Traffic Accident Hotspot Prediction using Machine Learning

1. Introduction & Motivation

Traffic accidents are a critical global issue, resulting in substantial economic, social, and human losses. Identifying and predicting accident-prone areas, commonly referred to as "hotspots," is essential for enhancing road safety and supporting effective urban planning. Traditional hotspot identification has often been reactive, relying on static statistical methods. However, the rise of machine learning (ML) enables proactive, data-driven predictions using historical crash data, weather conditions, spatial-temporal patterns, and road infrastructure features. This project leverages machine learning models to predict accident hotspots and deliver actionable visual insights to policymakers and urban planners.

2. Problem Definition

The primary objective of this project is to develop a machine learning-based predictive framework that:

- Predicts high-risk geographic zones for traffic accidents.
- Identifies key contributing factors such as time, weather, and road structure.
- Provides interpretable visual analytics, clustering outputs, and trend analysis.

The project also includes a second domain (breast cancer classification) to satisfy the requirement of analyzing two different application domains using ML.

3. Dataset Description

Traffic Accident Dataset:

- Source: Kaggle US Accidents Dataset (2016–2021)
- Initial size: 7.7 million records
- After cleaning: 96,026 high-quality records
- Key Features:
    - Temporal: Start_Time, End_Time, Hour, Day
    - Spatial: Start_Lat, Start_Lng, City, County
    - Environmental: Weather_Condition, Visibility, Temperature, Wind_Speed
    - Road Structure: Junction, Bump, Crossing, Amenity, No_Exit
    - Target: Severity (1 to 4)

Breast Cancer Dataset:

- Source: Kaggle (WDBC dataset)
- 569 rows, 32 features
- Features include radius, texture, smoothness, symmetry, etc.
- Target: Diagnosis (Malignant = 1, Benign = 0)

4. Methodology

4.1 Data Preprocessing

- Removed records with missing coordinates or severity values.
- Dropped irrelevant or high-cardinality columns.
- Applied one-hot encoding for categorical features.
- Used SMOTE to balance class distribution in the Severity variable.

4.2 Classification Models (Traffic Dataset)

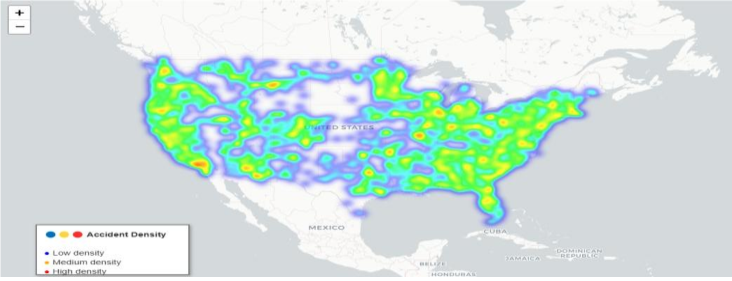
- Trained on 10K and full 96K dataset
- Models Used:
  - Random Forest
  - XGBoost (Final chosen model)
  - Gradient Boosting
  - AdaBoost
- Evaluation Metrics: Accuracy, Precision, Recall, F1-score
- Final XGBoost Accuracy: 67.43%

4.3 Clustering (Hotspot Detection)

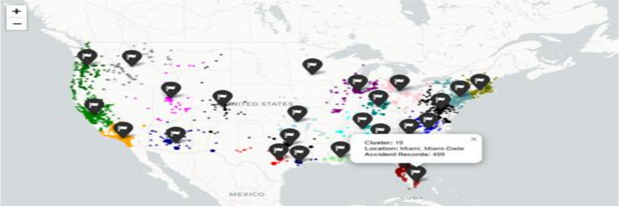
- Used KMeans with 25 clusters on coordinates.
- Clusters named by most frequent City & County.
- Final clustering visualized with Folium maps.

4.4 Visualization Insights (Traffic Dataset)

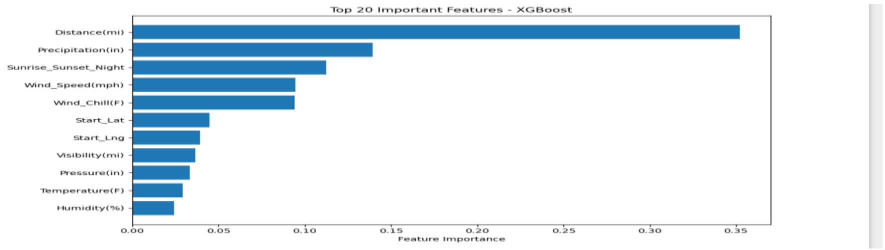
- Heatmaps of accident density across US.



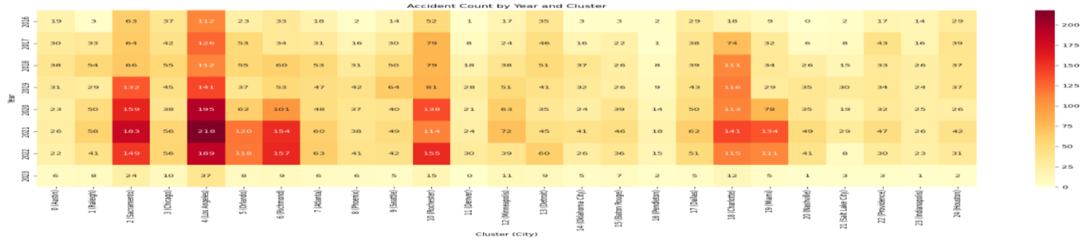
- Marker overlays to identify high-severity accidents.
- Clustering revealed the top 25 accident prone areas for 96k records.



- Feature importance visualized for XGBoost.

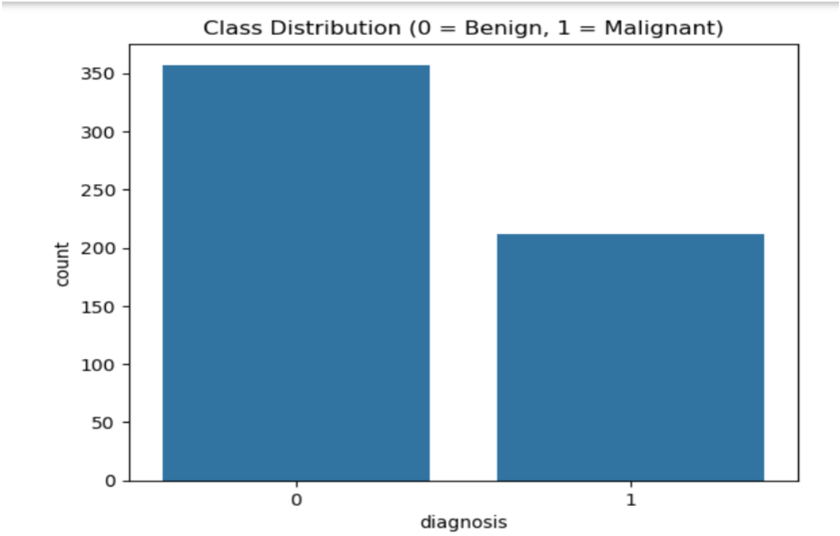


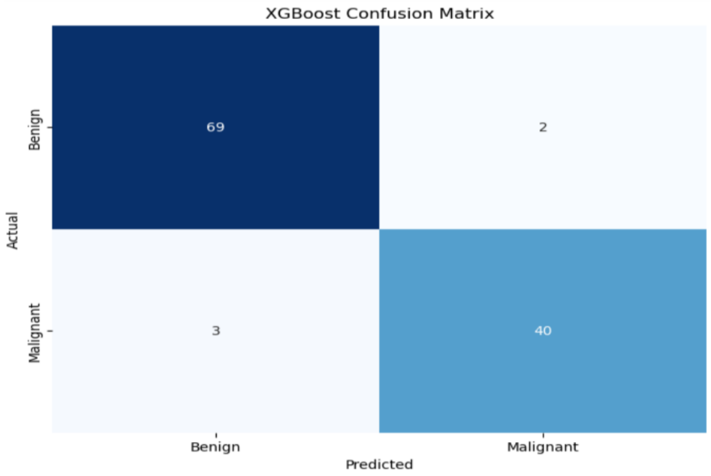
- Time series heatmap: Accident trends across clusters over years.

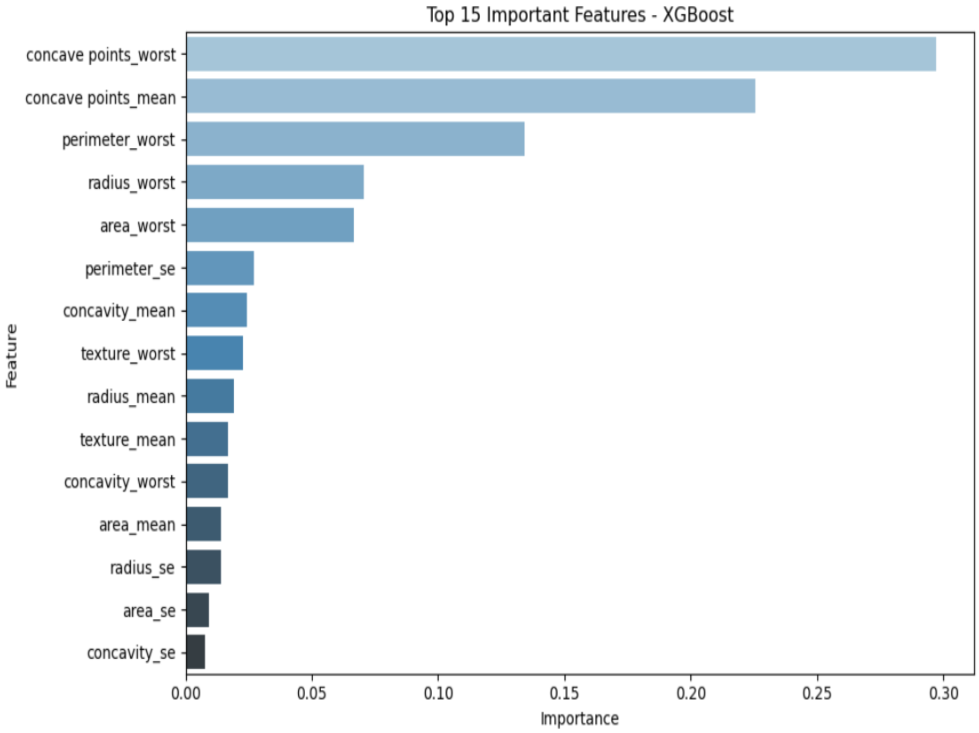


4.5 Breast Cancer Classification (2nd Domain)

- Encoded diagnosis: Malignant (1), Benign (0)
- Models: Random Forest and XGBoost
- Final Accuracy: Random Forest (96.5%), XGBoost (95.6%)
- Confusion matrix heatmap and feature importance plotted.
- EDA: Countplot, Top n features, and correlation heatmaps completed.







5. Results & Performance Analysis

Accidents Dataset:

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	57.72%	0.74	0.58	0.63
XGBoost (10K)	73.36%	0.75	0.73	0.74
XGBoost (Full)	69.57%	0.70	0.69	0.69
AdaBoost	32.52%	0.74	0.33	0.38
Gradient Boosting	52.76%	0.77	0.53	0.60

Breast Cancer dataset:

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	96.49%	0.97	0.93	0.95
XGBoost	95.61%	0.95	0.93	0.94

Efficiency:

- Full XGBoost model trained in ~42 seconds on 96K dataset.
- Clustering runtime: ~2.5 seconds
- GridSearchCV tuning for XGBoost took ~90+ seconds

6. Conclusion and Recommendations

This project successfully implemented a predictive system to identify accident-prone areas using machine learning techniques. The system combined classification models (with XGBoost showing the best results) and spatial clustering (KMeans) to detect and visualize traffic hotspots. Interactive maps and feature-based dashboards offered powerful visual tools for urban safety planners.

The inclusion of a second domain (breast cancer classification) further demonstrated the model's adaptability and the effectiveness of ML pipelines across domains.

Recommendations & Future Work:

- Improve spatial granularity by increasing cluster count or using DBSCAN.
- Build a lightweight real-time dashboard for public/authority access.
- Add new data dimensions like traffic volume, construction zones, or live feeds.
- Apply hyperparameter tuning with GridSearchCV for further gains.

7. References

- Kaggle US Accidents Dataset: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents
- Scikit-learn: https://scikit-learn.org
- XGBoost Documentation: https://xgboost.readthedocs.io
- Imbalanced-learn Toolkit: https://imbalanced-learn.org
- Folium Map Library: https://python-visualization.github.io/folium
- Breast Cancer Dataset (WDBC): https://www.kaggle.com/uciml/breast-cancer-wisconsin-data