אוניברסיטת
אריאל
בשומרון

Supervised and Unsupervised Road Risk Analysis

Lecturer: Chen Hajaj

Team Members: Nofar Mahrabi & Rotem Barel

https://github.com/RotemBarel1234/Road-_Risk-_Prediction-Clustering.git

Introduction

Road accidents continue to be a major concern globally, causing countless injuries and fatalities despite significant advancements in vehicle safety and road infrastructure. The loss of life and severe injuries resulting from these incidents highlight the urgent need for better tools to assess and mitigate accident risks. Understanding the factors that contribute to accidents and accurately predicting risk levels can lead to effective interventions that save lives and reduce harm on the roads.

The purpose of this study is to analyze various characteristics related to road accidents, including vehicle type, vehicle dimensions, accident density per square kilometer, and fatality counts across different age groups, in order to predict accident risk levels. By leveraging these features, the study aims to uncover patterns that can help identify high-risk situations and locations. The ultimate goal is to provide actionable insights that can lead to the reduction of road accidents and, most importantly, the preservation of human life. Additionally, the study explores anomalies in accident data to gain a deeper understanding of rare but potentially catastrophic scenarios, thereby offering new avenues for preventive strategies and public safety measures.

If we can successfully predict risk levels and identify high-risk factors, these insights could assist traffic authorities, urban planners, and policymakers in making data-driven decisions to improve road infrastructure, implement better safety regulations, and ultimately save lives.

Dataset and Features

We used a simple API call to read data from data.gov.il, which is updated at irregular intervals based on the availability of new accident reports. The dataset consists of 2,637 rows and 29 columns, providing detailed information on various characteristics related to road accidents.

1. **CITY** – Name of the municipality or locality where the accident occurred.
2. **CITYCODE** – Code representing the municipality or locality.
3. **ACC_INDEX** – Accident density per square kilometer, calculated as the ratio between the number of accidents within a polygon and its area in square kilometers.
4. **USETYPE** – Classification of the area based on primary land use.
5. **SUMACCIDEN** – Total number of accidents, including all types, during the defined time period within the polygon.
6. **INJTOTAL** – Total number of casualties (fatalities and injuries) in the accidents.
7. **DEAD** – Number of fatalities in the accidents.
8. **SEVER_INJ** – Number of people severely injured in the accidents.
9. **SLIGH_INJ** – Number of people with minor injuries in the accidents.
10. **INJ0_19** – Number of casualties (fatalities and injuries) aged 0-19.
11. **INJ20_64** – Number of casualties (fatalities and injuries) aged 20-64.
12. **INJ65_** – Number of casualties (fatalities and injuries) aged 65 and above.
13. **PEDESTRINJ** – Number of injured pedestrians involved in the accidents.
14. **TOTDRIVERS** – Total number of drivers involved in the accidents.
15. **VEHICLE** – Total number of vehicles involved in the accidents.
16. **BICYCLE** – Number of bicycles (including electric bicycles) involved in the accidents.
17. **MOTORCYCLE** – Number of motorcycles involved in the accidents.
18. **PRIVATE** – Number of private vehicles involved in the accidents.
19. **TRUCK** – Number of trucks over 3.5 tons involved in the accidents.
20. **TAZAREA** – Area of the traffic zone in square meters.
21. **POP_2018** – Population of the traffic zone in 2018.
22. **Shape_Area** – Geometric area of the traffic zone, used for internal data handling.
23. **Shape_Length** – Geometric perimeter length of the traffic zone, used for internal data handling.
24. **ID** – Internal ID, not intended for display.
25. **OBJECTID** – Internal object ID, not intended for display.
26. **OID** – Object identifier, not intended for display.
27. **USETYPECOD** – Internal code for area usage type, not intended for display.
28. **MAINUSE** – Primary land use type in the traffic zone.
29. **YEARMONTH** – The year and month in which the accidents occurred, used for temporal analysis.

During the initial preprocessing stage, several columns were removed as they were deemed irrelevant for the analysis. The columns 'OID' and 'YEARMONTH' were removed because they contained single-valued entries throughout the dataset, providing no informational value or variability. Additionally, the columns '_id', 'CITYCODE', 'USETYPECOD', and 'OBJECTID' were excluded since they served solely as unique identifiers and did not offer any analytical significance. The column 'USETYPE' was removed as well because it provided redundant information, and the column 'MAINUSE' was retained instead, as it offered a more detailed and consistent description of land use.

After removing irrelevant columns, we checked for duplicate rows and confirmed that there were no missing values in the dataset, ensuring data integrity and avoiding redundant or incomplete information that could bias the analysis.

Once data consistency was confirmed, we defined our target variable, 'SUMACCIDEN', representing the total number of accidents in a given area. We chose to classify the target variable into three distinct risk levels—low, medium, and high—based on percentiles (percentile-based binning) to ensure a balanced distribution of classes.

We discovered that 'INJTOTAL' is a perfect sum of the columns 'DEAD', 'SEVER_INJ', and 'SLIGH_INJ', indicating that it provides no additional information. Therefore, we decided to remove it to avoid redundancy and prevent potential issues that could affect the model's accuracy.

Next, we performed outlier analysis by examining the summary statistics of the dataset using the `describe` function. We focused on columns that displayed a wide range of values or a significant difference between the median and the maximum, as such discrepancies often indicate the presence of potential outliers. These columns primarily included variables related to population size, area, vehicle involvement, and casualty counts
For each selected column, we used histograms and box plots to visually inspect the distribution and identify outliers. Additionally, we displayed the rows with outlier values to examine whether other attributes in these rows could explain the presence of the outliers.

We analyzed the 'POP_2018' column and identified several outliers with high population values. Further investigation using the 'MAINUSE' column showed that these outliers were mainly from highly populated residential areas. As they reflect real-world characteristics, we retained them in the dataset.

We analyzed the 'TAZAREA' column and identified an extremely high maximum value (884,527,200), which appeared unusual compared to the rest of the data. Since this value likely resulted from a data entry error, we removed outliers exceeding 10,000,000 to ensure the accuracy of the analysis.

We analyzed the relationship between 'SUMACCIDEN' and key features like 'DEAD', 'PRIVATE', and 'VEHICLE', observing a positive correlation in each case. Outliers were retained as they represent real-world extreme cases and are important for accurately modeling accident severity.

We analyzed the relationship between 'TOTDRIVERS' and both 'SLIGH_INJ' (slight injuries) and 'SEVER_INJ' (severe injuries). High values in the injury columns corresponded to high values in TOTDRIVERS, indicating that these outliers likely represent real cases involving a large number of drivers. Therefore, we decided to retain these outliers, as they reflect genuine trends in the data.

We analyzed the 'ACC_INDEX' column and observed that higher values were strongly associated with high-risk areas, particularly in residential and industrial zones. Although occasional high values were found in open areas, they likely indicate specific hazardous locations. Therefore, we decided to retain all extreme values, as they provide valuable insights for identifying high-risk zones.

After investigating the outliers, we used density plots to analyze the distribution of the columns and observed non-normal distributions with tied values. Therefore, we chose Kendall Tau correlation, as it is more robust for handling tied values, long-tailed distributions, and potential outliers.

After analyzing the correlations, we removed highly correlated columns to reduce redundancy. Specifically, we kept 'VEHICLE' and removed 'TOTDRIVERS' due to their strong correlation (0.93). Additionally, we retained 'TAZAREA' and removed 'Shape_Area' and 'Shape_Length' since they were perfectly correlated. Lastly, we considered the high correlation (0.88) between 'PRIVATE' and 'VEHICLE', but decided to retain both for further analysis.

In the feature selection stage, we first chose to encode the target variable 'RISK_LEVEL' by mapping its categories (Low → 0, Medium → 1, High → 2) into numerical values. We then split the dataset into a training set (80%) and a test set (20%) using stratified sampling to preserve the distribution of the target variable.

We performed a Chi-Square test for the 'MAINUSE' feature and found a significant association with the target variable 'RISK_LEVEL' (p-value < 0.05), indicating its importance for further analysis.

Additionally, we applied the ReliefF algorithm to the numerical features, which helped identify the most relevant predictors for modeling.

We checked the class distribution of 'RISK_LEVEL' and confirmed that it was relatively balanced, enabling us to use standard classification methods without needing special techniques for handling class imbalance. For preprocessing, we built a pipeline that included several key steps to ensure robust data transformation and handling.

First, we used RobustScaler for numerical features. This choice was driven by the fact that our data contained outliers and non-normal distributions. Unlike standard scaling methods such as MinMaxScaler or StandardScaler is less sensitive to outliers, as it scales the data based on the interquartile range (IQR), ensuring that extreme values do not overly influence the scaling process.

For categorical features, we applied one-hot encoding to transform them into a numerical format suitable for machine learning models. This method was chosen to allow the model to effectively process categorical data without imposing any ordinal relationship between categories.

Additionally, we planned for future data preprocessing scenarios by incorporating strategies for handling missing values within the pipeline. Specifically:

For numerical features, we imputed missing values using the mean. This approach ensures that the central tendency of the data is maintained without introducing bias, especially when the data is relatively symmetric or has a normal-like distribution.
For categorical features, we imputed missing values using the most frequent (mode) value. This method helps preserve the distribution of categorical data and prevents the creation of new, potentially irrelevant categories.

By including these imputation strategies within the pipeline, we ensured that any future data with missing values could be processed consistently and automatically, without requiring manual intervention. This comprehensive pipeline design enabled us to handle a variety of real-world data issues, ensuring both reliability and scalability in our machine learning workflow.

Methodology

Supervised Learning: Classification Models

Since the project involved predicting the risk level of road accidents, a multi-class classification problem was formulated. The following classification algorithms were selected based on the characteristics of the dataset:

Logistic Regression
Logistic regression was chosen as a baseline due to its simplicity and interpretability. Despite its limitations in handling complex non-linear relationships, it provided a reliable reference for more complex models.
Rationale: The dataset included both numerical and categorical features with varying scales. Logistic regression offered quick convergence and interpretability, making it a suitable starting point.

Random Forest
Random forest was selected for its ability to handle non-linear relationships, resistance to overfitting, and robustness to noise and outliers. Rationale: Given the presence of potential outliers and noise in the dataset, random forest's ensemble nature allowed better generalization by averaging multiple decision trees.

Gradient Boosting

Gradient boosting was included because of its iterative approach to improve weak learners and its ability to capture complex patterns in the data.
Rationale: Since the dataset involved features with varying degrees of influence on the target variable, gradient boosting's strength in correcting errors of previous models made it a strong candidate.

Support Vector Machine (SVM)
SVM with an RBF kernel was included to explore non-linear decision boundaries.
Rationale: SVM is effective in high-dimensional spaces and performs well in cases where the data is not linearly separable. Given the potential for complex boundaries between classes in road accident data, SVM was a suitable choice.

k-Nearest Neighbors (kNN)
kNN was tested as a classification model because it is simple and effective, especially when decision boundaries are not linear. It predicts the class of a new sample based on the majority class of its nearest neighbors.
Rationale: Since road accident data likely contains complex patterns, kNN was a reasonable option. However, because it can be sensitive to noise and high-dimensional data, it was mainly used for comparison with other models.

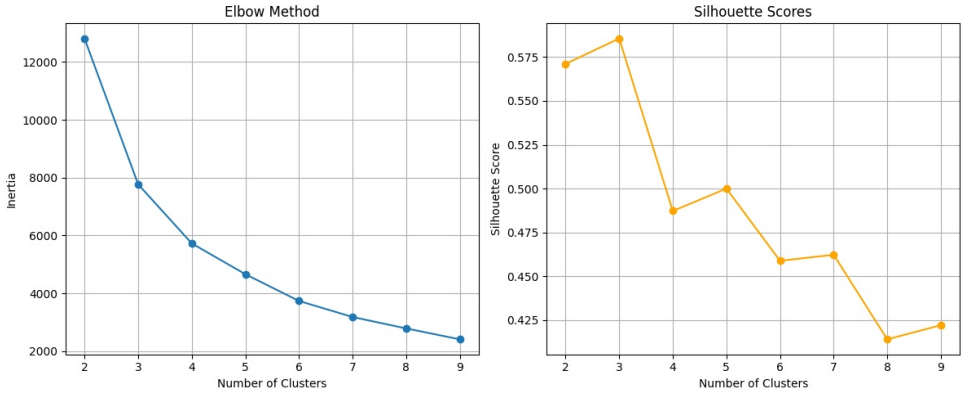
Unsupervised Learning: Clustering Models

KMeans Clustering
KMeans was selected as the primary clustering method due to its simplicity and effectiveness in partitioning data into distinct clusters by minimizing within-cluster variance. The algorithm was applied for various values of **K** (number of clusters), ranging from 2 to 9, to explore different possible groupings.
Evaluation Techniques:

- Elbow Method: This method was used to determine the optimal number of clusters by plotting the inertia (within-cluster sum of squares). The point where the inertia curve starts to level off indicates the most appropriate value for K.
- Silhouette Scores: Silhouette scores were calculated to assess the cohesion and separation of clusters. Higher scores indicated well-defined clusters.
  Both evaluation techniques suggested that **K=2 or K=3** clusters were optimal for the dataset, as these values provided a good balance between simplicity and meaningful separation of the data.

PCA
Given the high dimensionality of the dataset, Principal Component Analysis (PCA) was applied before re-evaluating the clustering results. PCA reduced the number of features while retaining the most significant variance, thereby minimizing noise and redundancy.
By transforming the dataset into principal components, PCA made it easier to visualize the clusters and improved clustering performance, as evidenced by higher silhouette scores post-PCA. Specifically:

- Before PCA, the silhouette score for **K=2** was 0.4891, while after PCA, it increased to 0.5711.
- Similarly, for **K=3**, the silhouette score improved from 0.4700 to 0.5855, indicating better-defined clusters with clearer separation.

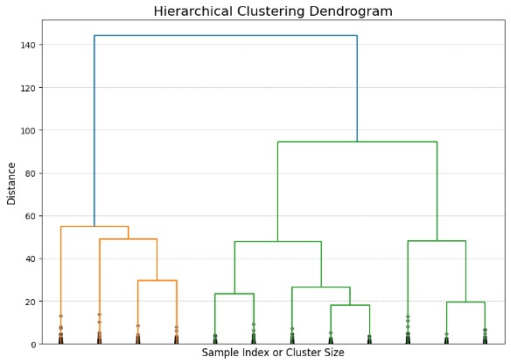


Hierarchical Clustering

To further validate the results obtained from KMeans, Hierarchical Clustering with the Ward linkage method was applied. This approach provided a different perspective on the clustering structure by visualizing the hierarchy of clusters through dendrograms.

Purpose and Results

- The first dendrogram, without specifying a predefined K, revealed a natural split into two primary clusters at a large distance, indicating that **K=2** could be an appropriate choice.
- Additional dendrograms for **K=2** and **K=3** were generated, showing further refinement in the clustering structure.
- This method confirmed that both **K=2** and **K=3** clusters were valid, with **K=3** offering finer-grained insights into accident risk patterns.



Experiments and Results

During the classification phase, we tested five different models: Gradient Boosting, Support Vector Machine (SVM), Logistic Regression, Random Forest, and k-Nearest Neighbors (kNN). After comparing the models, Gradient Boosting emerged as the best performer. Its parameters were tuned using Grid Search, and the following were selected:

- Learning rate was set to 0.2, balancing the speed of convergence and accuracy.
- Number of estimators was set to 100, providing enough iterations for optimal learning without excessive computation.
- Maximum depth was limited to 5 to prevent overfitting while still capturing complex patterns in the data.

In the clustering phase, we employed k-means and hierarchical clustering to explore underlying patterns in the data. The number of clusters, **k=3**, was chosen based on the elbow method and silhouette analysis, which indicated that three clusters offered the best separation. For hierarchical clustering, Ward linkage was used, minimizing variance within clusters and ensuring well-defined groupings.

Classification Metrics

To assess the performance of the classification models, we relied on the following metrics:

- **Accuracy**, which provided an overall measure of correct predictions.
- **Precision**, crucial for minimizing false positives when identifying high-risk areas.
- **Recall**, important for ensuring that all relevant high-risk areas were detected.
- **F1-Score**, a harmonic mean of precision and recall, offering a balanced view of the model's performance.

These metrics were chosen because of their relevance in handling potential class imbalance, which could otherwise skew the results.

Clustering Metrics

For the clustering analysis, traditional metrics like accuracy were not applicable. Instead, we used:

- **Silhouette Score** to evaluate how well the clusters were separated.
- **Within-Cluster Sum of Squares (WCSS)** to measure the variance within clusters, ensuring compact groupings.
- **Distance from Cluster Centroids** to identify anomalies—points farthest from the centroids were flagged for further investigation.

Findings

The findings of this project are twofold, reflecting the dual approach of classification and clustering. Gradient Boosting achieved the best results among all tested models, with an accuracy of 93.4% and precision, recall, and F1-score all approximately equal to 0.934.These results indicate that the model was highly effective in predicting high-risk areas, with minimal bias and variance. Compared to other models, Gradient Boosting demonstrated superior generalization across different datasets.

The clustering analysis revealed three distinct clusters: Cluster 0 represented balanced residential areas with low accident density, Cluster 1 identified high-risk urban areas requiring immediate safety interventions, and Cluster 2 corresponded to large open-area regions with moderate injury levels. Cluster 0 represented balanced residential areas with low accident density.

Additionally, 38 anomalies were detected across the clusters. These anomalies represent specific zones with unique characteristics that may warrant targeted investigations.

The performance of different algorithms varied significantly:

Gradient Boosting stood out due to its iterative approach of combining weak learners, which allowed it to capture complex patterns effectively. While SVM also performed well, it was computationally more expensive and sensitive to parameter tuning. Random Forest provided good results but tended to overfit with default parameters. Logistic Regression, though effective, lacked the flexibility to model non-linear relationships, and kNN struggled with noise and dimensionality, leading to lower performance.

In the clustering phase, k-means proved efficient and provided clear separations, though it was sensitive to initial centroid selection. Hierarchical clustering, while computationally intensive, offered better interpretability through dendrograms, which helped in validating the number of clusters.

The project achieved significant success in combining classification and clustering to provide both predictive insights and exploratory analysis. The classification model accurately identified high-risk areas, and the clustering process highlighted distinct patterns and anomalies, leading to actionable recommendations.

However, some limitations should be noted. The analysis relied on historical data, which may not fully capture future patterns or emerging risks. Additionally, some anomalies could be temporary, influenced by external factors such as road construction or temporary changes in traffic conditions.

Unexpectedly, several anomalies were found in low-risk residential areas, indicating potential hidden risks that might not have been evident in the initial analysis. Furthermore, hierarchical clustering revealed potential sub-clusters within high-risk areas, suggesting that further refinement of the clusters could be beneficial.

Future Directions

Moving forward, there are several key areas for improvement:

- Dynamic Model Evaluation: Implementing a real-time monitoring system to continuously evaluate model performance and update predictions based on new data.
- Enhanced Anomaly Detection: Incorporating additional contextual factors, such as weather conditions and time of day, could improve the accuracy and relevance of anomaly detection.

- Collaboration with Authorities: Partnering with local traffic authorities to validate the findings and implement targeted safety measures could greatly enhance the impact of this project.

By addressing both general patterns and specific cases, this project provides a comprehensive framework for improving traffic safety and identifying high-risk zones effectively.

Conclusion and Discussion

We invested significant effort throughout the past month to complete this project. Our work involved consistent collaboration and independent research, ensuring that we both developed a deep understanding of all aspects of the process. Although we divided the tasks according to the project guidelines, we regularly met every day or two to share insights, discuss progress, and merge our findings into a unified file, which allowed us to advance efficiently while maintaining high accuracy. Rotem was responsible for developing the functions for PCA and during the dimensionality reduction phase. This required detailed research into each parameter to ensure the most effective results. Nofar focused on the classification phase, where she built and fine-tuned the models and developed the key function for model evaluation. Both of us collaborated closely on the model evaluation function, revising and refining it several times to improve performance.
One of the most challenging and critical parts of the project was handling the outliers. We dedicated an entire week to this task, as it was crucial to ensure that we did not remove data points that could represent real and meaningful cases. This step required careful analysis and thoughtful consideration to balance data integrity with reliable clustering outcomes. Our mutual effort in this phase resulted in a robust method for identifying and addressing outliers without compromising the quality of the data.

In the clustering phase, we worked together to analyze the outputs from k-means and hierarchical clustering. Rotem implemented hierarchical clustering to compare its results with those of k-means, and together we discussed the findings and reached final conclusions. This joint effort ensured that all clustering decisions were well-supported and clearly documented.

As we approached the final stages of the project, Rotem worked on the presentation, while Nofar focused on writing the report. We worked really well as a team, helping each other whenever needed and making sure everything was done to the highest standard. Our combined efforts made this project a success.

.