Report: IML Assignment # 04

By

Name: Muhammad Tanzeel Saleem
Reg. No: 04071913026
Dataset: Student Performance

Introduction:

This work aimed to predict student performance using various machine learning algorithms applied to a dataset of student records. The objective was to determine whether a student would perform well or not. The algorithms used included logistic regression, neural network, decision trees, random forest, and Naïve Bayes. Principal Component Analysis (PCA) was utilized to investigate the impact of dimensionality reduction on model performance. The evaluation focused on precision and recall metrics, providing insights into the accuracy and effectiveness of the models in predicting student performance. By comparing different algorithms with and without PCA, the study sought to identify the most effective machine learning approaches for predicting student performance based on the available dataset.

Dataset:

The dataset approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires.

Methods used in this Assignment:

- **Logistic Regression:** Logistic regression is a linear classifier that models the relationship between the features and the probability of a game being won. It was applied both with and without PCA.

- **Neural Network:** A neural network is a powerful algorithm capable of capturing complex patterns in the data. A feedforward neural network with multiple hidden layers was trained for this task. PCA was applied to the dataset before training the neural network.
- **Decision Trees:** Decision trees partition the feature space based on a set of rules. Each branch represents a decision, and the leaf nodes correspond to the class labels. Decision trees were built and evaluated with and without PCA.
- **Random Forest:** Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It provides robustness against overfitting and improves generalization. Random forest models were trained using the dataset with and without PCA.
- **Naïve Bayes:** Naïve Bayes is a probabilistic classifier based on Bayes' theorem. It assumes independence between features and calculates the posterior probabilities. Naïve Bayes classifiers were trained and evaluated with and without PCA.

Results & Analysis

The following are the precision and recall results obtained for each algorithm with and without PCA:

Without PCA

Logistic Regression:

Precision: 0.8793

Recall: 0.9448

Neural Network:

Precision: 0.8802

Recall: 0.9357

Decision Tree

Precision: 0.8535

Recall: 0.8410

Random Forest

Precision: 0.8713

Recall: 0.9312

Naïve Bayes

Precision: 0.8667

Recall: 0.8794

With PCA

Logistic Regression:

Precision: 0.8582

Recall: 0.9549

Neural Network:

Precision: 0.8715

Recall: 0.9402

Decision Tree

Precision: 0.8225

Recall: 0.8253

Random Forest

Precision: 0.8511

Recall: 0.9538

Naïve Bayes

Precision: 0.8535

Recall: 0.8343

The analysis of the precision and recall results provides valuable insights into the performance of the algorithms with and without PCA. It is important to note that the impact of PCA varies across different algorithms. Logistic Regression and Random Forest benefit from PCA, particularly in terms of recall, suggesting that PCA helps in improving their ability to identify positive instances. The Neural Network remains relatively stable, indicating that it is less influenced by PCA. However, Decision Tree and Naïve Bayes show decreased performance in terms of precision and recall with PCA, suggesting that PCA might not be beneficial.

Conclusion:

It is concluded that precision and recall results with and without PCA reveals that Logistic Regression and Random Forest benefit from PCA, showing improved recall without significant impact on precision. The Neural Network remains stable regardless of PCA. However, Decision Tree and Naïve Bayes demonstrate decreased performance with PCA, suggesting it may not be beneficial for these algorithms. The impact of PCA

varies across algorithms, emphasizing the importance of considering specific dataset characteristics and other evaluation metrics when selecting an algorithm.