جامعة أم القرى
UMM AL-QURA UNIVERSITY

Report on Diabetes

Student name	Student ID	Group
Sara Abed Alzulfie	443002749	**2**
Raghad Adel Alzulfie	444003159	**1**

Instructor name: Umaima Falath
Date: 25-10-2024

Introduction

The Pima Indians Diabetes Dataset is a widely used dataset in machine learning and medical research that focuses on predicting the onset of diabetes in adult women of Pima Indian heritage. This dataset, originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases, includes a collection of diagnostic measurements related to diabetes risk factors, such as age, blood pressure, body mass index (BMI), and glucose levels. The primary goal is to develop predictive models that can identify whether a patient is likely to develop diabetes based on these features. Due to its balanced complexity and the real-world significance of diabetes, the dataset is often used as a benchmark for evaluating machine learning algorithms and for educational purposes in data science.

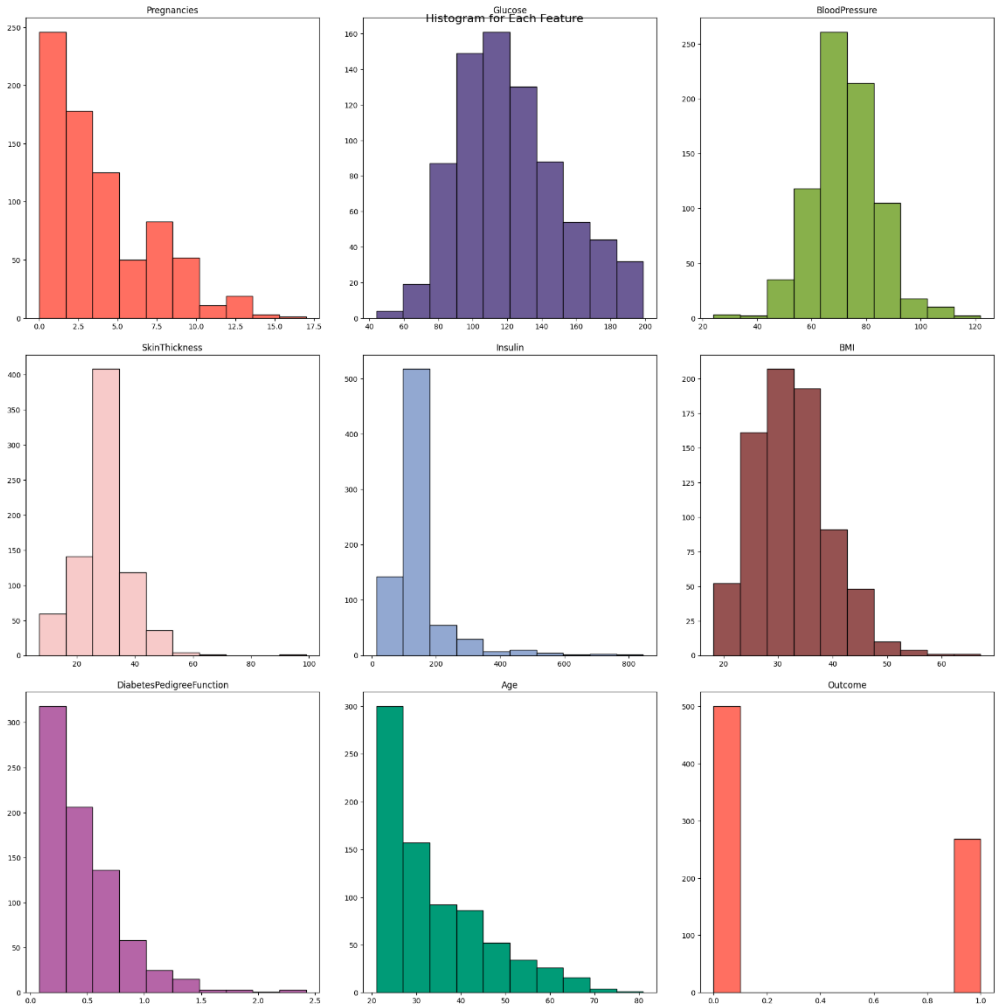
Characteristics of Data

- Pregnancies: Number of times the patient has been pregnant.
- Glucose: Plasma glucose concentration measured 2 hours after a glucose tolerance test.
- Blood Pressure: Diastolic blood pressure (mm Hg), the lower number in a blood pressure reading.
- Skin Thickness: Triceps skinfold thickness (in millimeters).
- Insulin: 2-hour serum insulin (measured in µU/mL).
- BMI: Body Mass Index
- Diabetes Pedigree Function: A function that scores the likelihood of diabetes based on family history.
- Age: the patient
- Outcome: the result of patients: 1 positive, 0 negative.

Data link; https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data.

Data Exploration

The data has 9 characters and 768 cases or record
While we were scraping the dataset we found that the minimum
On these columns, a value of zero does not make sense and thus indicates missing value;
Glucose
Blood Pressure
Skin Thickness
Insulin
BMI
It is better to replace zeros with nan since after that counting them would be easier and zeros need to be replaced with suitable value Aiming to impute nan values for the columns in accordance with their distribution so we filled it with mean this histogram shows the result:



Learning models.

Naive Bayes Classifier

The Naïve Bayes classifier is a supervised machine learning algorithm that is used for classification tasks. They use principles of probability to perform classification tasks. Naïve Bayes is part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category. Unlike discriminative classifiers, like logistic regression, it does not learn which features are most important to differentiate between classes.
eBook the data store for AI
Discover the power of integrating a data Lakehouse strategy into your data architecture, including enhancements to scale AI and cost optimization opportunities.
Overall Performance: 76% accuracy suggests a reasonably good model, but there's room for
Comparison
improvement.
Both Random Forest and Naive Bayes achieved similar accuracies in your case (around 77% and 76%, respectively). This suggests that both models are reasonably good at predicting diabetes based on the given features. The choice between the two might depend on factors like interpretability, computational cost, and the specific characteristics of your dataset.

learning models.

Random forest classifier

A random forest classifier.
A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Trees in the forest use the best split strategy, i.e. equivalent to passing splitter="best" to the underlying **DecisionTreeClassifier**. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

Interpreting the Results:
Overall Performance: 77% accuracy suggests a reasonably good model, but there's room for improvement.
Confusion Matrix Insights: Analyze the values in your confusion matrix. Higher TP and TN values indicate better performance. Focus on reducing FP and FN for a more accurate model. For example, if FN is high, it means the model is missing a significant number of actual diabetes cases, which could be problematic.
Further Considerations:
Feature Importance: Random Forests allow you to check feature importance. Identify the features that most strongly influence the predictions. This can provide insights into the factors driving diabetes prediction.
Hyperparameter Tuning: Experiment with different hyperparameters of the Random Forest model (e.g., number of trees, tree depth) to potentially improve its accuracy.
Data Balancing: If your dataset is imbalanced (more cases of one outcome than the other), consider techniques like oversampling or undersampling to address this and improve model performance

Conclusion

In this study, we applied two different classification algorithms—Naive Bayes and Random Forest—to predict the presence of diabetes. Both models achieved comparable accuracies, with Naive Bayes scoring 76% and Random Forest slightly outperforming it with 77%.

Naive Bayes Classifier:

Naive Bayes, being a generative model, worked reasonably well with an accuracy of 76%. While it assumes independence between features, which may not always hold in medical datasets, its simplicity and efficiency make it useful, especially for quick baseline models. However, this assumption limits its ability to model complex feature interactions, which is evident from the marginal performance gap compared to Random Forest.

Random Forest Classifier:

Random Forest performed slightly better with an accuracy of 77%. It's a more robust algorithm, particularly effective for handling non-linear relationships and interactions between features. Its use of multiple decision trees helps reduce overfitting, leading to more reliable results. Additionally, Random Forest offers valuable insights through feature importance, which can help identify the most influential predictors of diabetes. This can be particularly useful in understanding the key factors driving the disease. Moreover, further improvement in performance can potentially be achieved by hyperparameter tuning and addressing data imbalances.

Comparison and Final Thoughts:

Both models performed similarly, with Random Forest having a slight edge due to its ability to model complex patterns and provide insights into feature importance. Naive Bayes, while slightly less accurate, remains useful for its simplicity and speed. Depending on the context, whether the focus is on interpretability, computational efficiency, or predictive accuracy—either model could be appropriate.

In conclusion, for this diabetes classification problem, Random Forest offers a better balance of performance and interpretability. Still, with proper adjustments (e.g., hyperparameter tuning and balancing techniques), there is potential to improve both models' performances.

References

- https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data
- https://www.ibm.com/topics/naive-bayes
- https://www.kaggle.com/code/shrutimechlearn/step-by-step-diabetes-classification


جامعة أم القرى
UMM AL-QURA UNIVERSITY


جامعة أم القرى
UMM AL-QURA UNIVERSITY


جامعة أم القرى
UMM AL-QURA UNIVERSITY


جامعة أم القرى
UMM AL-QURA UNIVERSITY


جامعة أم القرى
UMM AL-QURA UNIVERSITY


جامعة أم القرى
UMM AL-QURA UNIVERSITY