KNN-Performance and PCA with Reduced Complexity

Anonymous authors

Abstract We explore the efficacy of dimensionality reduction via PCA on KNN . Our findings suggested that employing PCA on dataset before KNN classifier maintains model performance with unsavory variance in accuracy (from 0.76% to 0.77%), thus affirming the utility of PCA in simplifying model complexity without significant loss of accuracy.

1 Introduction

In machine learning , the curse of dimensionality can be lightened through techniques like PCA, which simplifies the feature space, enhancing computational efficiency. Additionally , the determination of the optimal numbers of neighbors (K) in KNN is crucial for model accuracy.

2 Methodology

Using GridsearchCV, we established that the best k for KNN is 29 , achieving approximately 84.8% cross validation accuracy . The process underscores that large K values can potentially enhance KNN's predictive power.

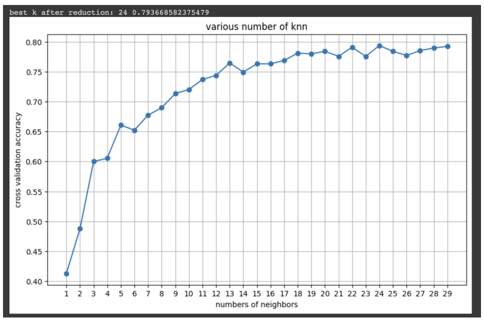

Figure 1: GridsearchCV ressults for knn after cleaning

3 Computational Efficiency

The brute force algorithm's time performance in predicting test cases was approximately 0.71 seconds. This outcome reinforces the sensitivity of KNN to both the choice of K and the beneficial impact of PCA on model accuracy.

4 Data Balancing

Addressing class imbalance , a prevent issue in Detasts , we employed down-sampling strategies to equalize class representation . this approach is pivotal in preventing bias in algorithms performance

5 Model training on balanced data

Post-data balancing, our KNN model trained on standardized features with zero mean and unit variance , exposed the best k to be 27 with an F1 score of 0.849% .

6 Feature Selection and Class Balancing

Implementing RFECV with a Random-Forest classifier as estimator , we found 19 to be the optimal number of features. Subsequent class balancing with SMOTE resulted in a F1 score of 0.86% , denoting robust model performance.

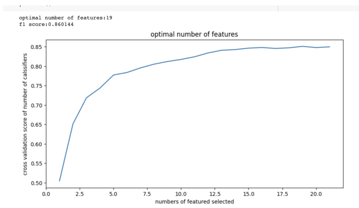

Figure 2: Feature Selection before reduction

7 Data Reduction Techniques

Although CNN-based data reduction expedited computations, it compromised classifier performance , as reflected by the diminished precision , recall and F1 scores. While best k on cleaned data was indemnified as 24 with 79.37% accuracy , post-PCA accuracy saw a minor increase.

8 Feature Selection on Reduced Data

Upon applying RFECV to the reduced data , 15 features were identified as optimal, with a corresponding F1 score of 74.79% , which was lower than the original dataset score.

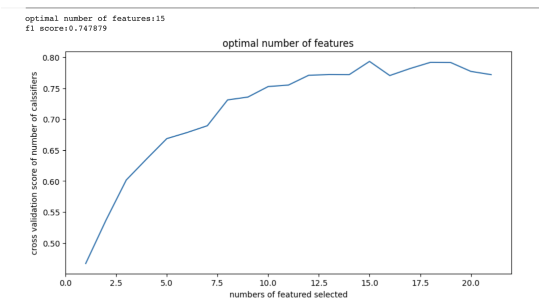

Figure 3: Feature selection on reduced data

9 Evaluation Metrics:

ROC and Precision-Recall curves for the cleaned data indicated AUC values below those derived from the original Dataset. This underscores the potential information loss sue to data cleaning with CNN.

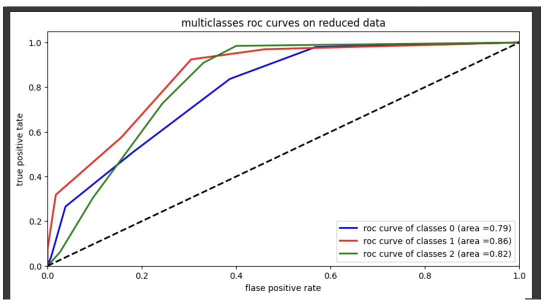

Figure 4: the ROC results after cleaning

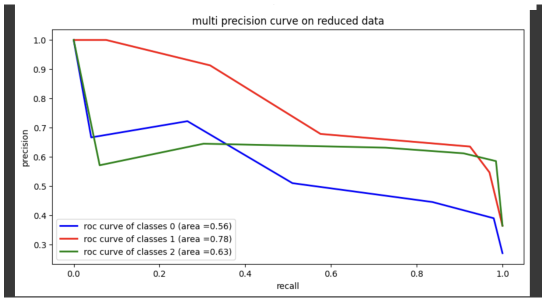

Figure 5: ROC and Precison curves after cleaning the data

10 Conclusion :

Our experiment elucidates the trade-offs involved in preprocessing for model simplification. While PCA and feature selection can enhance model interpretation, they may not always translate to improved performance, particularly when compared with the original Dataset. Future work may delve deeper into those preprocessing techniques to optimize the balance between model complexity, computational efficiency and accuracy. Thank you forgiving us the chance for this good experiment that helped us to understand a lot in the models performance.