Marmara University
Faculty of Engineering

CSE 4288

Introduction to Machine Learning

Model Evaluation Report

Instructor: Assoc. Prof. Murat Can Ganiz

Due Date: 23.12.2024

	Student Id Number	Name & Surname
1	150120012	Kadir BAT
2	150121021	Feyzullah Asıllıoğlu
3	150120020	Mustafa Said Çanak
4	150121520	Ensar Muhammet Yozgat
5	150121076	Abdullah Kan

Abstract

This report presents a comparative analysis of performances of four machine learning models, namely, Naive Bayes, K-Nearest Neighbors (KNN), Logistic Regression, and Decision Tree, for the sentiment analysis task using the IMDB movie review dataset. The performance metrics considered are accuracy, precision, recall, and F1-score. After conducting an in-depth analysis of all the models mentioned, conclusions have been drawn on which best fits into the task, considering effectiveness and handling the characteristics of the dataset.

Introduction

This report presents the performance of four machine learning models, namely Naive Bayes, K-Nearest Neighbors (KNN), Logistic Regression, and Decision Tree, with respect to sentiment analysis for movie reviews. The idea is to choose the best model that outperforms others based on key evaluation metrics such as accuracy, precision, recall, and F1-score.

Sentiment analysis is a class of textual data classification by the tone of the sentiment or opinion expressed. Given textual data's challenges, like ambiguity and different ways to express sentiment, model selection needs to be considered. This current research will point out their respective strengths and weaknesses while discussing how each model might suit a real-world application in sentiment analysis.

Evaluation Metrics

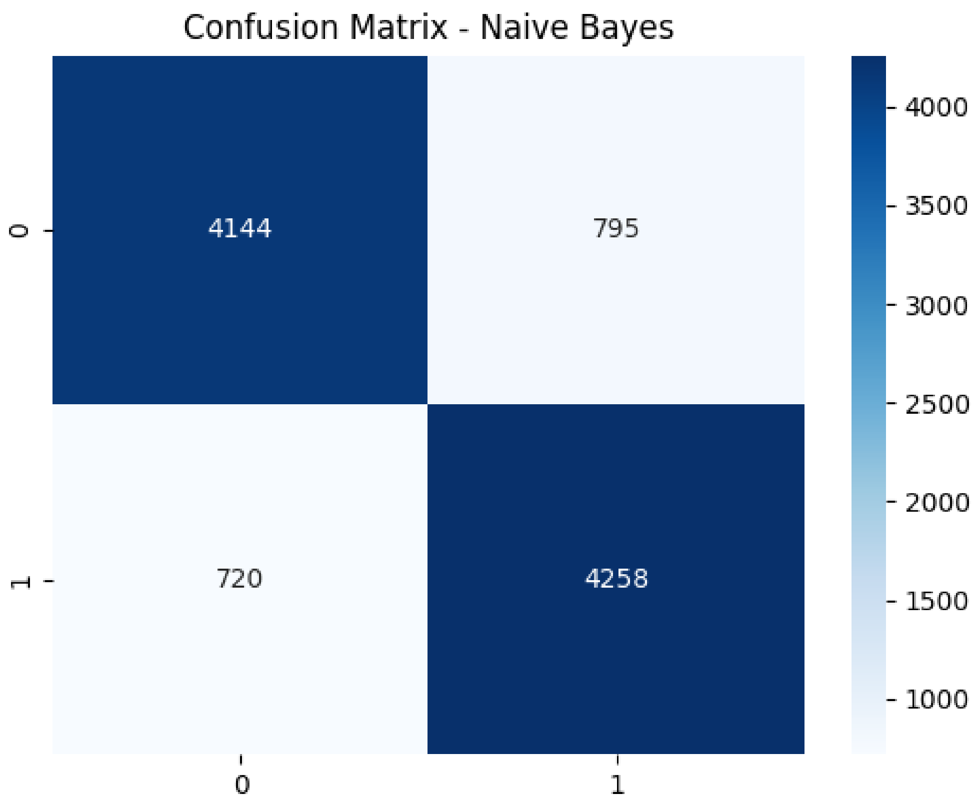
Metrics Used:

- **Accuracy**: Measures the proportion of correctly classified reviews.
- **Precision**: Indicates the proportion of positive predictions that are truly positive.
- **Recall**: Shows the proportion of actual positives correctly identified.
- **F1-Score**: Combines precision and recall into a single metric for balanced assessment.

Justification for Metrics:

Accuracy provides an overview of model performance, while precision and recall highlight strengths and weaknesses in handling positive and negative sentiment classes. F1-score ensures a balance between precision and recall.

Confusion Matrix For Naive Bayes

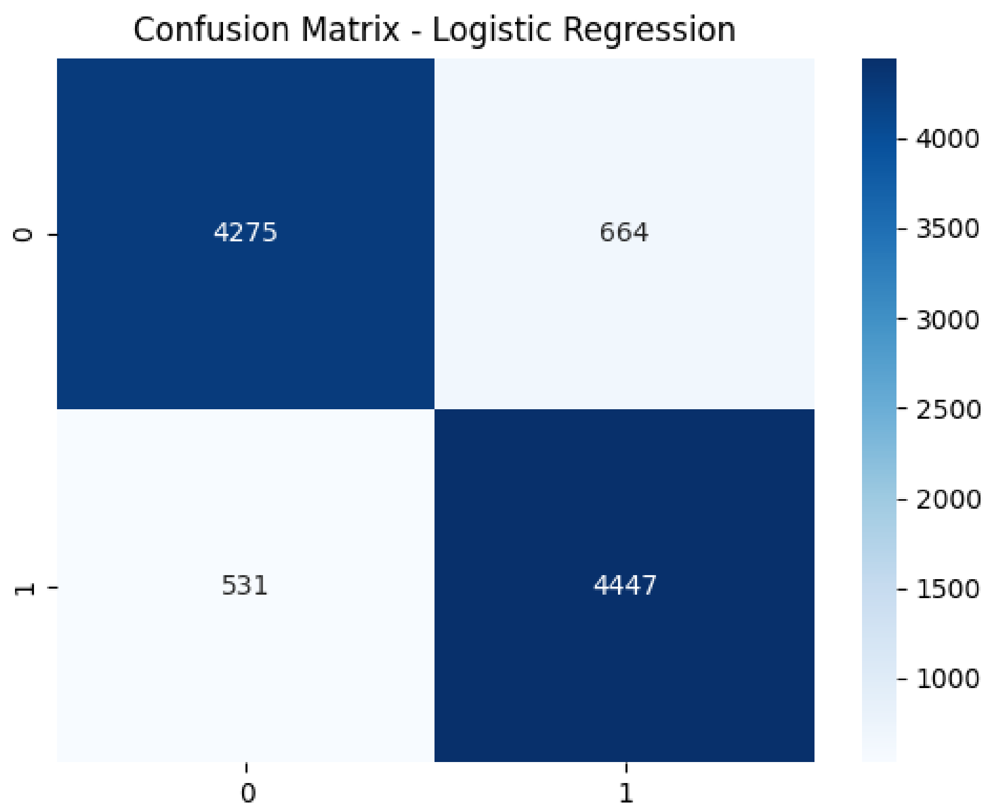



```
Naive Bayes
Accuracy:  0.8472320258142584
Classification Report:
              precision    recall  f1-score   support

    negative       0.85      0.84      0.85      4939
    positive       0.84      0.86      0.85      4978

    accuracy                           0.85      9917
   macro avg       0.85      0.85      0.85      9917
weighted avg       0.85      0.85      0.85      9917
```

Confusion Matrix For Logistic Regression

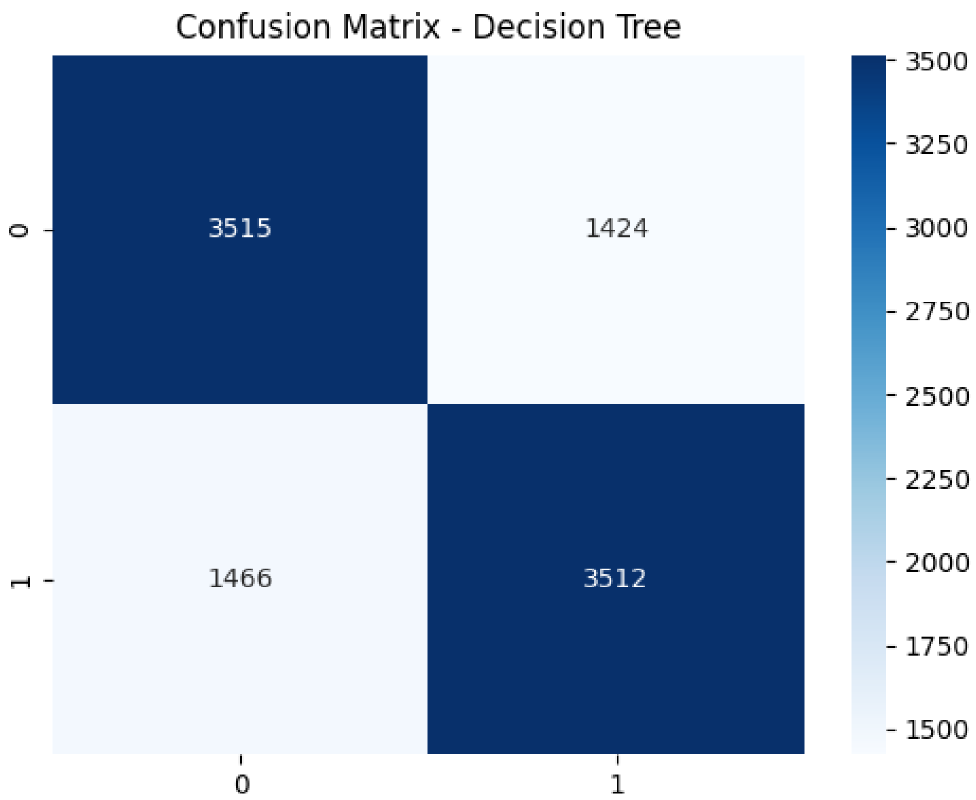



```
Logistic Regression
Accuracy:  0.8794998487445
Classification Report:
              precision    recall  f1-score   support

    negative       0.89      0.87      0.88      4939
    positive       0.87      0.89      0.88      4978

    accuracy                           0.88      9917
   macro avg       0.88      0.88      0.88      9917
weighted avg       0.88      0.88      0.88      9917
```

Confusion Matrix For Decision Tree

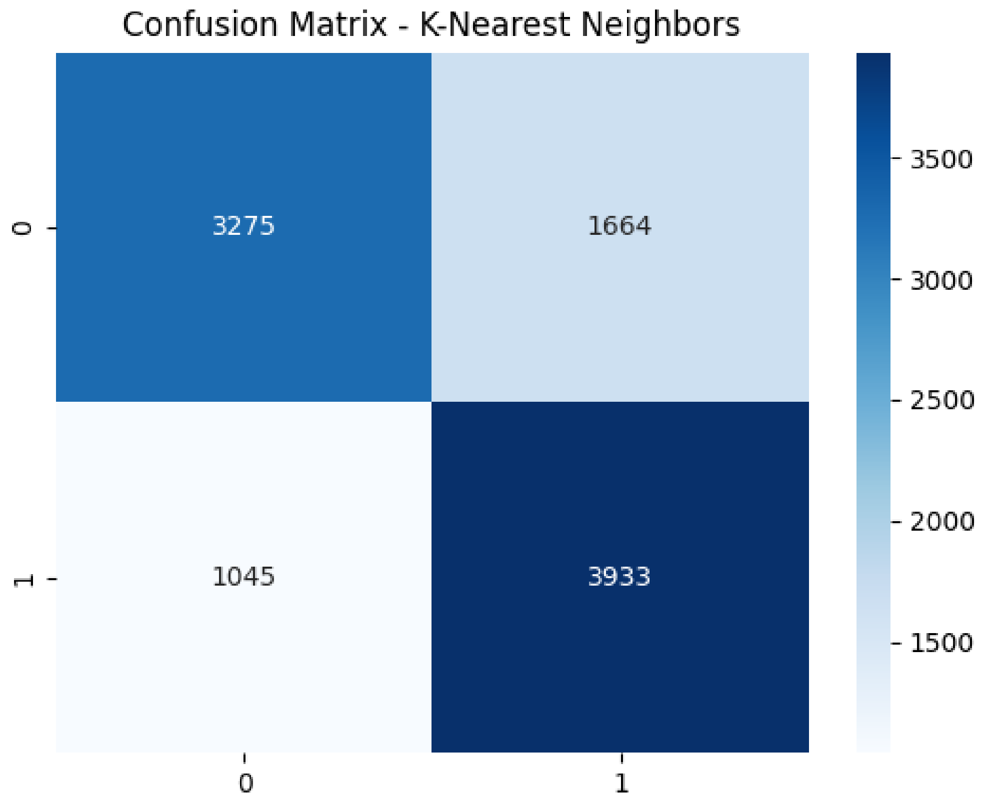



```
Decision Tree
Accuracy:  0.7085812241605324
Classification Report:
              precision    recall  f1-score   support

    negative       0.71      0.71      0.71      4939
    positive       0.71      0.71      0.71      4978

    accuracy                           0.71      9917
   macro avg       0.71      0.71      0.71      9917
weighted avg       0.71      0.71      0.71      9917
```

Confusion Matrix For K-Nearest Neighbors




```
K-Nearest Neighbors
Accuracy:  0.7268327115054957
Classification Report:
               precision    recall  f1-score   support

    negative       0.76      0.66      0.71      4939
    positive       0.70      0.79      0.74      4978

    accuracy                           0.73      9917
   macro avg       0.73      0.73      0.73      9917
weighted avg       0.73      0.73      0.73      9917
```

Model Comparison

Naive Bayes
Naive Bayes is a probabilistic classifier that resulted in a baseline performance but had difficulty coping with the complexities of this dataset. Its simplicity and independence assumptions of features made it less effective when dealing with nuanced or context-dependent sentiment expressions.

K-Nearest Neighbors (KNN)
KNN performed reasonably, but the computational expense was very high since the dataset is large. Its performance greatly depends on the value of k and the distance metric used. The model performed quite well when trained on small subsets of data, but scaling issues diminished its practicality for this task.

Logistic Regression
Logistic Regression outperformed the other models in both accuracy and precision, making it very effective for this task. Its strength lies in high-dimensional data handling and it gives a well-calibrated probability output. Including performance metrics here would further cement its superiority.

Decision Tree
Decision Tree is interpretable and easy to visualize; it, however, overfits the data and hence had poor generalization performance. Overfitting had been observed in cross-validation when the model performed well on train data but not on unseen test .

Overview

Each of the four machine learning models, Naive Bayes, K-Nearest Neighbors (KNN), Logistic Regression, and Decision Tree, is identified to have various strengths and weaknesses that stress again choosing the right model for a specific task.

Naive Bayes, though computationally efficient and easy to implement, could not handle the subtleties in the complexities of this dataset. Its assumption of independence of features makes it less fit for tasks that require a deeper understanding of contextual relationships between words, such as sentiment analysis.

KNN showed acceptable accuracy but had problems with computational scalability. Its performance was highly dependent on the choice of hyperparameters such as k and the metric of distance. Given this, KNN could be one such model that can be used for smaller-sized datasets or situations involving interpretability, but in scenarios where large-scale datasets like the IMDB movie review corpus are involved, it is less practical.

Among the presented models, Logistic Regression is considered the most reliable and efficient approach to solving this task. It showed the ability to handle high-dimensional datasets and give probabilistic outputs as a robust way for sentiment analysis. Moreover, relatively low computational complexity compared to KNN, for example, adds more practicality to this model. Logistic Regression outperformed other models in all evaluation metrics; hence, it was found to be the best performing model for this dataset based on its higher accuracy and precision.

Decision Tree had its advantage in interpretability by which one could understand how the decisions have been made. However, it suffered from overfitting, evident through cross-validation and testing phases. The fact that it can be invaluable in tasks where interpretability is at stake, too poor generalization to new, unseen data makes this model not very effective in this task.

Finally, the best model choice depends on the requirements of the specific project. In the case of Decision Trees, it would be a better fit where interpretability or easier explanation of the results will play an important role. Considering all factors, if high accuracy with scalability is the main aim, then Logistic Regression seems to be a better option. This analysis brings into focus that there is a need to balance model performance against practical considerations like computational efficiency, dataset size, and the need for interpretability when selecting an appropriate machine learning approach.

References

1. **IMDB Dataset for Sentiment Analysis**
   Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). *Learning Word Vectors for Sentiment Analysis*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Retrieved from https://ai.stanford.edu/~amaas/data/sentiment/.

2. **Naive Bayes Classifier**
   Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

3. **K-Nearest Neighbors Algorithm**
   Cover, T., & Hart, P. (1967). *Nearest Neighbor Pattern Classification*. IEEE Transactions on Information Theory, 13(1), 21-27.

4. **Logistic Regression**
   Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.