Assignment_2_Log_Reg

Michael Montanez

C0889555

Lambton College

AML 3104 - Neural Networks and Deep Learning

Prof. Ishant Gupta

Feb 27, 2024

Instructions:

Objective: The objective of this assignment is to build a predictive model to predict the likelihood of a patient having diabetes based on certain features.

Dataset: You will use the "diabetes" dataset provided. The dataset contains information about the medical history of patients, including features like Glucose level, Blood Pressure, BMI, etc., and a target variable indicating whether the patient has diabetes (1) or not (0).

Tasks:

- Explore the dataset to understand its structure and contents.
- Perform any necessary data preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features.
- Split the dataset into training and testing sets (e.g., 80% training and 20% testing).
- Build a Logistic Regression model to predict the likelihood of diabetes based on the features provided.
- Evaluate the model using appropriate metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
- Interpret the model coefficients to understand the impact of different features on the likelihood of diabetes.

Deliverables:

- Jupyter Notebook (or Python script) containing the code for data preprocessing, model building, and evaluation.
- Report summarizing the key findings, model performance metrics, and insights from the model coefficients.

Upload the notebook on moodle and github and share the link

Submission:

- Submit the Jupyter Notebook (or Python script) and the report summarizing your analysis.
- Include any additional insights, visualizations, or improvements you made to the model.

Resources:

- You can refer to Python libraries such as Pandas, NumPy, Scikit-learn for data manipulation, model building, and evaluation.
- Feel free to reach out for any clarifications or assistance during the assignment.

Deadline: Complete the assignment and submit it within [specified deadline].

Report:

Comprising 768 records and 8 features, including 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', and 'Age', the dataset encloses a diverse array of patient information. The target variable, 'Outcome', classifies patients as either diabetic (1) or non-diabetic (0). Statistical exploration reveals varying distributions and ranges across features, illustrating the dataset's complexity.

Data Preprocessing:

Before model construction, preprocessing steps were essential. This included handling missing values (There weren’t null values) and standardizing numerical attributes. Notably, zero values in critical features such as 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', and 'BMI' were substituted with their respective means to ensure data integrity because it didn’t make any sense for those features.

Model Development:

Logistic regression was employed to predict diabetes risk due to its simplicity and effectiveness in binary classification. The model was trained on 80% of the dataset, reserving 20% for testing purposes using the python library.

Model Evaluation:

The trained logistic regression model yielded the following evaluation metrics on the test set:

Accuracy: 78.57%

Precision: 76.32%

Recall: 54.72%

F1 Score: 63.74%

ROC AUC Score: 72.90%

These metrics offer insights into the model's classification accuracy and ability to balance precision and recall. While the model demonstrates satisfactory accuracy, enhancements in recall are warranted to capture more positive cases effectively.

Feature Importance Analysis:

An examination of model coefficients provides insights into feature importance. Notably, 'Glucose' and 'BMI' emerge as significant predictors positively associated with diabetes likelihood.

Conversely, 'BloodPressure' and 'Insulin' exhibit negative associations, suggesting potential protective effects against diabetes.

Conclusion and Recommendations:

In conclusion, the logistic regression model provides valuable insights into diabetes risk assessment. However, further refinement is necessary to improve recall and capture more positive cases accurately. Future work avenues may include exploring ensemble learning methods and integrating additional data sources for enhanced predictive performance. Also, with the coefficients, we can determine if we can delete some features or not.

Github link: https://github.com/mumuljuve/DiabetesMike