CSC-177: DATA ANALYTICS AND MINING

PROJECT 2: Classification Models Project

Due Date: Nov 17th 2023

Team Challengers (23)
1. Srujay Reddy Vangoor
2. Vaibhav Jain
3. Bashar Allwza
4. Varun Bailapudi
5. Uddayankith Chodagam

Executive Summary:

This report presents a comprehensive analysis of various classification algorithms applied to two distinct datasets. The project, conducted as part of the CSC177 course, is divided into two phases: Phase I involves applying a set of algorithms to a guided dataset, and Phase II centers around the application of these algorithms to a dataset independently chosen by the team. The primary aim is to evaluate the effectiveness of different classification algorithms in diverse scenarios, thereby gaining practical insights into their application in the field of data science.

1.0 Introduction

Classification algorithms are a cornerstone of machine learning, offering a multitude of approaches to categorize data. This study aims to not only apply these algorithms but also to critically analyze their performance under varying conditions, thereby contributing to the broader understanding of their practical utility in data science.

2. Methodological Framework

2.1 Phase I - Guided Dataset Analysis:

Preliminary Analysis and Preprocessing: The project commenced with the exploration of the provided dataset using Python and associated libraries. Key steps included data visualization for initial analysis and preprocessing activities like feature scaling and encoding.

Algorithm Implementation: The study involved the implementation of Logistic Regression, K-Nearest Neighbors, and Decision Trees, chosen for their relevance to the dataset's characteristics.

2.2 Phase II - Independent Dataset Analysis:

Dataset Curation and Cleanup: An autonomous selection of a secondary dataset was conducted, followed by data cleansing including the elimination of null values and duplicates.

Visualization and Feature Optimization: Employing advanced data visualization techniques, the study focused on understanding data correlations and outlier detection. Feature normalization was a critical step in preparing the dataset for algorithm application.

Algorithm Application and Tuning: A spectrum of models including SVC, KNN, Naive Bayes, Decision Tree, and Logistic Regression were applied. This phase emphasized parameter tuning for optimizing model performance.

3. Analysis and Observations:

3.1 Phase I - Guided Dataset Analysis:
Outcomes: The application of KNN and Decision Trees yielded significant insights. Detailed performance metrics analysis highlighted the models' strengths and weaknesses.
Learnings: The importance of feature selection in model efficacy was a key takeaway.

3.2 Phase II - Independent Dataset Analysis:
Insights Gained: Techniques such as data shuffling and outlier analysis were critical. The adaptability and accuracy of the SVC and Decision Trees in specific scenarios were noteworthy.
Comparative Evaluation: The models displayed varying degrees of effectiveness, underscoring the importance of choosing the right model based on the dataset's unique characteristics.

4. Conclusion:
The project offered valuable insights into the practical aspects of classification algorithms in machine learning. It highlighted the importance of data preprocessing and strategic model selection, providing a foundation for future research and application in the field.

5. Recommendations for Future Research:
Further exploration into ensemble learning techniques and advanced hyperparameter optimization methods is suggested. Additionally, delving into model interpretability and the significance of feature importance in classification tasks would be beneficial for comprehensive understanding and application.