Sudeshna Ghosh
(23122134)
3MSc.DS(B)

Predicting Diabetes with K-Nearest Neighbors: An Exploratory Analysis

Introduction:

Diabetes is a chronic disease that affects millions of people worldwide. Early detection and diagnosis are crucial for effective management and prevention of complications. In this study, we explore the use of the K-Nearest Neighbors (KNN) algorithm, a popular machine learning technique, to predict the likelihood of diabetes based on various features such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age.

Dataset and Preprocessing:

The analysis is based on a sample dataset containing 9 features and 10 instances. The data was first loaded into a pandas DataFrame, and the features were separated from the target variable (Outcome). The dataset was split into training and test sets using scikit-learn's train_test_split function, with a test size of 20%. The features were then scaled using StandardScaler to ensure that all features are on a similar scale and contribute equally to the model.

Exploratory Data Analysis:

Before diving into the modeling process, we performed exploratory data analysis to gain insights into the dataset. A correlation heatmap was generated to visualize the relationships between the features. This heatmap revealed that

certain features, such as Glucose and BMI, had stronger correlations with the Outcome variable than others.

Additionally, we created a pair plot to explore the pairwise relationships between all features and the distribution of each feature. The pair plot can help identify potential patterns or separability between the classes based on the feature combinations.

To further investigate the distributions of the individual features, we plotted histograms for each feature using Seaborn's distplot function. These distribution plots can reveal the shape and spread of each feature, which can be useful for identifying potential outliers or skewed distributions.

Model Training and Evaluation:

With the data preprocessed and exploratory analysis completed, we trained a KNN model using scikit-learn's KNeighborsClassifier. The number of neighbors was set to 3, but this parameter can be tuned further to improve model performance.

To evaluate the model's performance, we made predictions on the test set and calculated the confusion matrix and classification report. The confusion matrix provides a visual representation of the model's predictions, showing the number of true positives, true negatives, false positives, and false negatives. The classification report summarizes the precision, recall, and F1-score for each class, which are useful metrics for evaluating the model's performance.

Visualization of Decision Boundaries:

To better understand the decision boundaries of the KNN model, we visualized the decision boundaries using a subset of two features: Glucose and BMI. This visualization can help identify regions in the feature space where the model is likely to predict each class.

We plotted the decision boundaries for both the training and test sets, providing insights into the model's behavior and potential overfitting or underfitting issues.

Dimensionality Reduction: To further explore the relationships between features and improve the interpretability of the correlation heatmap, we applied Principal Component Analysis (PCA) to the correlation matrix. PCA is a dimensionality reduction technique that transforms the data into a new set of uncorrelated variables called principal components.

By reducing the dimensionality of the correlation matrix to two principal components, we were able to visualize the relationships between the original features and the principal components in a reduced correlation heatmap.

Conclusion:

This article presents an exploratory analysis of a diabetes dataset using the K-Nearest Neighbors algorithm. Through various visualizations and model evaluation techniques, we gained insights into the relationships between features, the distributions of individual features, and the performance of the KNN model in predicting diabetes.

While the sample dataset used in this analysis is relatively small, the techniques and visualizations demonstrated can be applied to larger datasets to potentially improve the model's performance and understanding of the underlying patterns in the data.

Future work could involve further tuning of the KNN model's hyperparameters, exploring other machine learning algorithms, or incorporating additional relevant features to enhance the predictive accuracy of the model.