Summary Report: Analysis of Road Accidents

Presenters:

- Ariel Koren - 318284239
- Yosef Bankurmono – 318455904

GITHUB:

Introduction

The issue of road accidents remains a significant challenge to public safety worldwide, with far-reaching consequences for human life. This project focuses on analyzing urban accident data to identify patterns and trends that can serve as a basis for improving road safety. We aim to create models for predicting high-risk accident areas and understand the impact of various characteristics, such as infrastructure, vehicle types, and demographic characteristics, on the severity of accidents.

Our main objective is to leverage machine learning techniques and advanced data analysis to examine the complex relationships between accident variables and their outcomes. In addition, we aim to generate practical insights that can assist local authorities and planning bodies in designing prevention policies and resource allocation. The use of real data allows us not only to understand the phenomenon in depth but also to demonstrate how technological tools can be integrated into decision-making processes. Within this project, we will examine questions such as: Which infrastructures and residential areas are more prone to accidents? How can data analysis technologies be used to reduce the number of accidents? And how can the impact of policy changes on accident severity be measured over time?

Data Set

The data set includes 1,174 records with 27 columns, describing accidents in various cities. Key features include the number of accidents (SUMACCIDEN), fatalities (DEAD), serious injuries (SEVER_INJ), and the types of vehicles involved.

The preprocessing process was complex and included several steps:

- **Handling Missing Values:** Missing values, mainly in the columns related to injuries and fatalities, were completed by calculating averages according to municipal areas (groupby).
- **Data Standardization:** Numerical data was standardized to prevent the influence of different scales on the model results.
- **Categorical Variable Encoding:** Categorical variables, such as vehicle types and districts, were encoded using One-Hot Encoding.
- **Feature Engineering:** Calculated columns were added, such as the percentage of serious injuries (PCT_SEVER), the percentage of fatalities (PCT_DEAD), and an encoded risk level (RISK_LEVEL_ENCODED).
- **Correlation Analysis:** Correlation analysis was performed between the features to identify possible relationships and remove highly correlated variables.

Methodology

During the project, we used various machine learning algorithms, including Random Forest, XGBoost, and Logistic Regression, due to their ability to handle structured data and the intuitive interpretation of the results.

The methodology included the following steps:

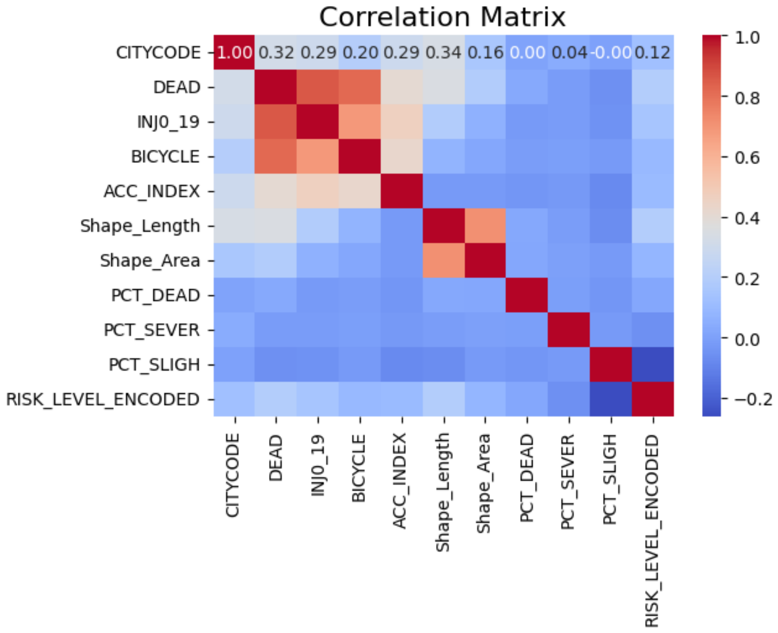
- **Preprocessing and Data Standardization:** Correction of missing values using groupby and Z-Score standardization for numerical data.
- **Feature Engineering:** Adding calculated columns (PCT_DEAD, PCT_SEVER, RISK_LEVEL_ENCODED) and correlation analysis.
- **Dimensionality Reduction:** Using PCA to reduce dimensions while preserving the critical information. For example, we reduced the data dimensions to five main components that retained over 90% of the variance.
- **Model Training and Cross-Validation:** Training models using cross-validation and tuning hyperparameters using Grid Search.
- **Evaluation:** Evaluating model performance using metrics such as Accuracy, Precision, Recall, and F1 score.

Experiments and Results:

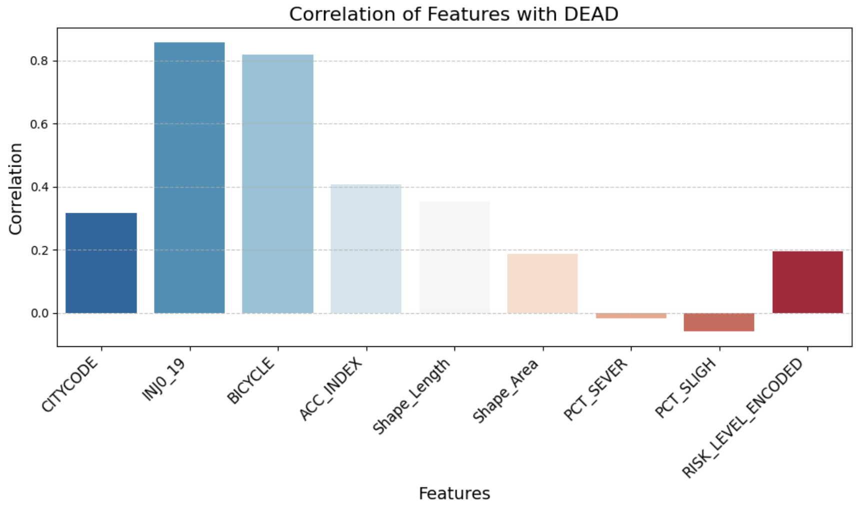
Parameter Selection

Initially, we worked with the following parameters: 'CITYCODE', 'DEAD', 'INJ0_19', 'BICYCLE', 'ACC_INDEX', 'Shape_Length', 'Shape_Area', 'PCT_DEAD', 'PCT_SEVER', 'PCT_SLIGH', 'RISK_LEVEL_ENCODED.'

The final parameter selection was done using VIF (Variance Inflation Factor) testing and correlation analysis. We filtered out columns with high VIF and analyzed their correlation with the target variable "DEAD". After visual inspection, the columns 'INJ0_19', 'BICYCLE', 'ACC_INDEX', 'Shape_Length', 'RISK_LEVEL_ENCODED' were chosen.



"The heatmap visualizes the impact of various variables on the DEAD column, which we then displayed in a bar plot for easier visual analysis."



Based on these data, we decided to proceed with only the 'INJ0_19', 'BICYCLE', 'ACC_INDEX', 'Shape_Length', and 'RISK_LEVEL_ENCODED' columns.

For Clustering - the following parameters were chosen after removing non-numerical columns: 'SUMACCIDEN', 'DEAD', 'SEVER_INJ', 'SLIGH_INJ', 'PEDESTRINJ', 'INJ0_19', 'INJ20_64', 'INJ65_', 'INJTOTAL', 'TOTDRIVERS', 'MOTORCYCLE', 'TRUCK', 'BICYCLE', 'PRIVATE', 'VEHICLE', 'ACC_INDEX', 1 'AreaSQKM', 'DISTRICT_דרום', 'DISTRICT_חיפה', 'DISTRICT_ירושלים', 'DISTRICT_מרכז', 'DISTRICT_צפון', 'DISTRICT_תל אביב', 'PCT_DEAD', 'PCT_SEVER', 'PCT_SLIGH', 'RISK_LEVEL_ENCODED', 'ACC_SEVERITY', 'VEHICLE_DENSITY'.

Evaluation Metrics

Model evaluation was performed using standard metrics:

- **Accuracy:** For example, Random Forest achieved an accuracy of 87%.
- **Precision and Recall:** For example, XGBoost achieved a Precision of 0.82 and a Recall of 0.78.
- **F1 Score:** For example, Random Forest achieved an F1 score of 0.85.
- **ROC-AUC:** For example, Random Forest achieved an AUC of 0.92.

Findings

The Random Forest model achieved the best performance among all the models tested, with an accuracy of 87%. This model excelled not only in its high accuracy but also in its ability to provide a high F1 score of 85%, indicating a good balance between precision and recall. In comparison, the XGBoost model demonstrated competitive performance with an accuracy of 85% and an F1 score of 83%, but required longer training times due to its dynamic parameter tuning. Logistic Regression, which served as a baseline model for comparison, achieved an accuracy of only 75% and a lower F1 score of 72%, suggesting it is less suitable for this complex problem.

Limitations and Insights

Several limitations were identified during the analysis. The imbalanced data posed a significant challenge, especially when certain categories were represented at a lower frequency than others. This led to a decrease in the performance of some models in rare categories. For example, the recall in rare categories in the XGBoost model was only 0.6. To address these limitations, it is recommended to explore techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to increase the diversity in rare categories and improve overall performance. In addition, examining more advanced models, such as Deep Learning, may be a further step in dealing with these complex data.

Clustering

The project also used clustering algorithms to identify possible groups and clusters within the data. The selected algorithms include KMeans, DBSCAN, and Agglomerative Clustering. The main objectives were to identify geographic groups with similar risk levels and analyze the characteristics that distinguish between the groups.

- **KMeans:** This method was chosen for its simplicity and ability to handle large datasets. The algorithm attempted to identify the optimal number of clusters using the Elbow method. For example, the choice of the number of clusters was determined by the clear break obtained in the Inertia graph.
- **DBSCAN**: This method was chosen for its ability to identify non-linear clusters and handle noise. During the analysis, the algorithm successfully identified areas with high accident concentrations that were considered "dense clusters," while other areas were marked as noise.
- **Agglomerative Clustering:** This algorithm was used for a hierarchical understanding of the clusters. Dendrogram analysis was used to select the most appropriate number of clusters, as well as to identify a geographic hierarchy between the areas.

Clustering Evaluation Metrics

Clustering was evaluated using the Silhouette Score and the Davies-Bouldin Index. The results are:

- **KMeans:**
  - Silhouette Score: 0.947
  - Davies-Bouldin Index: 0.312
- **DBSCAN:**
  - Silhouette Score: 0.094
  - Davies-Bouldin Index: 1.087
- **Agglomerative Clustering:**
  - Silhouette Score: 0.954
  - Davies-Bouldin Index: 0.254

Since Agglomerative Clustering and KMeans had the highest Silhouette scores, we continued to examine the clusters they created.

Conclusions from Clustering

The results led to the identification of three main clusters:

- **Cluster 0 - Low-Risk Areas:** This cluster includes local authorities with a very low number of accidents (SUMACCIDEN = 136 on average) and minimal injuries in all categories (DEAD = 0.011 on average, SEVER_INJ = 0.0058). The areas are characterized by a low-risk level (RISK_LEVEL_ENCODED ≈ 1.18). Despite the low number of accidents, the percentages of serious injuries (PCT_SEVER = 4.46) and fatalities (PCT_DEAD = 2.69) are relatively high. This cluster is mainly concentrated in areas in the north of the country (DISTRICT_ = צפון 40%).

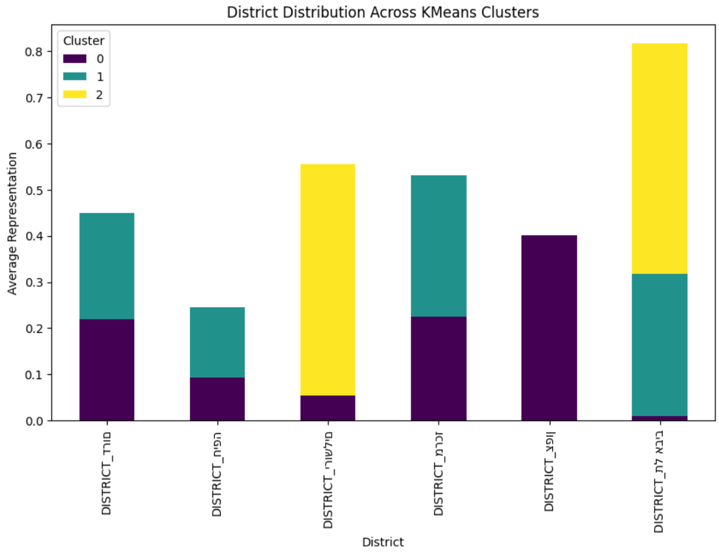
- **Cluster 1 - Medium-Risk Areas:** This cluster includes authorities with a medium number of accidents (SUMACCIDEN ≈ 6707) and medium levels of injuries (DEAD = 0.33 on average, SEVER_INJ = 0.24). The percentages of serious injuries (PCT_SEVER = 0.69) are lower compared to Cluster 0. The risk level (RISK_LEVEL_ENCODED = 2.0) indicates urban areas with complex infrastructures. This cluster is mainly concentrated in the Tel Aviv area (DISTRICT_22% = תל אביב).
- **Cluster 2 - High-Risk Areas:** This cluster includes authorities with the highest number of accidents (SUMACCIDEN ≈ 19125) and the highest levels of injuries (DEAD = 0.93, SEVER_INJ = 0.92). The percentages of slight injuries (PCT_SLIGH = 1.0) are relatively low compared to the total number. The risk level (RISK_LEVEL_ENCODED = 2.0) indicates a high concentration of accidents in crowded city centers. This cluster is mainly concentrated in Jerusalem and Tel Aviv (DISTRICT_Jerusalem = 50%, DISTRICT_50% = תל אביב).

Cluster Comparison

- Cluster 0: Few accidents but relatively high percentages of serious injuries, possibly due to a lack of suitable infrastructure in peripheral areas.
- Cluster 1: Medium levels of accidents and injuries, suitable for urban areas with diverse transportation.
- Cluster 2: A particularly high concentration of accidents in large city centers due to traffic congestion and inadequate infrastructure.

General Conclusion Regarding Clustering

The Clustering successfully distinguished between low, medium, and high-risk areas based on accident characteristics. The results emphasize the need for tailored approaches to different clusters: infrastructure reinforcement in Cluster 0, improved traffic management in Cluster 1, and congestion reduction projects in Cluster 2.



Project Implications

This project has yielded a prediction model that assists in estimating fatalities in relation to accidents and their direct influencing factors, thereby aiding in understanding how to prevent high-fatality risks in accidents. We also successfully categorized groups of parameters that affect the number and severity of accidents in different geographic areas, enabling a focus on the main causes of accidents and effective changes on the ground.

Team Member Roles

The work was carried out in close collaboration, with each team member focusing on improving specific parts of the overall project:

- Yosef Benkurmono: Building the full classification pipeline and preparing the presentation.
- Ariel Koren: Building the clusters and preparing the summary report.

Future Directions

- Investigating causal relationships to formulate better policies.
- Examining which influences led to the variables affecting the nature of the accident and creating proactive preventive measures.