# PREDICTIVE INSIGHTS FOR LESS ACCIDENTS

REPRESENTS : ASAF BENKORMONO AND ARIEL KOREN
DOCTOR : HEN HAGAG

# CHALLENGES WE FACED

01 Cleaning, normalizing, and organizing data involves several challenges.

02 Initially, handling missing values can be complex, especially when the pattern of missingness is not random

03 Choosing the appropriate imputation technique that give us the best results requires careful consideration.

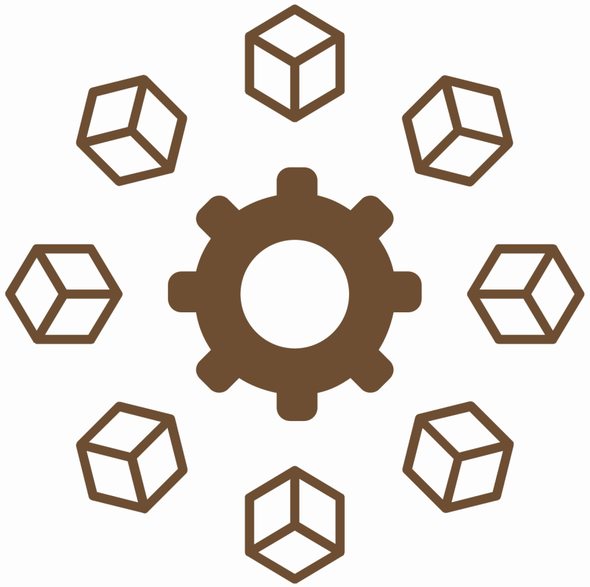
04 Feature engineering and selection present their own difficulties.

05 Determining the most relevant features and avoiding overfitting by reducing irrelevant or redundant features

# THE GOAL OF THE PROJECT

The primary goal of the project is to leverage advanced machine learning techniques and data analysis to examine the complex relationships between accident variables and their results. Additionally, we aim to generate practical insights that can assist local authorities and planning bodies in formulating prevention policies and allocating resources effectively.
This project focuses on analyzing urban accident data to identify patterns and trends that can improve road safety. Our goals include developing models to predict high-risk areas for accidents and understanding how factors like infrastructure, vehicle types, and demographics influence accident severity.

# OTHER TECHNIQUES



In the analysis of road accidents, common methods include Logistic Regression, KMeans, DBSCAN, and Neural Networks.

- Logistic Regression: A simple method but limited in handling complex data.
- KMeans: Suitable for clustering data but sensitive to noise.
- DBSCAN: Effective in detecting noise but struggles with overlapping clusters.
- Neural Networks: Powerful for large datasets but require high computational resources.

In this project, we selected Random Forest and XGBoost for their robustness and ability to identify important features, and KMeans and Agglomerative Clustering for analyzing risk zones and providing a hierarchical understanding of the clusters.

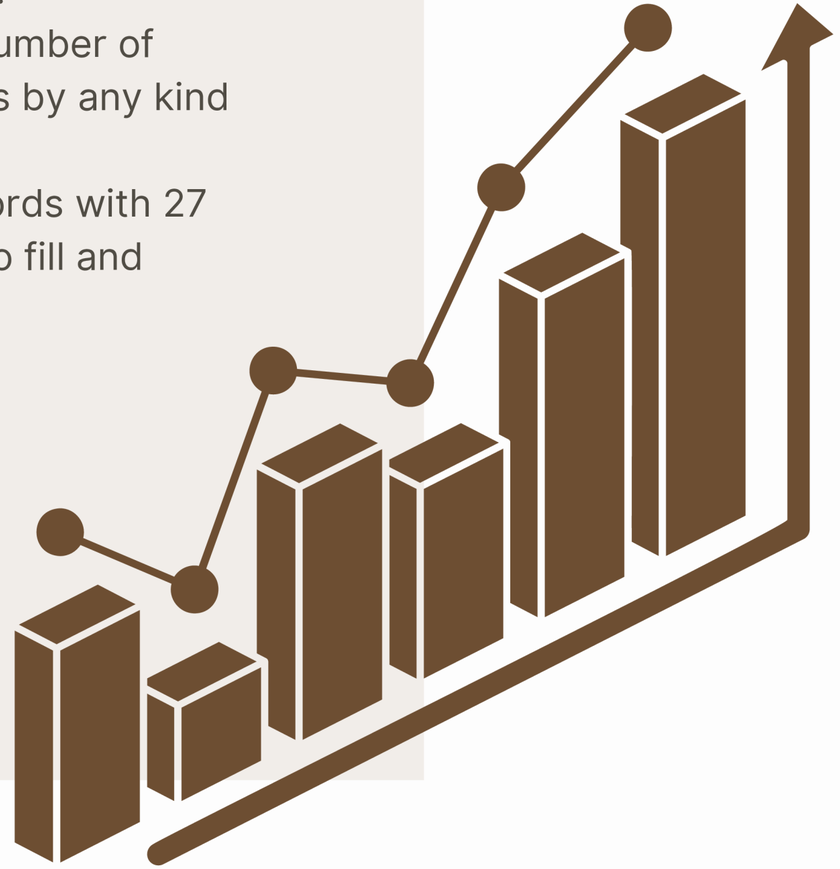
# OUR DATA

## DATA COLUMNS

- **SUMACCIDEN** – Number of accidents.
- **DEAD** – Number of fatalities.
- **SEVER_INJ** – Number of severe injuries.
- **BICYCLE** – Types of vehicles involved (e.g., bicycles, cars, etc.).
- **Shape_Length** – Geospatial data representing the length of the accident's location.
- **Shape_Area** – Geospatial data representing the area of the accident's location.
- **PCT_DEAD** – Percentage of fatalities out of total accidents.
- **PCT_SEVER** – Percentage of severe injuries out of total injuries.
- **PCT_SLIGH** – Percentage of slight injuries (if present).

## DATA STRUCTURE

The dataset contains accident data categorized by municipal areas.
It includes information on the number of accidents, fatalities, and injuries by any kind of vehicle .
The dataset contains 1,174 records with 27 columns,certain fields require to fill and normalize.

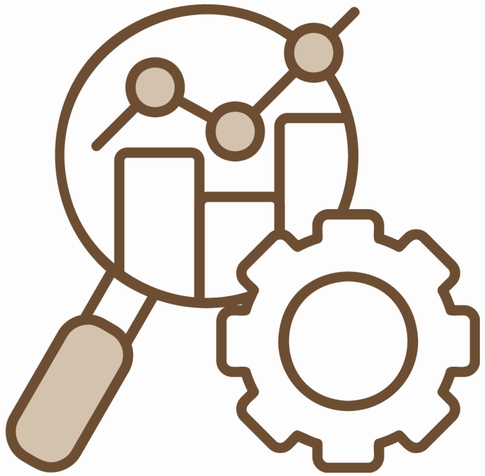


# METHODOLOGY

**ALGORITHMS AND ANALYSE**

We used both supervised and unsupervised machine learning algorithms to analyze accident data.
In supervised learning, we employed models such as Logistic Regression, Decision Tree, Random Forest, Ridge, and Gradient Boosting, with Gradient Boosting performing the best.
For unsupervised learning, we used KMeans, DBSCAN, and Agglomerative Clustering, with Agglomerative Clustering yielding the best results.
Missing values were filled using the groupby method, and numerical data was normalized.
We engineered new features like fatality and injury percentages and encoded risk levels.
PCA was applied for dimensionality reduction while retaining key information.
Models were trained with cross-validation, hyperparameters were tuned for improved accuracy, and model quality was assessed using model evaluation techniques.

# EXPERIMENTS AND EVALUATION

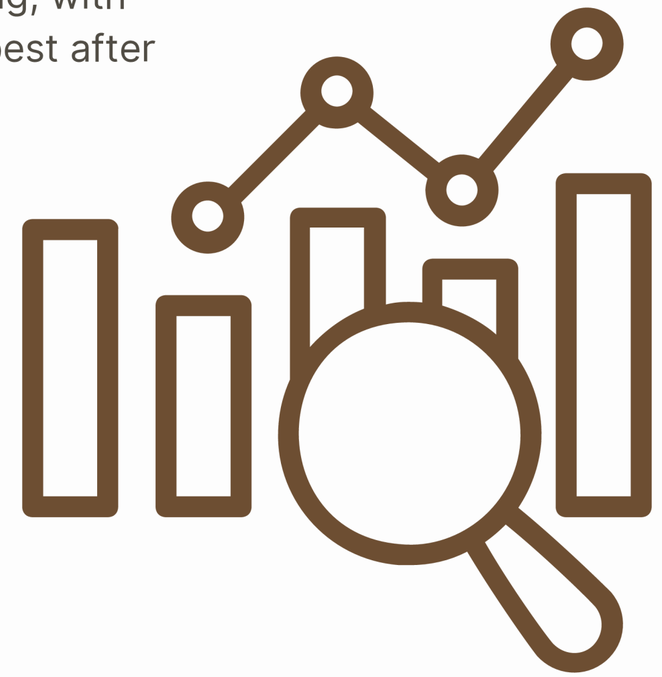


## SUPERVISED MODELS and evaluate

During the experiments and evaluation, various models were tested using performance evaluation techniques such as Accuracy, Precision, Recall, and F1 Score. The supervised models included Logistic Regression, Decision Tree, Random Forest, Ridge, and Gradient Boosting, with Gradient Boosting performing the best after careful check of overfitting .

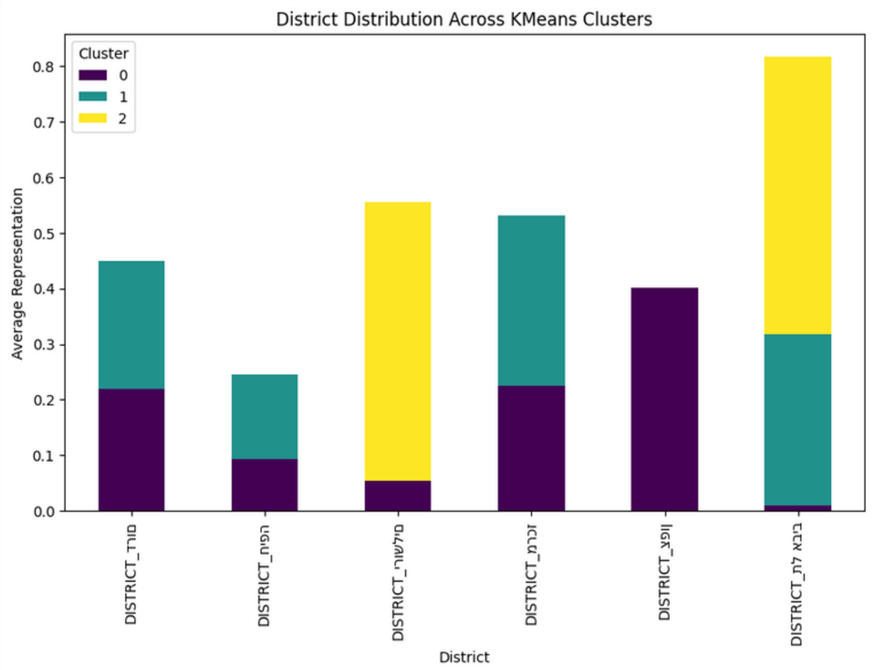
## UNSUPERVISED MODELS AND EVALUATE

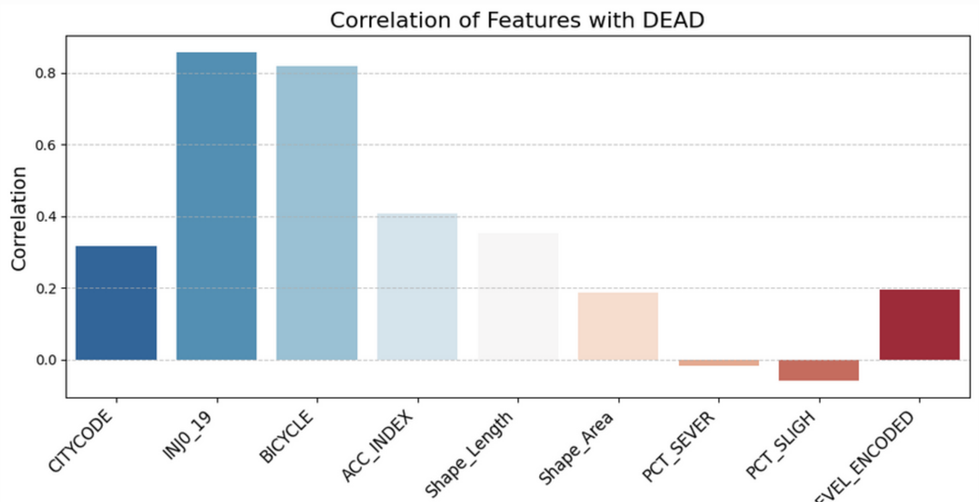
Models such as KMeans, DBSCAN, and Agglomerative Clustering were tested, with Agglomerative Clustering providing the best results.
The unsupervised models were evaluated using PCA and T-SNE for dimensionality reduction, as well as the Silhouette Score

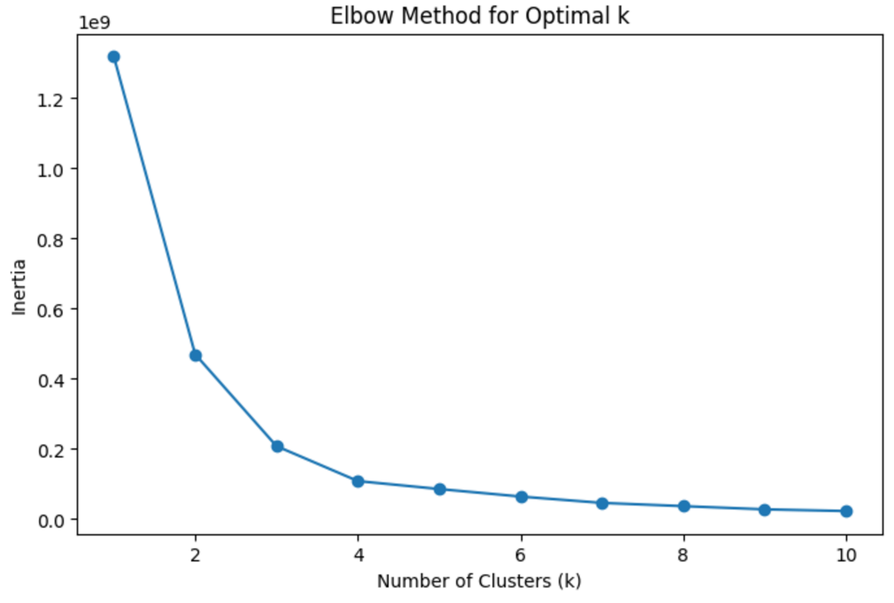


# RESULTS ACHIEVED

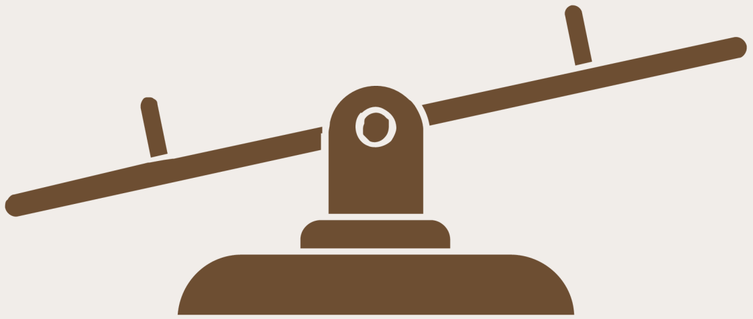
For unsupervised learning, three clustering models were tested.
KMeans identified three clusters based on accident and injury levels.
Cluster 0 represented low-risk areas, Cluster 1 was medium-risk with urban areas, and Cluster 2 had high-risk areas with densely concentrated accidents.
Similarly, Agglomerative Clustering identified the same three clusters. The quality of the clustering models was evaluated using Silhouette Score, PCA, and T-SNE visualizations, confirming their effectiveness in grouping accident data and identifying risk patterns.
For supervised models, we focused on predicting fatalities in accidents.
Logistic Regression, Decision Tree, Random Forest, Ridge, and Gradient Boosting were tested.
The Gradient Boosting model performed the best, effectively predicting the likelihood of fatalities using features like accident count, injury severity, vehicle types, and geographical data.



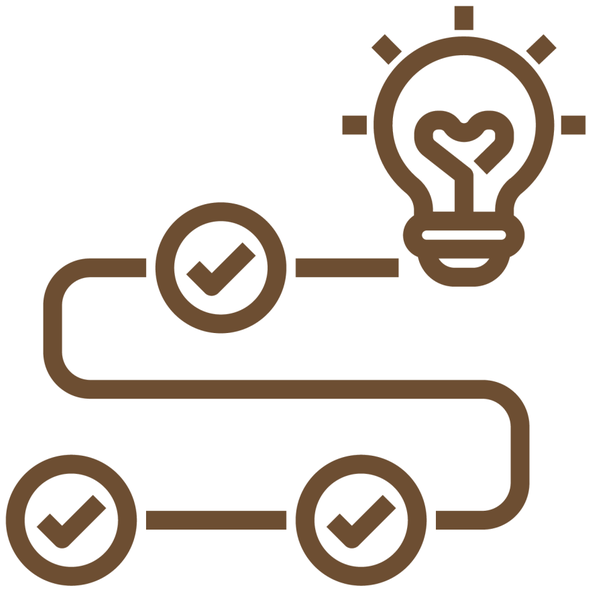



# COMPARISON BETWEEN TECHNIQUES



The experiments were designed to assess the impact of various features and hyperparameter configurations on model performance.
Hyperparameter Selection: The hyperparameters of each model were fine-tuned using grid search.
For example, the Random Forest model achieved optimal performance with 100 trees and a maximum depth of 10.
Evaluation Metrics: The primary metrics used were accuracy and the F1 score, which provided a balanced view of model performance on imbalanced datasets.
Findings: The Random Forest model performed the best, achieving 87% accuracy.
The XGBoost model showed competitive results but required longer training times.
Logistic Regression provided baseline results with an accuracy of 75%.
Limitations and Insights: While the models demonstrated good overall performance, they struggled with significantly imbalanced data.
It is recommended to explore resampling techniques, such as SMOTE, to address this issue in future models.



# CONCLUSION

This project successfully demonstrated the use of machine learning techniques to analyze traffic accident data.
Key contributions include identifying influential features, optimizing predictive models, and highlighting the need for data balancing in future studies.
The “Unsupervised” model Agglomerative Clustering, effectively uncovered hidden patterns in the data, particularly in distinguishing between high-risk and low-risk areas.
The “Supervised” models, especially Random Forest and Gradient Boosting, performed well in predicting fatalities and severe injuries.
Incorporating additional data, such as seasonal or environmental factors, could further enhance model performance and provide more comprehensive insights into traffic safety.

# THANK YOU

## A&A TEAM

**ARIEL KOREN - 318284239**

**ASAF BENKORMONO - 318455904**