Jefferson Veliz

Professor Jairam

Data mining for Business Analytics

August 8th, 2023

Lab #5

The motivations to dive into this lab was driven by a strong interest in understanding the intricate connections between diabetes and machine learning. Diabetes is a significant health concern affecting numerous individuals, and the prospect of leveraging machine learning to uncover potential patterns and insights in the data is truly captivating. The dataset at hand offers various features like pregnancies, blood sugar levels, and more. I am eager to explore these variables to decipher any potential correlations with the occurrence of diabetes. What really intrigued me was the chance to utilize a mix of supervised and unsupervised machine learning techniques.

Ensuring data quality is critical for reliable analysis. I diligently examined the dataset for missing values (na) and discrepancies in data types. This step aimed to uncover any potential data gaps that might hinder accurate analysis. Identifying and addressing missing values is crucial, as it can impact the validity of our findings and the performance of subsequent machine learning models. Additionally, verifying that data types align with the nature of the information they represent is essential for accurate computations and meaningful insights.

To understand how different variables within the dataset influence each other, I constructed a correlation plot using the seaborn library. The heatmap generated from the correlation matrix vividly illustrated the relationships between variables. The color-coded scheme provided an instant visual representation of the strength and direction of these relationships. Notably, the inclusion of actual correlation coefficients within the heatmap cells, formatted as percentages without decimal places, added a quantitative layer to our insights. The color map used (cool warm) further facilitated easy interpretation of the heatmap.

The classification report provides essential insights regarding the performance of our machine learning model in predicting diabetes outcomes. In the context of class 0 (non-diabetes), a precision value of 0.80 denotes that 80% of the predictions for non-diabetes instances were accurate. With a recall value of 0.89, the model effectively identified 89% of actual non-diabetes cases. The F1-score is noted at 0.85 for class 0. Notably, the dataset contains 157 instances belonging to class 0.

Shifting to class 1 (diabetes), the model showcases a precision value of 0.70, indicating that 70% of the predicted diabetes instances were accurate. The recall value of 0.53 highlights the model's ability to recognize 53% of real diabetes cases. Demonstrating a skillful equilibrium between precision and recall within this class, the F1-score stands at 0.60.

Additionally, I conducted a K-means clustering analysis on the dataset. The evaluation of the clustering results yielded a purity score of 65.89%. This score indicates the extent to which data points within the same clusters belong to the same true class. A higher purity score suggests a stronger alignment between the clusters and the actual classes, signifying the effectiveness of the clustering approach in capturing inherent patterns within the data.

Moreover, I continued to perform K-means clustering on the data. This method groups similar data points together. When I used different ways to measure the distances between points, I got different scores that show how well the groups match the real categories. For one way of measuring distance, called Manhattan distance, the score was 66.15%. Another way, called Euclidean distance, gave a score of 65.89%, as did squared Euclidean distance and Chebyshev distance. Another method, called Canberra distance, resulted in a score of 65.1%. Lastly, using the Chi-square distance, I also got a score of 65.1%. These scores tell us how well the groups represent the actual categories in the data.

When i compare the two methods, I find that the Logistic Regression model did a better job with accuracy and gave us detailed insights for each group. It used known information to make predictions and used clear measurements to tell us how well it did. On the other hand, K-means clustering showed us patterns in the data without knowing the groups beforehand. It didn't tell us about each group like the Logistic Regression did. Which method to choose depends on what we're trying to do. If I want accurate predictions and to understand why, we'd go with Logistic Regression. But if I just want to see patterns without knowing groups in advance, K-means clustering is the way to go.