Introduction

In today's rapidly changing landscape of vehicle technology, sensor data is playing a significant role that improves road safety and the driving efficiency. The current research looks into the possibility of using machine learning to forecast driving styles based on vehicle sensor readings based off the datasets obtained at [4]. The research will use three different classifiers - Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (kNN) - to determine the most efficient model in distinguishing different driving behaviors. Our methodology includes the intensive data cleaning, preprocessing and exploration of the data to confirm its quality and usefulness. Through this activity, we expect to give meaningful contribution into the development of smart driver assistance systems, one of the main purposes of predictive analytics which is the creation of a safer driving environment.

Data Processing

Data Cleaning

The dataset necessitated extensive preprocessing to rectify issues commonly associated with real-world data, including missing values and inconsistent formatting [8]. Numerical missing data were filled with the mean of their respective features, ensuring a consistent dataset without gaps that could skew analysis. Categorical data were similarly treated with the mode of their distribution, thus preserving the integrity and representativeness of the original dataset [3].

Data Analysis and Visulization

A cornerstone of the preprocessing stage was Exploratory Data Analysis (EDA), which utilized a suite of visualization tools. Boxplots provided a preliminary inspection for outliers, revealing data points that deviated significantly from the overall distribution, which could potentially influence the performance of the predictive models [6] [7]. Histograms were instrumental in understanding the distribution of each feature, offering insights into the data's skewness and kurtosis (Figure 4). The correlation heatmap served as a pivotal tool, revealing the degree of linear relationship between pairs of features (Figure 2). This was particularly important for identifying features with high multicollinearity, which could be candidates for removal or combination to improve model performance [6].

Model Training and Evaluation

Logistic Regression

The Logistic Regression model was appreciated as a base model because it is not only a simple model but also it demonstrates the effectiveness in classification problems which involve the selection from the binary alternatives. The model was suitably subjected to severe parameter optimization in order to avert the overfitting menace and give the model the ability to learn from any out of sight data.

Support Vector Machine (SVM)

The SVM was used specifically for this reason, because it could handle the high dimensionality belonging to the dataset. By modifying carefully the hyper-parameters for imputation of the class distribution the SVM features could illustrate that SVM can be applied for complex classification challenges [1].

K-Nearest Neighbours (kNN)

The advantage of the kNN algorithm compared to the other options is its ease of understanding, without which it would not be possible to make meaningful predictions. The final determination of an optimal number of neighbors was achieved through conducting various trials trying not to overfit and not to have local structure of data having a too much influence on the algorithm [2].

Cross-Validation and Parameter Tuning

We subsequently decided cross validation was indispensably not only to asses the models' robustness and accuracy, but also evaluate its reliability across different data segments. Apart from that, it was used to optimize the hyperparameters and get an idea about how the models performed so that the models did not overfit and the results remained valid in that case [5].

Model Evaluation

The models' performance was quantitatively assessed using accuracy, precision, recall, and F1-scores. Additionally, ROC curves and corresponding AUC metrics provided a visual and statistical representation of each model's ability to classify driving styles effectively. The ROC curves, in particular, demonstrated the trade-off between sensitivity and specificity, with the kNN model displaying a commendable AUC of 0.93, indicating its superior performance in distinguishing between the two driving styles (Figure 10) [3].

Results

Evaluating machine learning models unveiled distinct predictive strengths. The kNN model excelled, surpassing Logistic Regression and SVM in accuracy and F1-scores (Figures 6, 7, and 8), with its higher AUC demonstrating better classification effectiveness (Figure 10). Logistic Regression and SVM showed room for improvement, as their lower AUCs pointed to potential gains from refined class balance and feature scaling (Figure 7).

The Distribution of Driving Styles (Figure 4) indicated an imbalance in the dataset, necessitating adjustments in model training, particularly for SVM. The kNN model, however, showed robustness against such imbalances.

Feature interrelationships, as visualized in the Correlation Heatmap (Figure 1) and scatter plot matrix (Figure 2), highlighted multicollinearity, prompting a more selective feature utilization strategy to enhance model interpretability.

Boxplots and histograms (Figures 3 and 5) underscored the varying scales of data, advocating for standardization to maintain feature parity in model influence.

This concise evaluation underlines the importance of comprehensive model assessment—balancing statistical evaluation with visual analysis to refine current models and guide future improvements in driving behavior prediction.

Figures

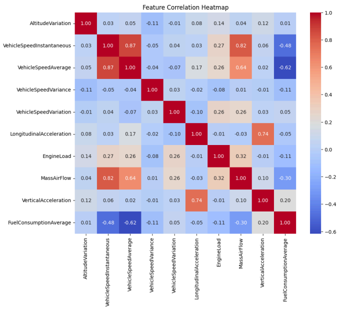


Figure 1

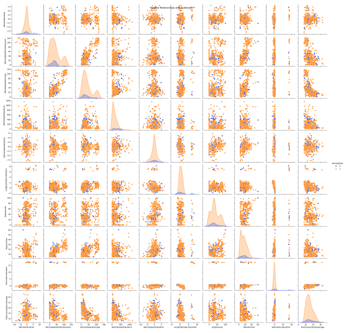


Figure 2

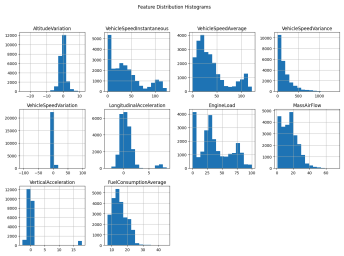


Figure 3

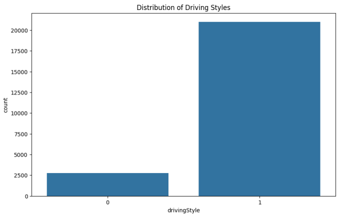


Figure 4

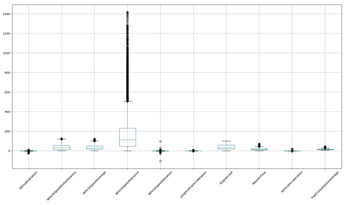
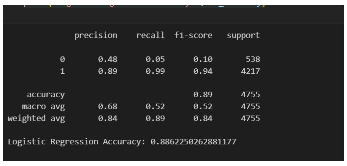


Figure 5



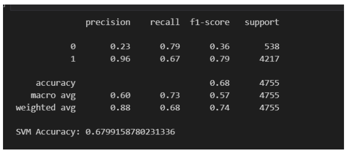
```
              precision    recall  f1-score   support

           0       0.48      0.05      0.10       538
           1       0.89      0.99      0.94      4217

    accuracy                           0.89      4755
   macro avg       0.68      0.52      0.52      4755
weighted avg       0.84      0.89      0.84      4755

Logistic Regression Accuracy: 0.8862250262881177
```

Figure 6



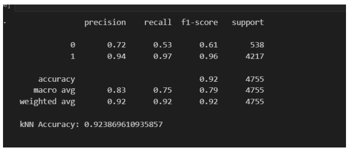
```
              precision    recall  f1-score   support

           0       0.23      0.79      0.36       538
           1       0.96      0.67      0.79      4217

    accuracy                           0.68      4755
   macro avg       0.60      0.73      0.57      4755
weighted avg       0.88      0.68      0.74      4755

SVM Accuracy: 0.6799158780231336
```

Figure 7



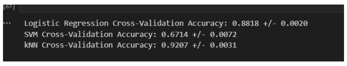
```
              precision    recall  f1-score   support

           0       0.72      0.53      0.61       538
           1       0.94      0.97      0.96      4217

    accuracy                           0.92      4755
   macro avg       0.83      0.75      0.79      4755
weighted avg       0.92      0.92      0.92      4755

kNN Accuracy: 0.923869610935857
```

Figure 8



```
[67]
...  Logistic Regression Cross-Validation Accuracy: 0.8818 +/- 0.0020
     SVM Cross-Validation Accuracy: 0.6714 +/- 0.0072
     kNN Cross-Validation Accuracy: 0.9207 +/- 0.0031
```

Figure 9

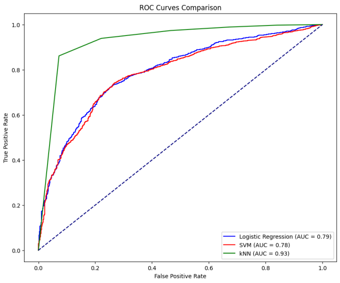


Figure 10

Conclusion

This research has illuminated the critical role of machine learning in interpreting vehicle sensor data to predict driving styles. Through comparative analysis, the kNN model emerged as the most proficient, showcasing superior accuracy and robustness against imbalanced datasets. Notably, the exploration underscored the significance of meticulous data preprocessing and feature selection in optimizing model performance. Our findings advocate for the integration of more diverse features and the exploration of ensemble and deep learning techniques to further refine predictive accuracy. Reflecting on this journey, it becomes clear that the fusion of data science and vehicle technology holds immense promise for revolutionizing driving safety and efficiency. Future research should extend beyond the current dataset to include contextual factors, offering a more holistic understanding of driving behavior and its implications for driver-assistance technologies.

References

[1] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[2] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21-27, 1967.

[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[4] Gloseto, "Traffic Driving Style Road Surface Condition," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/gloseto/traffic-driving-style-road-surface-condition

[5] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation," Machine Learning Mastery, Aug. 9, 2018. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/

[6] Matplotlib: Visualization with Python. [Online]. Available: https://matplotlib.org/

[7] Seaborn: statistical data visualization. [Online]. Available: https://seaborn.pydata.org/

[8] Pandas: Python Data Analysis Library. [Online]. Available: https://pandas.pydata.org/