Diabetes Prediction Analysis Report

1. Introduction

Diabetes is a chronic health condition characterized by high blood sugar levels. Early diagnosis and management are crucial to prevent complications and improve patient outcomes. This project utilizes a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases to develop an Artificial Neural Network (ANN) that predicts whether a patient has diabetes based on certain diagnostic measurements. The dataset consists of 9 features for each patient, including physiological measurements and medical history.

2. Problem Statement

The primary objective of this project is to develop a predictive model that can accurately determine whether a patient has diabetes. Given the complexity of the disease and the dataset's constraints (e.g., all patients are females of Pima Indian heritage, aged at least 21 years), the challenge lies in selecting and tuning a model that effectively captures the underlying patterns in the data to provide reliable predictions.

3. Approach

Data Collection and Cleaning

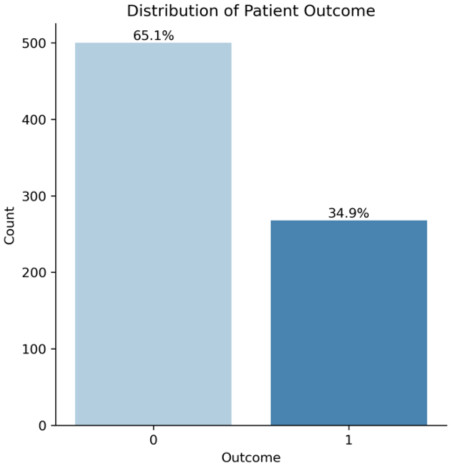
The dataset includes the following features:

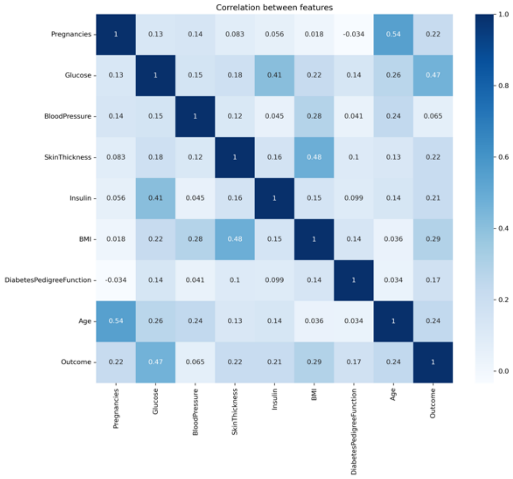
- **Pregnancies**: Number of pregnancies
- **Glucose**: Glucose level in blood
- **BloodPressure**: Blood pressure measurement
- **SkinThickness**: Skin thickness
- **Insulin**: Insulin level in blood
- **BMI**: Body mass index
- **DiabetesPedigreeFunction**: Diabetes pedigree function (genetic risk indicator)
- **Age**: Age in years
- **Outcome**: Binary variable indicating the presence (1) or absence (0) of diabetes

Data cleaning involved addressing missing values, particularly for the `SkinThickness` and `Insulin` columns, by replacing zero entries with the mean of each column. This step was necessary to ensure the completeness of the dataset and prevent errors during model training.

Exploratory Data Analysis (EDA)

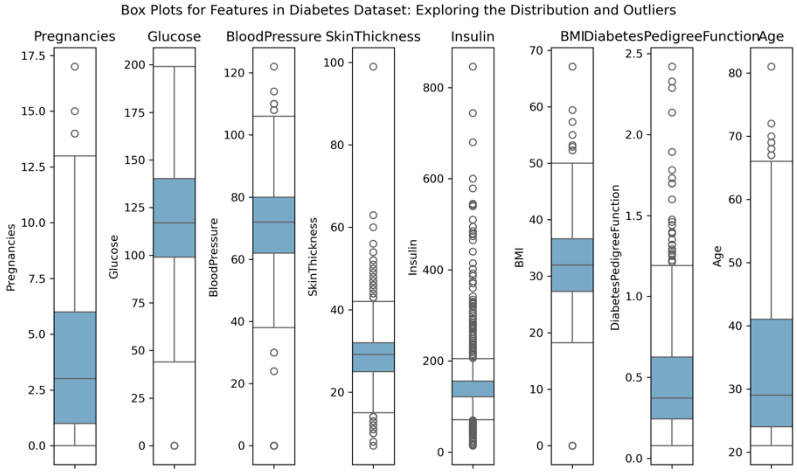
- **Outcome Distribution**: The dataset shows an imbalance, with 65.1% of patients not having diabetes (Outcome = 0) and 34.9% having diabetes (Outcome = 1).
- **Feature Distributions**: Most features exhibit right-skewed distributions, with notable outliers in `Pregnancies`, `Insulin`, and `SkinThickness`.





- **Correlation Analysis**: The highest correlation with the diabetes outcome is found in glucose levels (0.47), followed by BMI (0.29) and age (0.24).

- **Outlier Analysis**: The percentage of outliers in the dataset is low (0.03%), indicating that their impact on the ANN's performance is minimal.



4. Modeling and Evaluation

Initial Model

The initial ANN was constructed using Keras, comprising one hidden layer. The model was trained on 80% of the data and evaluated on the remaining 20%. The initial accuracy was 75.32%, with notable precision and recall differences between diabetic and non-diabetic predictions.

- **Non-Diabetic Predictions**: Precision = 0.80, Recall = 0.82, F1-score = 0.81
- **Diabetic Predictions**: Precision = 0.66, Recall = 0.64, F1-score = 0.65

Model Optimization

- **Cross-Validation**: Implemented K-fold cross-validation (5 splits), yielding a mean accuracy of 65.31% with a standard deviation of 0.34%.
- **Hyperparameter Tuning**: Grid search identified optimal parameters: batch size of 10, 150 epochs, `he_uniform`initialization, and `Adam` optimizer.
- **Improved Model**: Added batch normalization and dropout layers to prevent overfitting. This model achieved a cross-validation mean accuracy of 76.22% with a standard deviation of 2.63%.

Advanced Training Techniques

- **Early Stopping and ReduceLROnPlateau**: Applied to enhance model convergence and generalization. The final model reached an improved test accuracy of 77.27%.
- **Confusion Matrix**: Demonstrated balanced predictions, with a true negative rate of 85, false positive rate of 14, true positive rate of 34, and false negative rate of 21.

5. Final Model Explanation

The final ANN architecture consists of multiple layers with dropout and batch normalization. The model successfully predicts diabetes outcomes by capturing complex patterns in the dataset, demonstrating improved performance over the initial model. The use of advanced training techniques and hyperparameter tuning contributed to its robustness and accuracy.

6. How to Improve the Model

- **Data Imbalance**: Address class imbalance using techniques like Synthetic Minority Over-sampling Technique (SMOTE), resampling, or adjusting class weights during training.
- **Feature Transformation**: Apply transformations (e.g., log transformation) to normalize skewed features, potentially improving model performance.

7. Personal Conclusion

The project effectively demonstrated the application of ANN in predicting diabetes outcomes. The iterative approach, involving data preprocessing, exploratory analysis, model building, and optimization, highlights the importance of systematic model development. The improved model offers a reliable tool for diabetes prediction, emphasizing the potential of machine learning in healthcare applications.

8. Ideas for Further Research

- **Feature Engineering**: Explore additional features or interactions between existing features to enhance predictive accuracy.
- **Ensemble Models**: Investigate ensemble techniques, such as stacking or boosting, to combine multiple models and potentially improve performance.
- **Longitudinal Data**: Incorporate temporal data to capture disease progression over time, providing insights into patient trajectories.

9. Concrete Recommendations for Implementation

- **Integration into Healthcare Systems**: Deploy the ANN model as a decision-support tool for healthcare professionals, aiding in early diabetes diagnosis.
- **User-Friendly Interface**: Develop a user-friendly interface for clinicians to input patient data and receive predictions seamlessly.
- **Continuous Monitoring and Updates**: Regularly update the model with new data to maintain accuracy and relevance, adapting to changes in patient demographics and clinical practices.