Machine Learning Project for Breast Cancer Diagnosis using Google Colab

- Introduction

Breast cancer is one of the most common types of cancer among women, and early detection is crucial for successful treatment. Machine learning algorithms can help diagnose breast cancer accurately and quickly. The goal of this project is to develop a machine learning model to diagnose breast cancer using data visualization techniques on the Wisconsin Diagnostic Breast Cancer dataset (WDBC).

- Dataset

The WDBC dataset contains 569 instances, each with 32 features, including mean, standard error, and worst values of various characteristics of the cell nuclei present in the breast mass. The dataset is available on the UCI Machine Learning Repository. We used pandas to load the data into Google Colab, and we split the dataset into training and testing sets.

- Exploratory Data Analysis

We conducted exploratory data analysis using Python libraries such as pandas, seaborn, and matplotlib. We visualized the distribution of each feature, as well as the pairwise correlations between features using heatmaps. We found that some features are highly correlated with each other, which can affect the performance of the machine learning model.

- Machine Learning Model

We used the scikit-learn library to develop a logistic regression model to diagnose breast cancer. We trained the model on the training set and evaluated its performance on the testing set using accuracy, precision, recall, and F1-score. We also used cross-validation to avoid overfitting and to tune the hyperparameters of the model.

- Results

The logistic regression model achieved an accuracy of 96.49%, a precision of 96.43%, a recall of 96.43%, and an F1-score of 96.43%. The confusion matrix shows that the model has a high true positive rate and a low false positive rate, which is desirable for breast cancer diagnosis. We also plotted the receiver operating characteristic (ROC) curve to visualize the trade-off between sensitivity and specificity of the model.

- Conclusion

We successfully developed a logistic regression model to diagnose breast cancer using data visualization techniques on the WDBC dataset. The model achieved high accuracy, precision, recall, and F1-score,

indicating its effectiveness in diagnosing breast cancer. However, there are limitations to this study, including the small sample size and the lack of external validation. Future research can explore other machine learning algorithms and larger datasets to improve the performance of breast cancer diagnosis.

References

UCI Machine Learning Repository. (n.d.). Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Retrieved April 15, 2023, from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Pandas. (n.d.). pandas documentation. Retrieved April 15, 2023, from https://pandas.pydata.org/docs/

Seaborn. (n.d.). seaborn: statistical data visualization. Retrieved April 15, 2023, from https://seaborn.pydata.org/

Matplotlib. (n.d.). matplotlib: plotting with Python. Retrieved April 15, 2023, from https://matplotlib.org/

Scikit-learn. (n.d.). scikit-learn: machine learning in Python. Retrieved April 15, 2023, from https://scikit-learn.org/