Driver behavior profiles for road safety analysis

Arushi Biswas[1], Deshmukh Aourva Deepak[2]

***Abstract*—This project investigates driver behavior profiling for road safety analysis using accelerometer and gyrometer data. Through visualization-based exploratory data analysis, data cleansing, and advanced processing techniques, insights into the dataset's characteristics and relationships are gained. Future work includes model development and deployment for real-world road safety applications.**

I. INTRODUCTION

Driver behavior profiling plays a crucial role in road safety analysis, aiding in the identification of risky driving patterns and the development of proactive safety measures. This report delves into the analysis of accelerometer and gyrometer data to understand driver behavior. Through visualization-based exploratory data analysis, data cleansing, and advanced processing techniques, insights into the dataset's characteristics and relationships are derived. The project aims to develop predictive models for driver behavior classification, contributing to enhanced road safety initiatives. With a focus on data-driven approaches, this report sets the stage for leveraging advanced analytics to address critical road safety challenges.

II. PROBLEM STATEMENT

To develop a comprehensive analytical framework for profiling driver behavior using accelerometer and gyrometer data, with the overarching objective of enhancing road safety measures.

III. OBJECTIVES

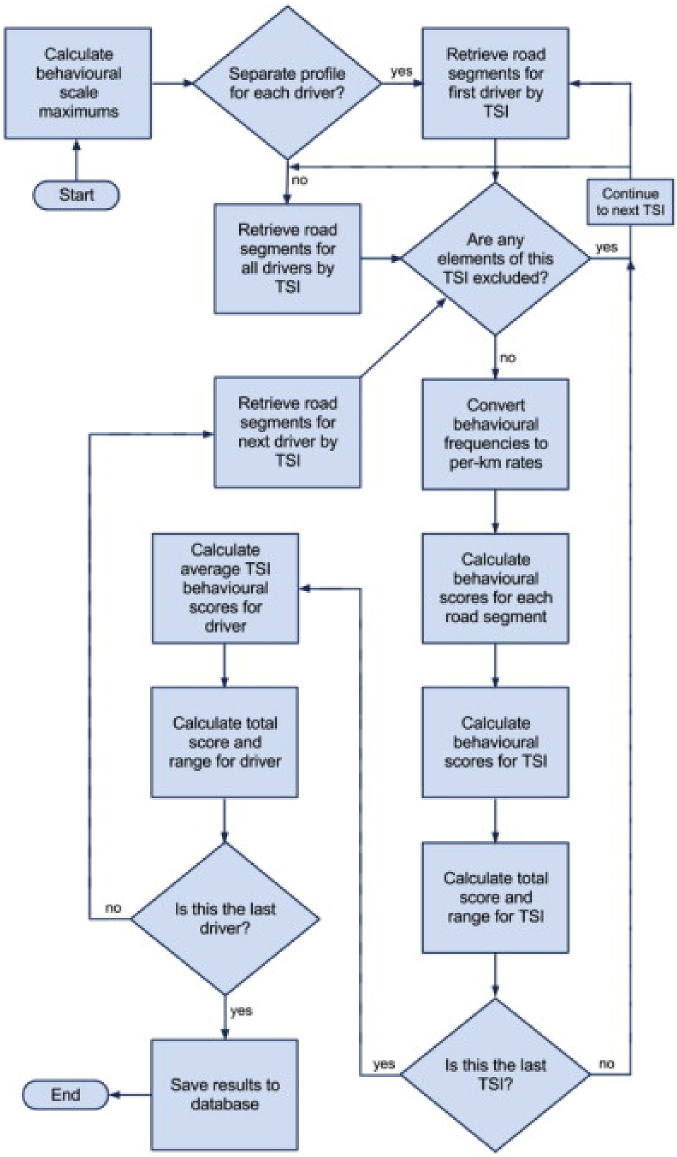
The principal objective of this project is to employ data science methodologies to analyze and model driver behavior based on sensor data captured from vehicles. By delving into driver behavior patterns, the aim is to discern potential risk factors and formulate predictive models to bolster road safety measures.

1. Analyze and preprocess accelerometer and gyrometer data to extract meaningful features and prepare the dataset for modeling.

2. Develop and train predictive models capable of classifying driver behavior based on sensor data, with a focus on identifying potential risk factors for traffic accidents.

3. Evaluate the performance of the developed models and utilize them to inform policy decisions, improve driver training programs, and contribute to the development of intelligent transportation systems aimed at enhancing overall road safety.

IV. EXISTING METHODOLOGY


Calculate behavioural scale maximums
Start
Separate profile for each driver?
yes
no
Retrieve road segments for first driver by TSI
Continue to next TSI
Retrieve road segments for all drivers by TSI
Are any elements of this TSI excluded?
yes
no
Retrieve road segments for next driver by TSI
Convert behavioural frequencies to per-km rates
Calculate average TSI behavioural scores for driver
Calculate behavioural scores for each road segment
Calculate total score and range for driver
Calculate behavioural scores for TSI
no
Is this the last driver?
yes
Calculate total score and range for TSI
End
Save results to database
yes
Is this the last TSI?
no

V. PROPOSED ENHANCEMENTS

1.Feature Engineering: Explore additional features derived from accelerometer and gyrometer data, such as velocity, jerk, or frequency domain features, to capture more nuanced aspects of driver behavior and improve the predictive performance of the models.

2.Hyperparameter Tuning: Perform systematic hyperparameter tuning for the predictive models using techniques such as grid search or random search to optimize model performance and fine-tune model parameters for better results.

VI. DATASET DESCRIPTION

The dataset provided encompasses accelerometer and gyrometer readings obtained from drivers during diverse driving scenarios. Each data point is associated with a target variable indicative of the driver behavior classification. The features encapsulate GyroX, GyroY, GyroZ, AccX, AccY, and AccZ, representing measurements of angular velocity and acceleration along distinct axes.

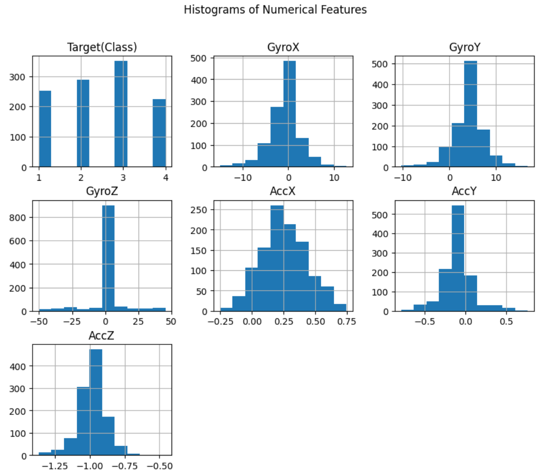
VII. APPROACH

This project has adhered to a structured workflow, comprising exploratory data analysis (EDA), data preprocessing, advanced processing, and modeling phases:

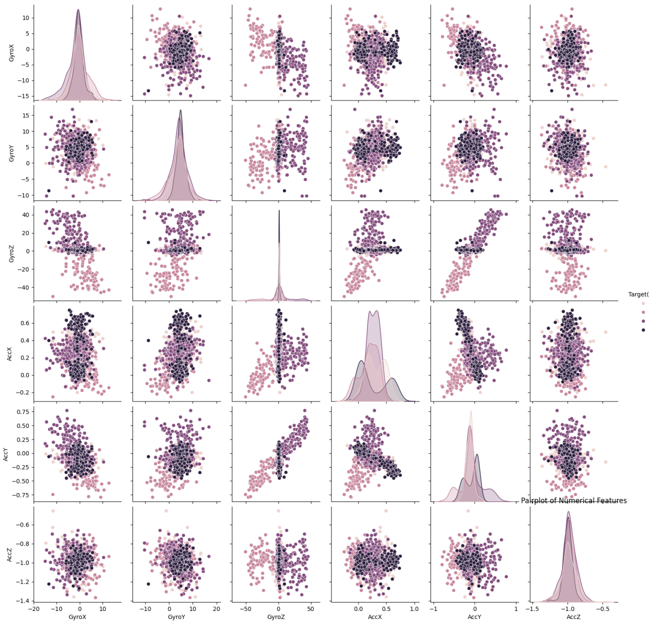
A. *EXPLORATORY DATA ANALYSIS (EDA)*

The initial stage entailed conducting visual and statistical analyses to glean insights into the distribution, variability, and correlations within the dataset. Scrutinizing the relationships between sensor readings and target behavior classifications furnished preliminary comprehension and steer subsequent preprocessing endeavors.

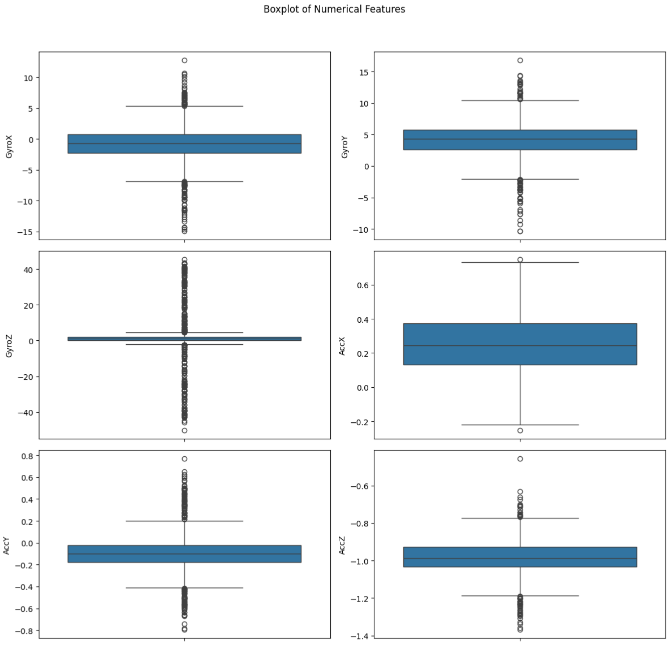
- Histograms: Histograms provide a graphical representation of the distribution of a single numerical variable. They can help identify the shape, central tendency, and spread of the data.



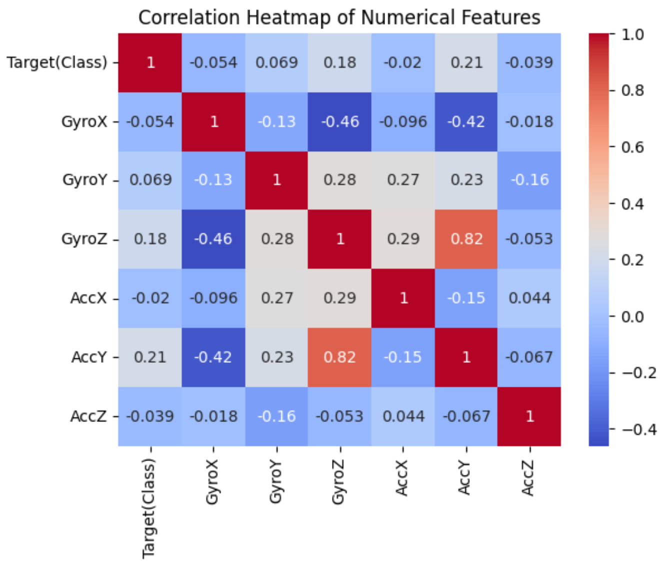
- Pairplot of Numerical Features:
  The pairplot displays scatter plots of all numerical features against each other, with separate colors representing different target classes. By examining the scatter plots, we can observe the relationships and distributions between the features and explore potential patterns or clusters within the data. For example, we can look for clusters of data points corresponding to different driver behavior classes and assess the separability of these clusters.



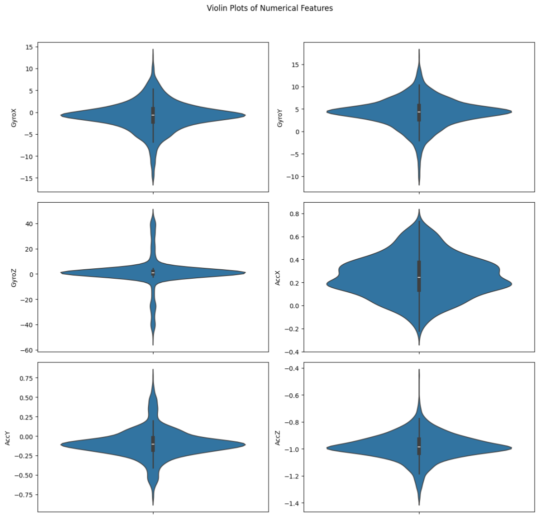
- Boxplot of Numerical Features by Target Class:
  The boxplots illustrate the distributions of each numerical feature grouped by target class.We have examined whether there are significant variations in accelerometer and gyrometer readings (GyroX, GyroY, GyroZ, AccX, AccY, AccZ) between different driver behavior classes.



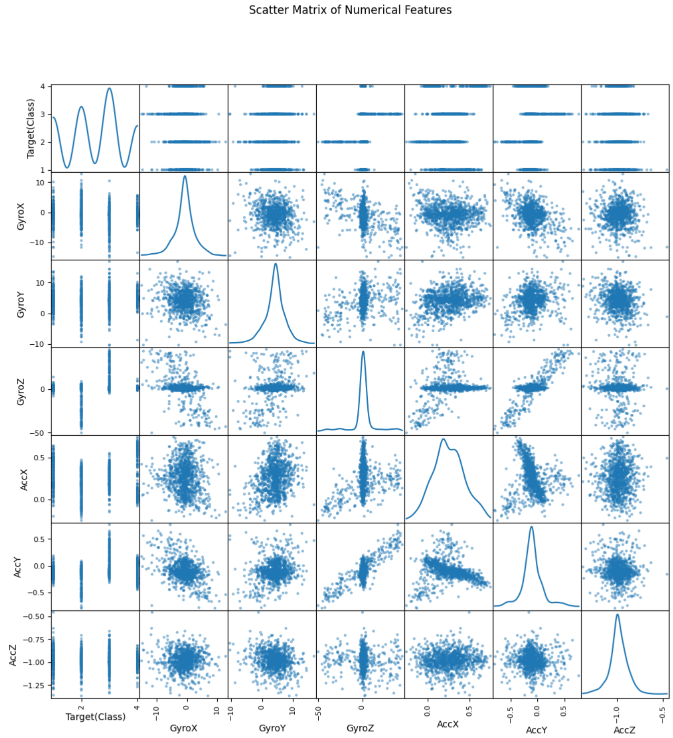
- Correlation Heatmap A correlation heatmap displays the pairwise correlation coefficients between numerical variables in the dataset. It has helped identify relationships and dependencies between variables, which can guide feature selection and modeling decisions.



- Violin Plots: Violin plots are similar to boxplots but also display the probability density of the data at different values. They provided insights into the distribution and spread of the data, similar to boxplots, but with additional information about the density.



- Scatter Matrix: A scatter matrix displays scatter plots of all pairwise combinations of numerical variables in the dataset. It allowed for a quick visual inspection of relationships and potential correlations between variables.



- Observations
  1.Distribution of Numerical Features:
  The histograms and violin plots illustrate the distributions of accelerometer and gyrometer readings (GyroX, GyroY, GyroZ, AccX, AccY, AccZ). From these visualizations, we observe the shapes of the distributions, including any skewness, peaks, or outliers present in the data.
  2.Relationships Between Features:
  The pairplot and scatter matrix reveal the relationships between pairs of numerical features. We can observe patterns, trends, or clusters in the scatter plots, indicating potential correlations or dependencies between variables.
  3.Correlation Between Features:
  The correlation heatmap displays the pairwise correlation coefficients between numerical features. We identify the strength and direction of correlations between features, helping to identify potentially redundant or highly correlated variables.
  Overall, the visual explorations conducted during EDA provided valuable insights into the dataset's structure, distributions, relationships between variables, and potential patterns or trends present in the data. These conclusions served as a foundation for subsequent data preprocessing, feature engineering, and model development stages in the driver behavior profiling project.

B. DATA CLEANSING AND PREPROCESSING

After performing data cleansing and preprocessing on the dataset, several results were achieved.

- No Missing Values:

There were no missing values in our dataset. If there were missing values in the dataset, they would have been dropped during the data cleansing step. The message "No missing values to be dropped." would be printed, indicating that the dataset is complete and ready for preprocessing.

- Feature Scaling:
  The numerical features (GyroX, GyroY, GyroZ, AccX, AccY, AccZ) have been scaled using StandardScaler to standardize their values. This means that each feature now has a mean of 0 and a standard deviation of 1, ensuring that they are on a similar scale. This preprocessing step is crucial for many machine learning algorithms, especially those based on distance metrics or gradient descent optimization.
- Splitting the Dataset:
  The dataset has been split into features (X) and the target variable (y). The features (X) contain all columns except the target variable 'Target', while the target variable (y) contains only the 'Target' column. This separation is necessary for supervised learning tasks, where the goal is to predict the target variable based on the features.
- Training and Testing Sets:
  The dataset has been further split into training and testing sets. By default, 80 percent of the data is allocated to the training set and 20 percent to the testing set. This division allows for the evaluation of the model's performance on unseen data, helping to assess its generalization ability.
- Shapes of Training and Testing Sets:
  The shapes of the training and testing sets are printed, indicating the number of samples and features in each set. This information helps verify that the data splitting was performed correctly and provides insights into the dataset's size.
- Observation Overall, after data cleansing and preprocessing, the dataset is in a suitable format for further analysis and modeling. The numerical features have been scaled, missing values have been handled, and the dataset has been split into training and testing sets for machine learning model development and evaluation.

C. ADVANCED PROCESSING

After performing advanced processing on the preprocessed dataset, several results are obtained:

- Feature Engineering:
  New features, namely 'GyroMagnitude' and 'AccMagnitude', have been engineered by calculating the magnitude of the gyroscope and accelerometer readings, respectively. These new features capture additional information about the magnitude of motion sensed by the sensors, which may be relevant for understanding driver behavior patterns.
- Dimensionality Reduction (PCA):
  Principal Component Analysis (PCA) has been applied to reduce the dimensionality of the dataset. The original six-dimensional feature space (GyroX, GyroY, GyroZ, AccX, AccY, AccZ) has been projected onto a two-dimensional space defined by the first two principal components ('PC1' and 'PC2'). Dimensionality reduction facilitated visualization and helped identify dominant patterns or trends in the data.
- Correlation Matrix:
  The correlation matrix shows the pairwise correlations between the numerical features in the dataset. Each cell in the matrix represents the correlation coefficient between two features, ranging from -1 to 1. Positive values indicate a positive correlation (as one feature increases, the other also tends to increase), while negative values indicate a negative correlation (as one feature increases, the other tends to decrease). A correlation coefficient close to zero suggests little to no linear relationship between the features. The correlation matrix helps identify potential dependencies or redundancies between features, guiding feature selection and model interpretation.
- Covariance Matrix:
  The covariance matrix provides information about the covariance between pairs of numerical features. Covariance measures the extent to which two variables change together. Positive values indicate that the variables tend to increase or decrease together, while negative values indicate an inverse relationship. The magnitude of the covariance reflects the strength of the relationship between variables, but it is not standardized like the correlation coefficient. The covariance matrix complements the correlation matrix by providing insights into the variability and joint distribution of the features.

VIII. RESULT ANALYSIS

- The project has provided a comprehensive understanding of the dataset, including its distributions, relationships between variables, and key patterns or trends.
- Visualization-based EDA revealed insights into the dataset's characteristics, guiding subsequent preprocessing steps.
- Data cleansing and preprocessing ensured that the dataset was suitable for modeling, with missing values handled and features standardized.
- Advanced processing techniques, including feature engineering, dimensionality reduction, and correlation analysis, enhanced the dataset by introducing new features and reducing dimensionality while preserving important information.
- Overall, the project has equipped us with a well-prepared dataset and valuable insights to build predictive models for driver behavior profiling, contributing to road safety analysis and decision-making.

IX. CONCLUSION

The project has successfully undertaken a comprehensive analysis of driver behavior data for road safety analysis.

Through visualization-based exploratory data analysis (EDA), data cleansing and preprocessing, and advanced processing techniques, valuable insights have been gained into the dataset's characteristics, relationships between variables, and potential patterns. The key findings and accomplishments of the project include:

- Understanding the Dataset: Visualization-based EDA provided a deep understanding of the dataset's distributions, correlations, and trends, laying the foundation for subsequent analysis.
- Data Cleansing and Preprocessing: Missing values were handled, and numerical features were standardized through feature scaling, ensuring the dataset's readiness for modeling tasks.
- Advanced Processing: Feature engineering introduced new features capturing the magnitude of gyroscope and accelerometer readings, while dimensionality reduction using PCA reduced the dataset's dimensionality while preserving its variance.

X. FUTURE WORK

While the project has achieved significant milestones, there are several avenues for future exploration and enhancement:

- Model Development: Build predictive models to classify and predict driver behavior based on the preprocessed dataset. Explore various machine learning algorithms such as classification algorithms (e.g., logistic regression, random forests) and deep learning models (e.g., neural networks) to achieve optimal performance.
- Feature Selection: Conduct further feature selection and analysis to identify the most informative features for modeling. Techniques such as recursive feature elimination, feature importance ranking, and domain knowledge incorporation can aid in selecting the most relevant features.
- Deployment and Integration: Once developed and validated, deploy the models into real-world applications for road safety analysis. Integrate the models with existing traffic management systems or driver assistance technologies to enhance road safety measures and provide actionable insights to stakeholders.

XI. REFERENCES

- Accident Analysis and Prevention Volume 76, March 2015, Pages 118-132
- Design of the In-Vehicle Driving Behavior and Crash Risk Study Transportation Research Board, Washington, D.C (2011)
- Development of a method for detecting jerks in safety critical events Accid. Anal. Prev., 50 (2013), pp. 83-91, 10.1016/j.aap.2012.03.032

Team15-dsreport.pdf

ORIGINALITY REPORT

13%	3%	3%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Florida International University Student Paper	1%
2	remittancesreview.com Internet Source	1%
3	Submitted to Berlin School of Business and Innovation Student Paper	1%
4	Submitted to BPP College of Professional Studies Limited Student Paper	1%
5	Submitted to University of Hertfordshire Student Paper	1%
6	Submitted to University of Sunderland Student Paper	1%
7	Submitted to CSU, Long Beach Student Paper	1%
8	Submitted to Macquarie University Student Paper	1%
9	Submitted to National College of Ireland	