
The KNN model showed the lowest accuracy (70.99%) but had the highest recall of the three models at 52.87%, indicating it was better at identifying true positives than the other models. Its precision (63.89%) and specificity (81.94%) were the lowest, which could indicate a trade-off where the model captures more true positives at the expense of correctly identifying true negatives. The F1 score was comparable to the other models, but the ROC AUC of 0.759 was the lowest, suggesting its overall ability to discriminate between the classes is less effective than the other models.

Among the three machine learning models evaluated for diabetes prediction, the Random Forest model emerges as the most effective choice. Despite exhibiting similar accuracy to Logistic Regression, Random Forest outperforms both Logistic Regression and KNN in several key metrics. It demonstrates slightly better precision and a marginally higher F1 score compared to Logistic Regression, indicating its ability to achieve a more balanced trade-off between precision and recall. Moreover, Random Forest achieves a higher specificity than both Logistic Regression and KNN, suggesting its superior capability in correctly identifying true negatives. Although its ROC AUC is slightly lower than that of Logistic Regression, the overall performance of Random Forest underscores its effectiveness in discriminating between diabetes and non-diabetes cases.

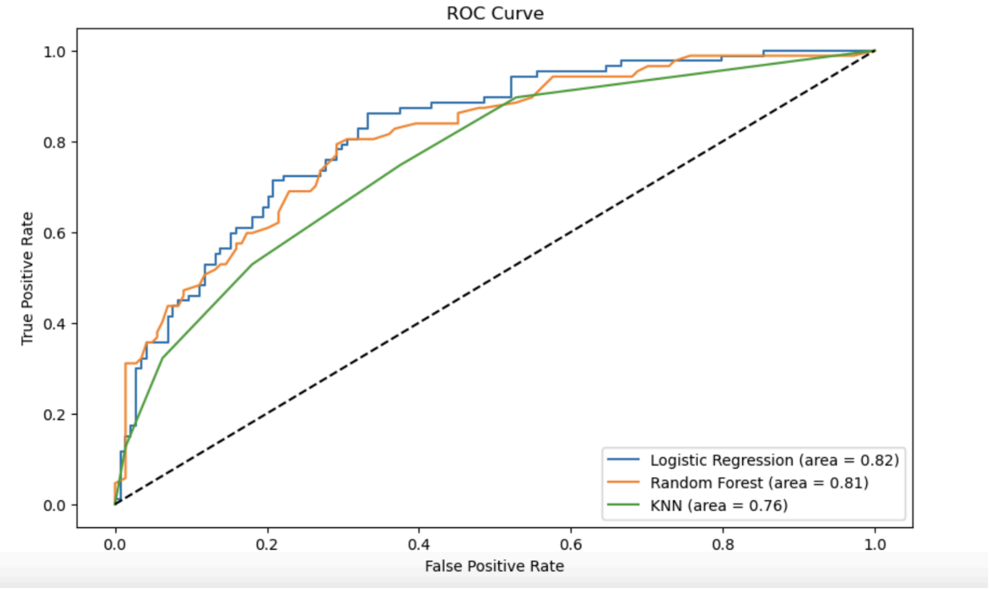


Figure 3. Visualization of ROC curves of the 3 machine learning models.

CONCLUSION

From our analysis of the Pima Indians Diabetes dataset, we've learned how different machine learning models perform in predicting diabetes outcomes, with the Random Forest classifier showing the most promising results in terms of accuracy, precision, sensitivity, specificity, F-score, and area under the ROC curve. The exploration of feature correlations revealed that certain variables such as Glucose and BMI have stronger associations with the diabetes outcome, suggesting their critical role in model predictions. For future extensions, it would be beneficial to investigate feature engineering techniques to enhance model performance, such as creating interaction terms or extracting new features from existing data. It would be beneficial to work with a dataset that contains more data on more features than the 8 in this dataset. Additionally, applying advanced methods like Gradient Boosting or stacking different models could potentially improve prediction accuracy. Exploring the application of machine learning models to personalize diabetes management plans or predict other metabolic disorders could also be a valuable direction for expanding the impact of this research in medical science and healthcare.

Sources

1. https://www.kaggle.com/datasets/nikhilnarasimhan3264/pima-indians-diabetes
2. Joshi, R. D., & Dhakal, C. K. (2021, July 9). Predicting type 2 diabetes using logistic regression and machine learning approaches. International journal of environmental research and public health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8306487/
3. Rao, S. (2023, June 13). What is a precision recall curve and how is it used for machine learning?. Artificial Intelligence +. https://www.aiplusinfo.com/blog/what-is-a-precision-recall-curve-and-how-is-it-used-for-machine-learning/
4. Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B., Yu, J., Li, C., Yu, F., & Ren, Z. (2022, November 15). Machine learning models for data-driven prediction of diabetes by lifestyle type. International journal of environmental research and public health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9690067/
5. https://www.kaggle.com/code/bhavyagoyal867/pima-diabetes-prediction
6. Gulati, A. P. (2024, March 17). Diabetes prediction using machine learning. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/
7. Aman-Preet-Singh-Gulati. (n.d.). Diabetes-prediction-using-ML/Pima diabetes blog.ipynb at main · Aman-Preet-Singh-Gulati/diabetes-prediction-using-ML. GitHub. https://github.com/Aman-Preet-Singh-Gulati/Diabetes-prediction-using-ML/blob/main/PIMA%20diabetes%20blog.ipynb