ML Project- Unsupervised and Supervised Learning

Subject: Predicting High-Risk Traffic Areas in Israel
Course: Advanced Topics in Machine Learning
Lecturer: Chen Hagag
Team Members: Nofar Mahrabi & Rotem barel

The Problem

Road accidents continue to occur frequently due to various factors such as human error, poor road conditions, and inadequate infrastructure.

- There is growing awareness of the need to identify high-risk traffic areas, especially in regions where road and transportation infrastructure is underdeveloped.

- Understanding the patterns that contribute to these areas becoming hazardous is crucial for improving infrastructure, urban planning, and road safety.

- Analyzing these high-risk zones can provide critical insights to help reduce accident rates and save lives.

Original Goals of the Project

Identify high-risk traffic areas:
Predict zones in Israel with a high likelihood of road accidents using machine learning.

Investigate the relationships between various factors and accident severity:
Study how road conditions, accident numbers, and vehicle types affect accident severity.

Detect anomalies in accident patterns:
Identify unusual trends in accident data for better safety measures.

Provide actionable insights for urban planning and road safety improvements:
Provide recommendations for infrastructure upgrades and improved road safety policies.

General Methods and Techniques Used

- **Data Collection:**
  Integrating traffic, Region, and road condition data from multiple sources.

- **Data Cleaning:**
  Handling missing values, removing outliers, and standardizing data.

- **Feature Engineering:**
  Creating new variables and categorizing accident data for improved analysis.

- **Exploratory Data Analysis (EDA):**
  Visualizing patterns using graphs, heatmaps, and trend analysis.

- **Geospatial Analysis:**
  Mapping accidents and identifying high-risk areas with clustering.

Data Overview

Index	POP_2018	city	MAINUSE	TAZAREA	SUMACCIDEN	DEAD	SEVER_INJ	SLIGH_INJ	PEDESTRINJ	INJ0_19	...	TOTDRIVERS	MOTORCYCLE	TRUCK	BICYCLE	PRIVATE	VEHICLE	ACC_INDEX	Shape_Length	Shape_Area	RISK_LEVEL
0	9948	ראשון לציון	מגורים	852978	50	0	14	44	28	13	...	76	10	0	3	31	62	58.618	3765.78645	8.53E+05	High
1	6378	טבריה	מגורים	1475818	10	0	1	19	1	4	...	21	1	0	2	8	13	6.776	6981.50653	1.48E+06	Medium
2	12277	ירושלים	מגורים	346241	17	0	4	20	10	8	...	26	1	0	0	12	21	49.099	3958.16108	3.46E+05	High
3	5341	הרצליה	מגורים	499512	10	1	1	9	5	1	...	16	1	0	1	9	12	20.02	3065.80174	5.00E+05	Medium
4	0	קרית אתא	שטח פתוח	508601	46	2	13	112	0	19	...	101	1	2	0	54	83	90.444	3692.33382	5.09E+05	High
5	4250	אשדוד	מגורים	339317	20	0	4	27	5	5	...	35	0	0	0	14	32	58.942	2799.18919	3.39E+05	High
6	2992	טל שחר נחשון בקוע נווה שלום	מגורים	28874994	20	1	10	53	0	11	...	43	1	2	0	16	35	0.693	34926.8309	2.89E+07	High
7	1737	תל אביב - יפו	מגורים	230938	64	1	16	62	5	3	...	134	17	0	2	48	98	277.13	1966.35898	2.31E+05	High
8	19257	בני ברק	מגורים	435224	26	0	6	23	21	15	...	35	4	0	3	14	29	59.739	2970.81937	4.35E+05	High
9	0	דאהרייה	לא ידוע	58946757	6	0	0	14	0	2	...	11	0	0	0	7	9	0.102	40071.5074	5.89E+07	Medium

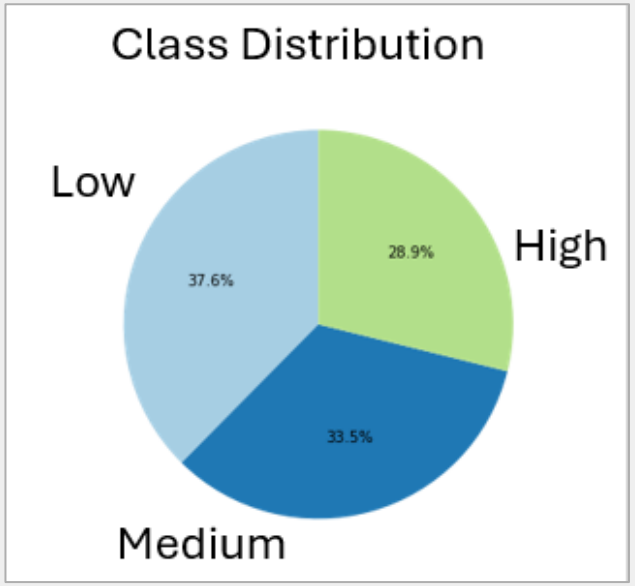
Data Overview

Feature	Description
CITY	Municipality name
CITYCODE	Municipality code
TAZAREA	Traffic zone area (m^2)
POP_2018	Population in 2018
ACC_INDEX	Accident density per km^2
SUMACCIDEN	Total accidents in the area
INJTOTAL	Total casualties (fatalities + injuries)
DEAD	Number of fatalities
SEVER_INJ	Severe injuries
SLIGH_INJ	Minor injuries
INJ0_19	Casualties aged 0-19
INJ20_64	Casualties aged 20-64
INJ65_	Casualties aged 65+
VEHICLE	Total vehicles involved
TOTDRIVERS	Total drivers involved
BICYCLE	Bicycles involved
MOTORCYCLE	Motorcycles involved
PRIVATE	Private vehicles involved
TRUCK	Trucks over 3.5 tons involved
PEDESTRINJ	Injured pedestrians
USETYPE	Area land use classification
MAINUSE	Primary land use type
YEARMONTH	Year and month of accidents

Supervised Model – Methodology

For the supervised classification task of predicting road accident risk levels, the following models were employed:

- **Logistic Regression**: Used as a simple and interpretable baseline to evaluate mixed feature types.
- **Random Forest**: Chosen for its robustness to noise, ability to handle non-linear relationships, and strong generalization.
- **Gradient Boosting**: Applied to capture complex patterns and refine predictions iteratively, leveraging varying feature importance.
- **SVM (Support Vector Machine)**: Effective for non-linear class separations and high-dimensional data.
- **kNN (k-Nearest Neighbors)**: Used for its simplicity in identifying local patterns, serving primarily as a benchmark model.



Unsupervised Model – Methodology

- **Dimensionality Reduction (PCA)**
  1. Reduced feature dimensions while retaining key variance.
  2. Simplified data for better clustering and reduced noise.

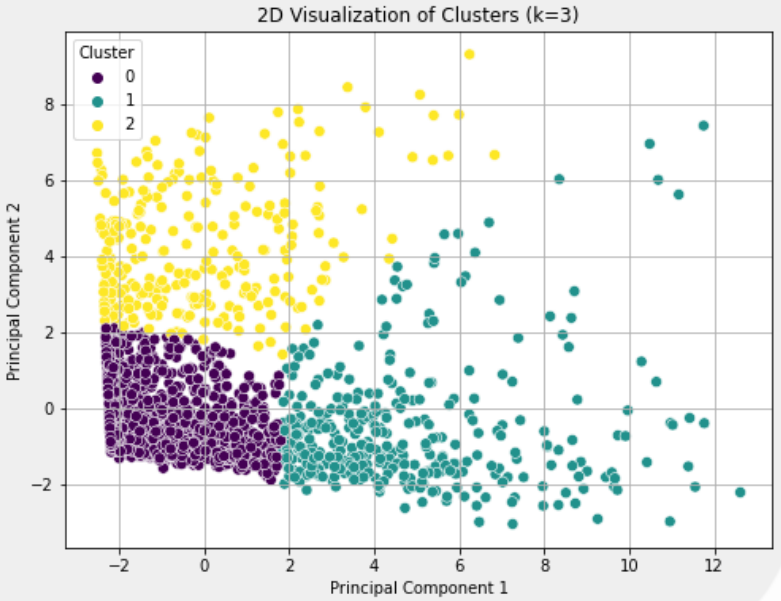
- **Choosing Clusters (Elbow Method)**
  1. Determined the optimal number of clusters by analyzing the Within-Cluster Sum of Squares (WCSS).

- **K-Means Clustering**
  1. Grouped traffic zones based on accident patterns to identify high-risk zones.

- **Cluster Evaluation (Silhouette Score)**
  1. Used the Silhouette Score to evaluate the quality of clustering.
  2. Ensured that clusters are well-separated and cohesive.

Why PCA, K-Means, and Silhouette Score?

- **PCA:** Reduces dimensionality, improving clustering accuracy.
- **K-Means:** Efficiently groups similar traffic zones to uncover patterns.
- **Silhouette Score:** Validates the clustering structure and ensures meaningful results.



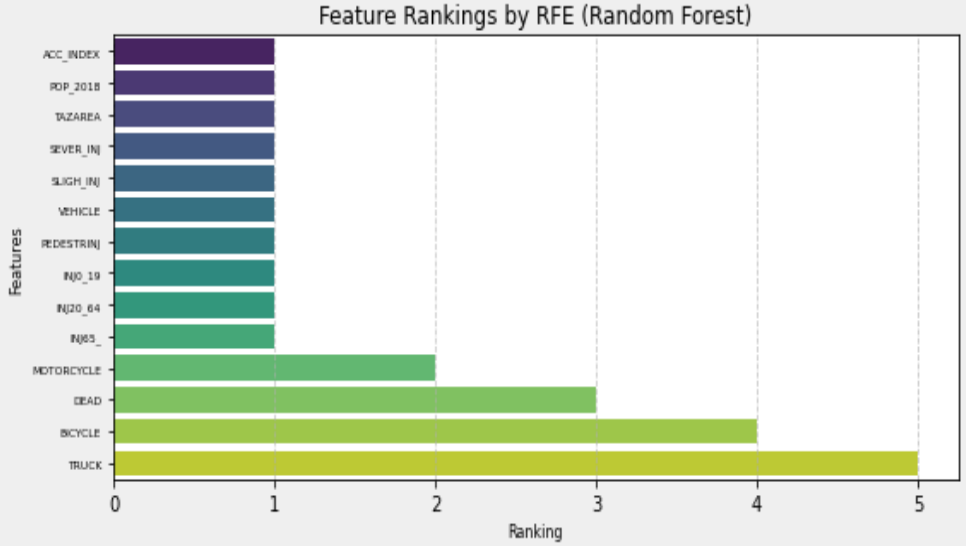
Experiments and Evaluation Techniques

Model Type	Model/Technique	Description	Evaluation Metrics
Supervised	Gradient Boosting	Classification model to predict accident severity. Hyperparameter tuning (GridSearchCV) optimized learning rate, number of estimators, and max depth.	Accuracy, Precision, Recall, F1-score, Cross-validation
Unsupervised	K-Means with PCA	PCA reduced dimensionality; K-Means grouped traffic zones. Tested cluster numbers (K) from 2 to 9.	Elbow Method, Silhouette Score

Results Model Supervised

Model Evaluation Results with Best Parameters:

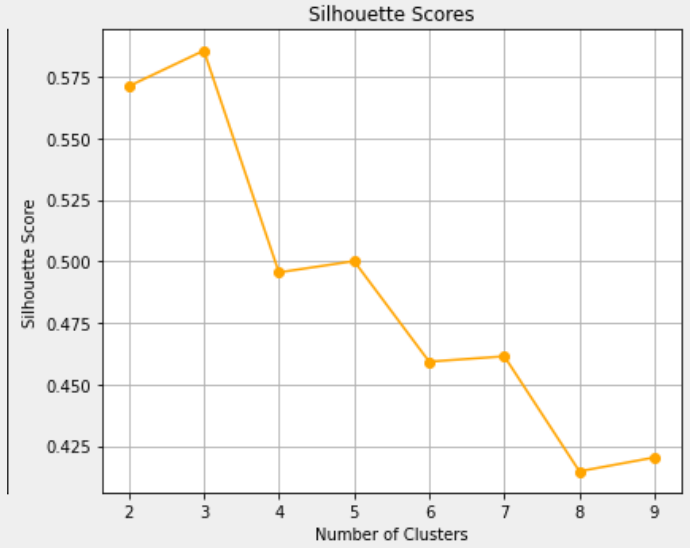
	Model	Precision	Recall	F1-Score	Accuracy
0	Gradient Boosting	0.938430	0.938462	0.938191	0.938462
1	SVM	0.936376	0.936264	0.936268	0.936264
2	Logistic Regression	0.933871	0.934066	0.933891	0.934066
3	Random Forest	0.930628	0.929670	0.929891	0.929670
4	kNN	0.892348	0.892308	0.891563	0.892308

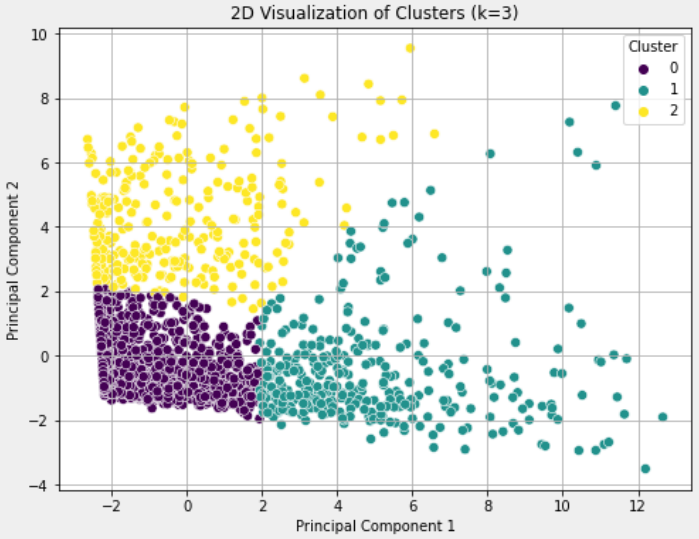


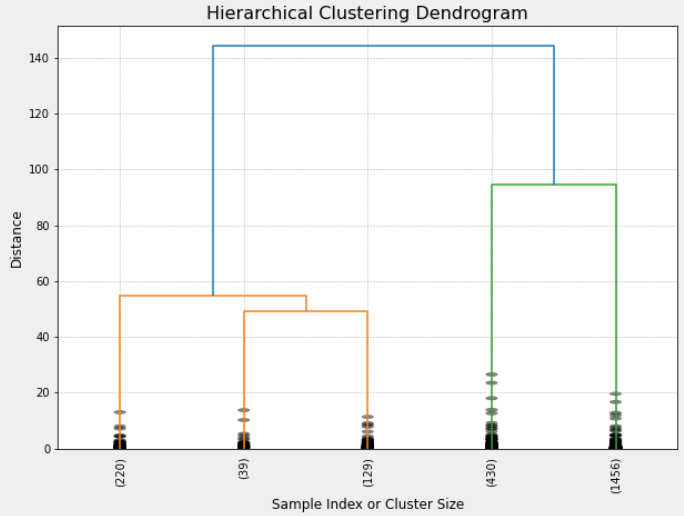
- **Model Performance**
  Gradient Boosting achieved the best performance with 93.8% accuracy.
  Random Forest and SVM performed slightly worse but remained strong.
  kNN showed the lowest performance due to noise sensitivity.

- **Feature Importance**
  ACC_INDEX, POP_2018, VIHICLE, TAZAREA had the highest predictive impact.

Results Model Unsupervised

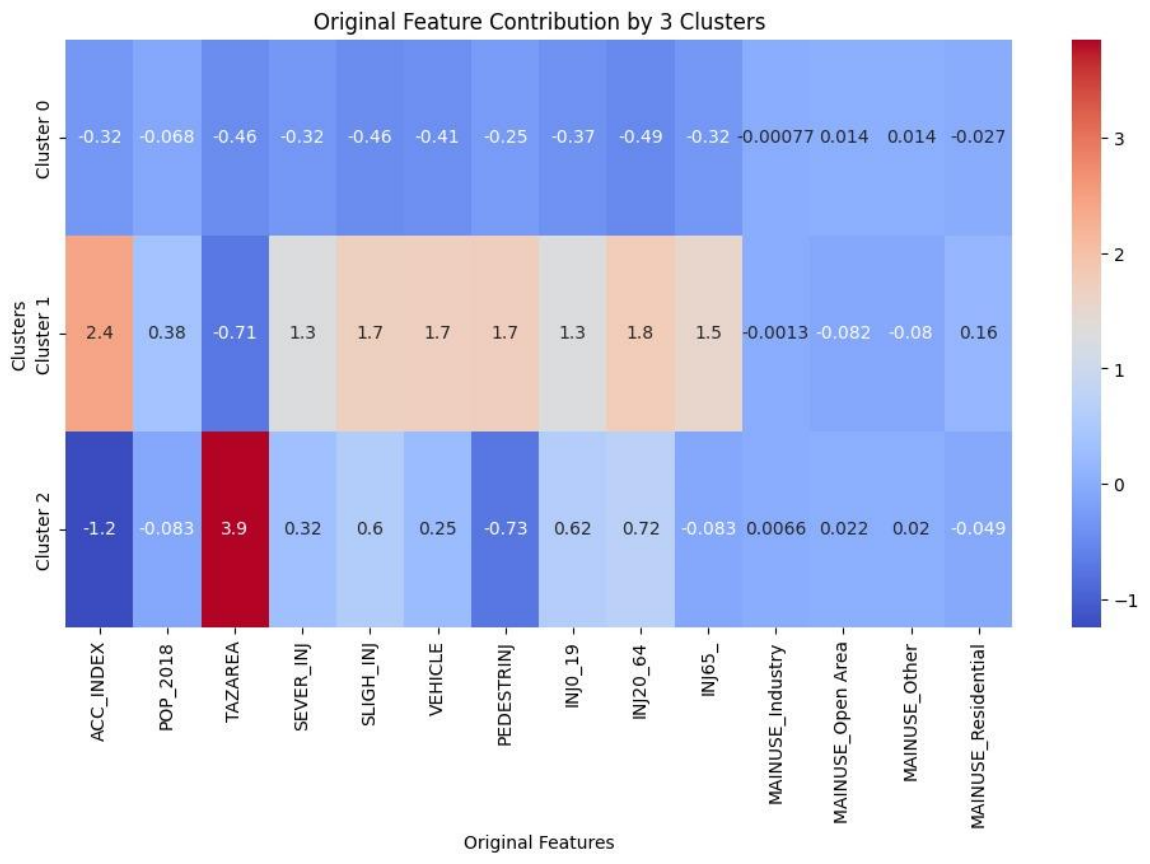






- **Silhouette Scores:**
  The highest silhouette score was achieved with three clusters, indicating that the optimal number of clusters is **K = 3**.
- **2D Cluster Visualization:**
  The PCA-reduced data shows distinct separations between the three clusters, further confirming the optimal value of **K = 3**.
- **Hierarchical Clustering Dendrogram:**
  The dendrogram validates the presence of three main clusters, supporting the findings from the silhouette score analysis.

Results Model Unsupervised



- **Cluster 0:** Dominated by features like PEDSTRINJ and USETYPE, indicating these areas have moderate accident risks due to human activity and land use factors.
- **Cluster 1:** Strongly influenced by TZAAREA (Geographic Area) and POP_2018, suggesting that densely populated regions contributing significantly to higher accident risks.
- **Cluster 2:** Key contributions come from VEHICLE and SEVER_INJ, indicating clusters with high accident severity involving multiple vehicle.

Comparison Table for Supervised Model

Feature/Model	Gradient Boosting	Random Forest	SVM	kNN	Logistic Regression
Learning Method	Boosting (combining weak models)	Bagging (ensemble of weak models)	Classic (Kernel-based)	Distance-based	Classic (Linear)
Non-linearity Handling	Very good	Good	Excellent (via kernel)	Limited	Weak
Noise Sensitivity	Moderate	Low	High	High	Low
Parameter Usage	Improves with hyperparameter tuning	Low dependency on parameters	Sensitive to C and γ	Sensitive to k	Minimal sensitivity
Small Group Adaptability	Good	Good	Limited	Limited	Limited
Handling Missing Data	Good with transformations	Good	Poor	Poor	Good
Run Time	Relatively slow	Faster	Slow	Fast	Fast
Model Interpretability	Complex to understand	Relatively easy to understand	Complex	Simple	Simple

Comparison Table for Unsupervised Model

Criterion	K-Means	Hierarchical Clustering
Algorithm Type	Partition-based	Hierarchical
Clustering Process	Iteratively updates cluster centroids	Builds a tree-like structure (dendrogram)
Number of Clusters	Requires pre-specifying (e.g., k=3)	No need to pre-specify
Computational Complexity	Efficient for large datasets	Slower, especially with large datasets
Cluster Shape	Assumes spherical clusters	Can detect non-spherical clusters
Scalability	High scalability with large data	Limited scalability
Interpretability	Less interpretable	More interpretable via dendrogram
Use Case	Simple and scalable segmentation tasks	Exploring hierarchical relationships

Conclusion – Supervised Model

- **Model Performance:**
  - Gradient Boosting achieved the highest accuracy and F1-Score, making it the top-performing model.
  - SVM and Logistic Regression performed well and are viable alternatives.
  - Random Forest showed good results but was slightly less effective.
  - kNN lagged, showing poor suitability for this dataset.

- **Feature Importance**
  - VEHICLE was the most impactful feature, significantly influencing accident risk prediction.
  - Other key features: SLIGH_INJ, ACC_INDEX.
  - Minimal influence from demographic features like TAZAREA and POP_2018.

- **Model Selection Rationale**
  - Recall was prioritized to ensure high-risk areas were correctly classified, minimizing safety risks.
  - Gradient Boosting was selected for its ability to handle complex patterns and outliers effectively.

Conclusion – Unsupervised Model

- **Cluster Analysis**
- **Cluster 0: Balanced Residential Areas**
  - Low accident density, high residential zones.
  - Recommendations: Minor pedestrian safety improvements and regular monitoring.
- **Cluster 1: High-Risk Urban Areas**
  - High accident density, severe injuries, high vehicle involvement.
  - Recommendations: Strict traffic enforcement and infrastructure upgrades.
- **Cluster 2: Large Open-Area Regions**
  - Moderate injury levels, low population density.
  - Recommendations: Enhance long-distance infrastructure and pedestrian safety.

- **Anomalies**

38 anomalies detected across clusters, requiring focused investigation:

  - Cluster 0: 25 anomalies due to temporary hazards.
  - Cluster 1: 9 anomalies indicating exceptionally high-risk zones.
  - Cluster 2: 4 anomalies linked to geographic conditions or outdated infrastructure.

Thank you for listening!