Classification in Healthcare Dataset

Objective

The primary objective of this assignment is to provide you with hands-on experience in performing classification on real-world datasets with machine learning methods you have learned in class.

Instructions

Download dataset

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. The dataset consists of several medical predictor variables and one target variable, Outcome.

Download the dataset here - link
Dataset Reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8943493/

Classification using Machine Learning Methods

Using Google Colab or Jupyter Notebook, write Python code to classify the above dataset. The classification process should include EDA, feature selection, and model training & evaluation.

Exploratory Data Analysis

1. **Data Loading:** Load the dataset into your notebook.
2. **Data Inspection:** Display the first few rows of the dataset, check for missing values, and provide summary statistics.
3. **Data Visualization:** Create relevant plots (e.g., histograms, box plots, scatter plots, correlation matrices) to visualize the data distribution and relationships between features.

4. **Data Preprocessing:** If necessary, perform data preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features.

Feature Selection (Filtering / Statistics)
After completing the EDA, implement feature selection techniques using filtering and statistical methods. The goal is to identify and select the most relevant features for machine learning models. Include the following steps:

- Feature Ranking: Use statistical methods (e.g., correlation coefficients, mutual information) to rank the features based on their importance.
- Feature Selection: Select a subset of the most important features based on the ranking obtained in the previous step.

Which features will you propose using in the machine learning model based on statistical analysis? Why use these features?

Model Selection, Training, and Evaluation
In this part of the assignment, you will select, train, and evaluate machine learning models on the provided the above dataset to predict the target variable (diabetes diagnosis). This process will help you gain practical experience in building and assessing classification models for healthcare applications.

1. **Model Selection:** Begin by selecting at least two different classification algorithms (Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, Neural Networks) to build your predictive models. ***Briefly explain your reasons for choosing these algorithms.*** Consider factors such as simplicity, interpretability, and performance.
2. **Data Splitting:** Split your dataset into a training set and a testing set. You will use the training set to train your models and the testing set to evaluate their performance.
3. **Model Training:** Train each of your selected models on the training dataset. Make sure to follow best practices, including handling missing data, feature scaling, and encoding categorical variables as needed. Tune the

hyperparameters of your models for optimal performance. You can use techniques like grid search or random search for hyperparameter tuning.

4. **Model Evaluation:** Evaluate the performance of each model using appropriate classification metrics. Common metrics include accuracy, precision, recall, F1-score, and ROC-AUC. Create a confusion matrix for each model to visualize the true positives, true negatives, false positives, and false negatives.
5. **Comparison and Selection:** Compare the performance of your selected models. Discuss which model(s) performed better in terms of the chosen evaluation metrics. Consider the practical implications of model performance in a healthcare context. ***Which model would be more suitable for real-world deployment?***
6. **Interpretability and Explainability:** For each model, analyze the features' importance or contribution. Explain which features are most influential in making predictions. **Discuss the interpretability of your models and any potential challenges in explaining their decisions in a healthcare setting.**

You can use any Python libraries you want. Some good ones include:

- Data structure: pandas,
- Math: numpy, scipy
- Visualizations: matplotlib, sns
- ML Libraries: sklearn, tensorflow, pytorch

Documentation and Analysis

Provide a written analysis of your findings. Discuss why certain features were selected or excluded based on the EDA and feature selection results. Discuss your selection of the machine learning model based on the training and evaluation results.

Submission Guidelines

- Submit your assignment as a Jupyter Notebook file or Google Colab Notebook link in EMAS2.
- Include clear and organized code, comments, and explanations.
- Clearly state which dataset you chose.
- Submit your analysis as part of the notebook file, there's no need to create a separate document.

Grading Criteria

The assignment will be graded based on the following criteria:

- Problem and dataset description (5%)
- Completeness and correctness of EDA process (10%)
- Implementation of feature selection techniques (15%)
- Data splitting, model selection, and training (25%)
- Model evaluation and hyperparameter tuning (25%)
- Documentation and clarity of code (10%)
- Quality of written analysis (10%)

Note

If you need any assistance, feel free to ask me. You can use any resource you need to work on this assignment (including GPT), but please make sure you understand what you are putting into the notebook. ***Be honest with yourself, your future self will be glad you did.***