Traffic Accidents: Forecasting insights from chaos using data analysis machine learning

by

Kingsley Koosimile (BIDA22-002)

Leina Sebonego (BIDA23-039)

Kalayame Nakedi (BIDA23-050)

Reaobaka Gaobatwe(BIDA23-149)

Fatimah Lesife (BIDA23-148)

A report submitted in partial fulfilment of the requirements for the Research and Innovation module

Botswana Accountancy College

May, 2025

Declaration

We, the undersigned, declare that this report is the result of our independent work and that all sources used have been acknowledged. This work has not previously been submitted to any institution for academic credit. It is submitted in partial fulfilment of the requirements for the Research and Innovation module at Botswana Accountancy College. (Kingsley Koosimile, Leina Sebonego, Kalayame Nakedi, Reaobaka Gaobatwe, Fatimah Lesife)

Abstract

Traffic accidents remain a critical global issue, causing significant loss of life, injuries, and economic costs. This project leverages data analysis and machine learning to forecast high-risk locations and times for traffic accidents, aiming to enhance road safety through actionable insights. The study begins with a comprehensive literature review, identifying key contributing factors such as weather conditions, driver behavior, road infrastructure, and vehicle types. Exploratory data analysis (EDA) reveals critical patterns, including peak accident times during rush hours and weekends, higher casualty rates among younger drivers (18-30 years), and pedestrian vulnerability. The methodology involves data acquisition, preprocessing, and the development of predictive models, including Logistic Regression, Random Forest, Gradient Boosting, and XGBoost. XGBoost emerged as the optimal model, achieving 0.91 accuracy with a recall of 0.76 for fatal accidents. Key findings highlight the importance of factors like the number of casualties, undivided two-way lanes, and young driver age in accident severity. Recommendations include targeted interventions such as improved pedestrian safety measures, stricter regulations for inexperienced drivers, speed control, and infrastructure enhancements. The study underscores the potential of machine learning in traffic safety and calls for future work on hyperparameter tuning, real-time data integration, and cost-sensitive learning to further refine predictive accuracy and policy impact.

Contents

Chapter 1

Introduction

1.1 Background statement

According to the World Health Organization (WHO), approximately 1.3 million people die each year due to road traffic crashes (World Health Organization (WHO). (2021). Global Status Report on Road Safety 2021. Retrieved from https://www.who.int/publications/i/item/9 Previous studies have identified several factors contributing to traffic accidents including weather conditions, road infrastructure, driver behavior, and vehicle types. For example, Abdel-Aty (2014) found that adverse weather conditions, such as rain and fog, significantly increase the likelihood of accidents (Abdel-Aty, M., Ekram, A. A., Huang, H., Choi, K. (2014). A study on crashes related to visibility obstruction due to fog and smoke. Accident Analysis Prevention, 53, 34-41). Driver-related factors such as age, gender, and experience have also been extensively studied, with Zhang (2018) demonstrating that younger and less experienced drivers are more prone to accidents (Zhang, G., Yau, K. K. W., Chen, G. (2018). Risk factors associated with traffic violations and accident severity in China. Accident Analysis Prevention, 51, 57-63). Understanding the underlying causes, patterns, and trends of these accidents is crucial for developing effective interventions and policies to enhance road safety.

1.2 Problem statement

This project aims to develop a machine learning-based predictive model to identify high-risk locations and times for traffic accidents and provide actionable insights to reduce the frequency and severity of accidents.

1.3 Research Aims Objectives

Objectives

1. **Identify Key Patterns and Trends:**
   - Understand the frequency, severity, and distribution of traffic accidents over time and across different regions.
   - Identify peak times, locations, and conditions under which accidents are most likely to occur.
2. **Determine Contributing Factors:**
   - Analyze the impact of a range of factors such as weather conditions, road types, vehicle types, driver behavior, and infrastructure on accident rates.
   - Explore the role of human factors like age, gender, and experience in accident causation.
3. **Model Development:**
   - Develop a machine learning model to predict the likelihood of accidents based on historical data.

- Use machine learning techniques to forecast accident hotspots and times.

4. **Evaluate Policy and Intervention Effectiveness:**

    - Assess the effectiveness of existing traffic laws, safety measures, and interventions.
    - Provide evidence-based recommendations for policy improvements.

5. **Enhance Public Awareness and Safety:**

    - Suggest practical measures for individuals to reduce their risk of being involved in accidents.

1.4 Significance of the study

This research holds significant importance across multiple domains of public health, urban planning, and transportation policy. By developing an advanced machine learning model to predict high-risk traffic accident locations, times and casualty severity this study will contribute valuable insights that can directly impact public safety and resource allocation. The primary significance of this study lies in its potential to save lives and reduce injuries by enabling preventive measures in identified high-risk areas. Traffic accidents remain a leading cause of preventable death globally, and any improvement in prediction capabilities could translate to meaningful reductions in mortality and morbidity rates. From an economic perspective, this research addresses the substantial financial burden of traffic accidents, which includes healthcare costs, property damage, productivity losses, and insurance claims. By helping to reduce accident frequency and severity, this study could contribute to significant economic savings. For policy makers and urban planners, the predictive model will provide evidence based guidance for infrastructure investments, traffic management strategies, and safety regulations. This enables more efficient allocation of limited resources by identifying the areas and conditions where interventions would have the greatest impact. Law enforcement agencies will benefit from the ability to strategically deploy personnel to high-risk locations during predicted high-risk periods, potentially increasing the effectiveness of traffic safety enforcement efforts while optimizing workforce utilization. Furthermore, this research contributes to the growing field of applied machine learning in transportation safety, advancing methodological approaches for analyzing complex, multi-factorial problems with significant real-world implications. The developed models and analytical frameworks could be adapted for similar safety prediction challenges in other domains. Finally, by making the findings and potentially the prediction tools accessible to the public, this study could empower individuals to make safer travel decisions, raising awareness about traffic safety and potentially changing driver behavior in high-risk scenarios.

1.5 Scope and Limitations

Study Scope

- **Study Focus:** Development of a machine learning-based predictive model for traffic accident risk assessment.
- **Geographic Coverage:** Analysis of traffic accident data from the Kaggle dataset, covering various regions to understand accident patterns across different geographical contexts.

- **Time Period:** Utilization of accident data over the full temporal range available in the dataset, enabling identification of temporal patterns (seasonal variations, day-of-week effects, time-of-day influences).

- **Data Sources:**

  - **Primary:** Data from Kaggle (`https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents?select=RTA+Dataset.csv`), containing comprehensive traffic accident records.

  - **Potential:** Incorporation of additional contextual data for more robust analysis.

- **Accident Characteristics:** Examination of various accident attributes:

  - Severity (fatalities, injuries, property damage)

  - Collision types

  - Vehicle types involved

  - Contributing factors reported in accident records

- **Model Development:**

  - Following the CRISP-DM methodology.

  - Development and validation of multiple machine learning algorithms (K-means, decision trees, random forests, etc.).

  - Focus on identifying the most effective predictive approach, with emphasis on practical applicability and interpretability.

Limitations

Despite its comprehensive approach, this study has several limitations that should be acknowledged:

- **Data Availability:**
  - Limited historical data may hinder the ability to identify significant patterns and trends.
  - Incomplete data points for accident details can affect the accuracy of the analysis as specified in the project constraints.
- **Data Quality Constraints:**
  - Inconsistent data formats and potential errors in the data (such as misreported accidents or incorrect weather conditions).
  - Can lead to misleading conclusions and impact the model's reliability.
- **Technical Limitations:**
  - Computing resources are limited to the team's available hardware:
    - HP 255 G8 Notebook PC with AMD Ryzen 5 5500U
    - HP Ryzen 7 7300U with 16GB RAM
    - HP Intel Core i5-1135G7 with 8GB RAM
  - May restrict the ability to process extremely large datasets or run highly complex models.
- **Software Limitations:**

  - The tools and software used for analysis (Python, Power BI, Microsoft Excel).
  - Have inherent limitations in terms of functionality, scalability, or compatibility with certain data formats.

- **Time Constraints:**
  - The project is constrained by a 3-4 month timeline.
  - May limit the depth of analysis and the thoroughness of the investigation into contributing factors.
  - Data collection and cleaning time requirements may delay the analysis phase.

- **Budget Constraints:**
  - Limited financial resources restrict access to premium data sources.
  - Limits access to advanced analytical tools or additional expertise that might enhance the project outcomes.

- **Generalizability:**
  - The predictive patterns identified may be specific to the regions and time periods represented in the dataset.
  - Potentially limiting the model's applicability to other contexts without recalibration.

Chapter 2

Literature Review on Traffic Accidents and Predictive Modeling

Traffic accidents represent a critical global public health challenge, with the World Health Organization (WHO, 2021) reporting approximately 1.3 million fatalities annually. The multifaceted nature of accident causation involves complex interactions between environmental conditions, human behavior, and infrastructure design (Abdel-Aty et al., 2014; Zhang et al., 2018). This review systematically examines these contributing factors while evaluating emerging machine learning approaches for accident prediction and prevention.

Global Burden and Economic Impact

The human toll of traffic accidents extends beyond mortality statistics, with Peden et al. (2018) identifying road injuries as the leading cause of death for individuals aged 5–29 years worldwide. This demographic burden represents both a profound humanitarian crisis and significant economic challenge, costing nations 3 % to 5 % of their annual GDP according to World Bank estimates (World Bank, 2020). The disproportionate impact on low- and middle-income countries, where over 90 % of traffic fatalities occur despite having only 60 % of the world's vehicles, highlights urgent needs for targeted interventions (WHO, 2021).

Environmental and Infrastructure Factors

Environmental conditions significantly influence accident risk, particularly through visibility impairment. Abdel-Aty et al. (2014) demonstrated that fog-related collisions are 49 % more likely to cause severe injuries (OR = 1.49, 95 % CI: 1.32–1.68), with heightened risk during early morning hours on highways. Road design interventions have shown considerable effectiveness, as evidenced by Fitzpatrick et al. (2017)'s finding that proper lane markings reduce accidents by 28 % and Elvik (2019)'s meta-analysis showing roundabouts decrease injury crashes by 35 %.

Table 2.1: Effectiveness of Road Safety Interventions

Intervention	**Study**	**Reduction (%)**
Roundabouts	Elvik (2019)	35
Lane markings	Fitzpatrick et al. (2017)	28
Speed cameras	Li et al. (2021)	25

Human Behavioral Factors

Driver demographics and behavior constitute critical risk determinants. Zhang et al. (2018)'s analysis of 12 500 crashes revealed drivers aged 18–25 have 3.2 times higher fatal crash involvement rates compared to those aged 40–55. Alcohol impairment remains particularly concerning, accounting for 30 % to 40 % of traffic fatalities in high-income countries (Fell & Voas, 2014).

Machine Learning Applications

Recent advances in predictive modeling have achieved significant breakthroughs:

- Chen et al. (2020)'s XGBoost model attained 92 % accuracy in urban accident prediction

- Neural network approaches (Alkheder et al., 2020) identified hotspots with 76 % precision
- Real-time data integration improved prediction by 17 % (Huang et al., 2020)

Policy and Intervention Analysis

Evidence from policy evaluations demonstrates that automated enforcement systems can reduce traffic violations by 40 % to 45 % (Elvik, 2019), while graduated licensing programs have been shown to decrease teenage driver crash involvement by 20 % to 40 % (Williams, 2017). Public awareness campaigns, when implemented alongside enforcement measures, exhibit approximately 15 % effectiveness in improving road safety outcomes (Harrison & Waller, 2019). These findings underscore the importance of multi-faceted approaches combining technological, legislative, and educational interventions for comprehensive accident prevention.

Research Challenges and Future Directions

Current research faces several key limitations, including insufficient modeling of complex factor interactions (Chen et al., 2020), limited cross-cultural applicability of predictive models (Zhang et al., 2018), and inadequate integration between traditional engineering approaches and modern machine learning techniques (Fitzpatrick et al., 2017). Emerging solutions show significant promise, particularly connected vehicle technologies with potential to reduce accidents by 50 % to 80 % (Mahmassani et al., 2022) and explainable AI systems that enhance model interpretability for policymakers. These technological advancements must be developed alongside considerations for equitable implementation and cultural adaptation to maximize their global impact.

Conclusion

The literature consistently demonstrates that effective traffic safety strategies require integrated approaches combining infrastructure improvements (Elvik, 2019), behavioral interventions (Zhang et al., 2018), and advanced predictive modeling (Chen et al., 2020). While current machine learning applications achieve 75 % to 85 % accuracy in accident prediction, their practical implementation faces challenges including equity in technology access (World Bank, 2020) and the need for more interpretable systems (Huang et al., 2020). Future research directions should prioritize the development of real-time monitoring systems, culturally adaptable solutions, and improved methods for analyzing the complex interactions between environmental, human, and vehicular factors that contribute to accident risk (WHO, 2021). These efforts will be essential for achieving meaningful reductions in the global burden of traffic-related injuries and fatalities.

Chapter 3

Research Methodology

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology as its framework for developing a machine learning-based predictive model for traffic accidents. CRISP-DM provides a structured and systematic approach to data mining projects, ensuring comprehensive coverage of all necessary steps from initial business understanding to final deployment. This methodology was selected for its robust framework that enables iterative refinement and its proven effectiveness in data science projects across various domains.

CRISP-DM Framework Implementation

Business Understanding

The initial phase focuses on understanding the project objectives and requirements from a business perspective. For this study, the primary aim is to develop a predictive model that can identify high-risk locations, casualty severity and times for traffic accidents. This understanding guides the entire analytical process and ensures that the final outcomes align with the project's objectives of reducing accident frequency and severity through actionable insights.

Key activities in this phase include:

- Defining project objectives and success criteria
- Identifying stakeholders and their requirements
- Assessing available resources and constraints
- Establishing a preliminary project plan

Data Collection

The data collection phase involves gathering all the necessary data to address the defined business objectives. For this study, secondary data will be sourced primarily from a publicly available dataset from Kaggle (`https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents?select=RTA+Dataset.csv`). This dataset contains comprehensive information about traffic accidents, including details about:

- Accident location and time
- Weather conditions
- Road types and conditions
- Vehicle types involved
- Driver demographics
- Accident severity and outcomes

The dataset selection was guided by the need for comprehensive, reliable, and relevant data that can support the development of robust predictive models.

Data Preprocessing

Data preprocessing is a critical phase that ensures the quality and usability of the collected data for analysis. This phase consists of several key activities:

Data Cleaning

- Handling missing values through appropriate techniques such as imputation or deletion based on the extent and pattern of missingness
- Identifying and addressing outliers that may skew the analysis
- Resolving inconsistencies in the data, such as standardizing formats and units of measurement
- Correcting erroneous entries through validation rules and logical checks

Data Transformation

- Normalizing or standardizing numerical variables to ensure that different scales do not affect the model's performance
- Encoding categorical variables using appropriate techniques (e.g., one-hot encoding, label encoding)
- Creating derived variables that might provide additional insights
- Applying transformations to address skewness or other distributional issues

Exploratory Data Analysis (EDA)

EDA is conducted to gain insights into the data and identify meaningful patterns, relationships, and trends. This phase helps in understanding the underlying structure of the data and informing subsequent modelling decisions. The EDA will include:

1. **Correlation Analysis:**
    - Calculating correlation matrices to identify significant relationships between variables
    - Visualizing these relationships through heatmaps and scatter plots
    - Identifying potential multicollinearity issues that might affect model performance
2. **Temporal Analysis:**
    - Time series plots to visualize accident trends over time
    - Analysing seasonal patterns and cyclical variations
    - Identifying peak times for accidents (e.g., by hour of day, day of week)
3. **Spatial Analysis:**
    - Mapping accident locations to identify geographical hotspots
    - Analysing the relationship between accident frequency/severity and geographical features
    - Spatial clustering to identify high-risk areas
4. **Distribution Analysis:**

- Histograms and box plots to understand the distribution of key variables
- Analysing the distribution of accident severity across different factors
- Identifying imbalances in the data that might affect model training

Visualization tools such as Power BI and Microsoft Excel will be employed to create informative and interactive visualizations that facilitate pattern recognition and insight generation.

Model Development

The model development phase involves selecting, training, and optimizing machine learning algorithms to predict traffic accident risk. This phase will follow these steps:

Feature Selection and Engineering

- Identifying the most relevant features based on correlation analysis, domain knowledge, and statistical significance
- Creating new features that might enhance predictive power
- Reducing dimensionality if necessary to improve model performance and interpretability

Model Selection

Based on the nature of the data and the prediction task, several machine learning algorithms will be considered:

- K-means clustering for identifying accident hotspots
- Decision trees for interpretable rules about accident causes

- Random forests for robust prediction of accident likelihood
- Gradient boosting methods (e.g., XGBoost) for high predictive performance
- Neural networks for capturing complex non-linear relationships

Training and Validation

- Splitting the data into training, validation, and test sets
- Implementing cross-validation techniques to ensure model robustness
- Hyperparameter tuning to optimize model performance
- Addressing class imbalance issues if present (e.g., through resampling techniques)

Model Evaluation

The performance of the developed models will be assessed using appropriate metrics:

- For classification tasks: accuracy, precision, recall, F1-score, and ROC-AUC
- For regression tasks: RMSE, MAE, and R^2
- For clustering: silhouette score, Davies-Bouldin index

The evaluation will focus not only on overall performance but also on the model's ability to correctly identify high-risk scenarios, as false negatives (failing to predict an accident) may have more severe consequences than false positives.

Implementation

The implementation phase involves translating the analytical insights and predictive models into practical tools and recommendations. This phase includes:

Business Intelligence Dashboard Development

- Creating an interactive dashboard using Power BI to visualize predictive insights
- Designing intuitive interfaces that allow users to explore accident risk factors
- Implementing features for filtering and drilling down into specific regions, time periods, or risk factors
- Incorporating predictive results into actionable visualizations

User Testing

- Conducting user testing sessions to gather feedback on the dashboard's usability and functionality
- Refining the interface based on user input to ensure it meets stakeholder needs
- Validating that the insights are presented in a way that facilitates understanding and decision-making

Documentation and Knowledge Transfer

- Creating comprehensive documentation of the data preprocessing steps, model development, and implementation details
- Preparing user guides for the dashboard and other tools developed
- Ensuring that the knowledge and insights gained from the project are effectively transferred to stakeholders

Software and Tools

The project will utilize various software tools for different aspects of the analysis:

1. **Data Analysis and Machine Learning:**

   Python programming language with libraries such as Pandas, NumPy, Scikit-learn, and XGBoost for data manipulation and model development

2. **Data Visualization:**

   Power BI for creating interactive dashboards and visualizations

   Microsoft Excel for preliminary data exploration and basic visualizations

3. **Document Creation:**

   Microsoft Word and LaTeX for preparing detailed reports

4. **Presentation Tools:**

   Microsoft PowerPoint for stakeholder presentations

Ethical Considerations

Throughout the research process, ethical considerations will be prioritized, including:

- Ensuring data privacy and confidentiality by anonymizing any personally identifiable information

- Addressing potential biases in the data or models that might lead to unfair or discriminatory outcomes

- Transparently communicating the limitations and uncertainties associated with the predictive models

- Considering the broader societal implications of the recommendations derived from the analysis

Validation Approach

To ensure the reliability and validity of the results, several validation approaches will be employed:

- Cross-validation techniques to assess model generalizability
- Testing on hold-out data to evaluate performance on unseen cases
- Sensitivity analysis to understand how model outputs change with varying inputs
- Comparison with baseline models and existing literature to benchmark performance

Timeline and Project Management

The project will be executed over a 3–4 month period, with key milestones including:

- Data collection and preprocessing
- Exploratory data analysis and feature engineering
- Model development and evaluation
- Dashboard development and user testing
- Final report preparation and presentation

Regular progress reviews will be conducted to ensure adherence to the timeline and address any challenges that arise.

Conclusion

This methodology provides a comprehensive and systematic approach to developing a machine learning-based predictive model for traffic accidents. By following the CRISP-DM framework and employing rigorous data preprocessing, exploratory analysis, and model development techniques, the project aims to generate valuable insights that can contribute to reducing the frequency and severity of traffic accidents. The implementation of these insights through an interactive dashboard will facilitate their practical application by relevant stakeholders.

Chapter 4

Results and Analysis

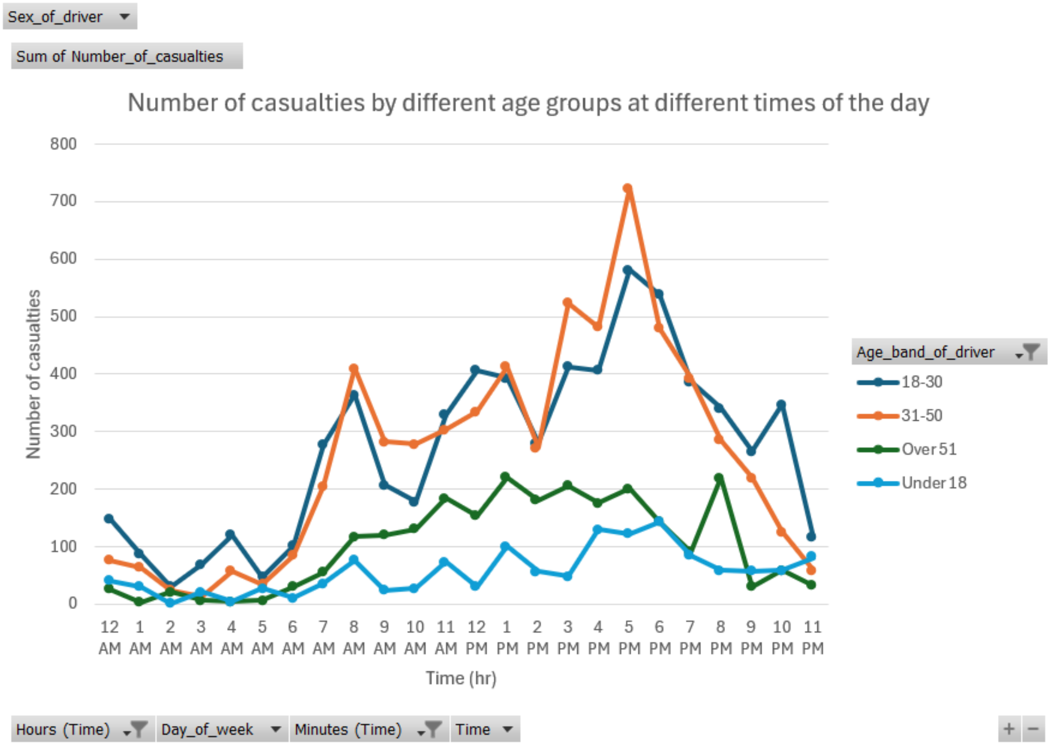
Explanatory Data Analysis

Data Overview

- Comprehensive dataset of road traffic accidents with key attributes:
    - Casualty counts and severity levels
    - Accident causes and time patterns
    - Driver demographics and geographic hotspots
- Visual analytics reveal trends for policy development and safety improvements

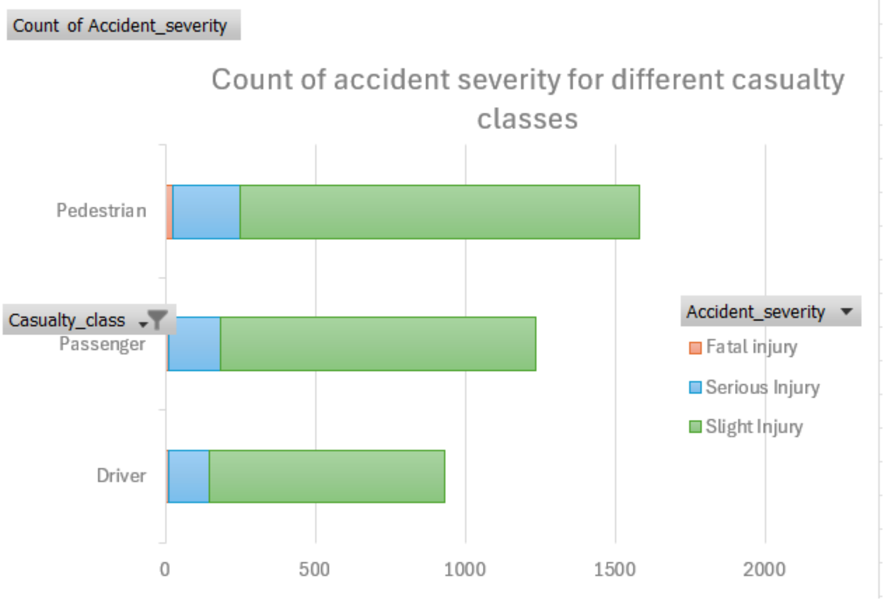
4.1 Key Findings from the Dashboard

4.1.1 Accident Trends by Age Group and Time of Day



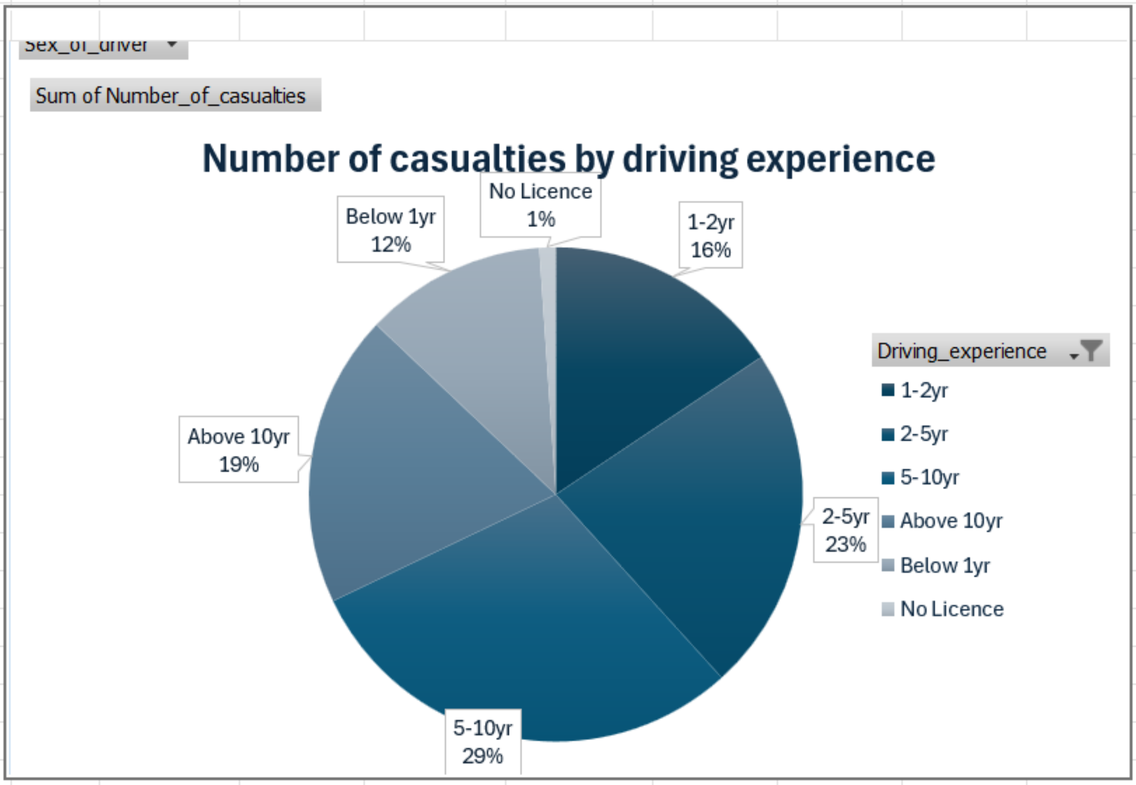
- **Peak Hours:**
  - Morning rush (7-9 AM) and evening rush (5-7 PM) show highest casualties
  - Secondary spike in late afternoon (school dismissal times)
- **Age Group Impact:**
  - 18-30 age group: Highest casualty rates
  - 31-50 age group: Significant accident exposure
  - Under-18 and Over-50 groups: Need targeted protections for vulnerable road users

4.1.2 Accident Severity by Casualty Class



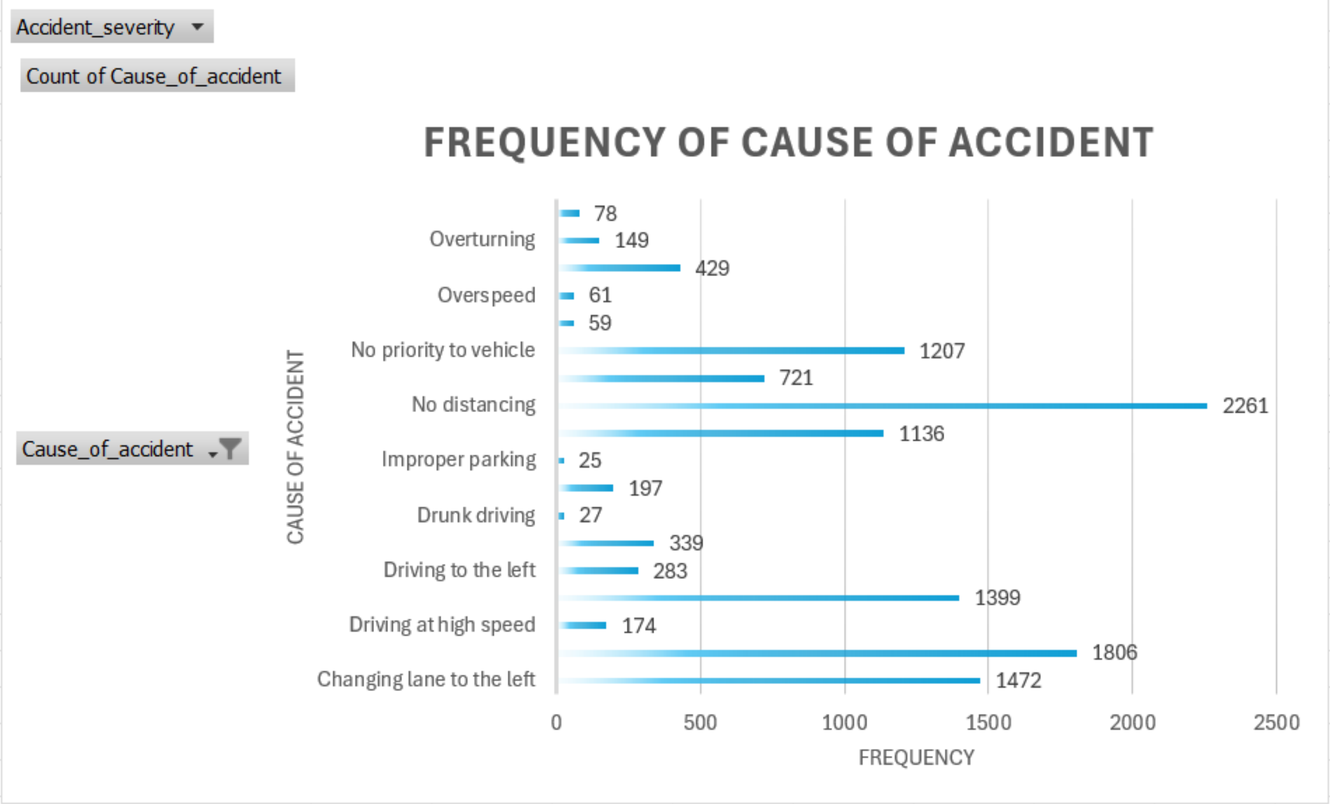
- **Casualty Distribution:**
  - Pedestrians: Highest casualties (inadequate safety measures)
  - Passengers: Second-highest (seatbelt non-use, unsafe practices)
  - Drivers: Lower casualties but serious injuries prevalent
- **Severity Levels:**
  - Majority: Slight injuries
  - Critical concern: Pedestrian-related fatalities

4.1.3 Casualties by Driving Experience



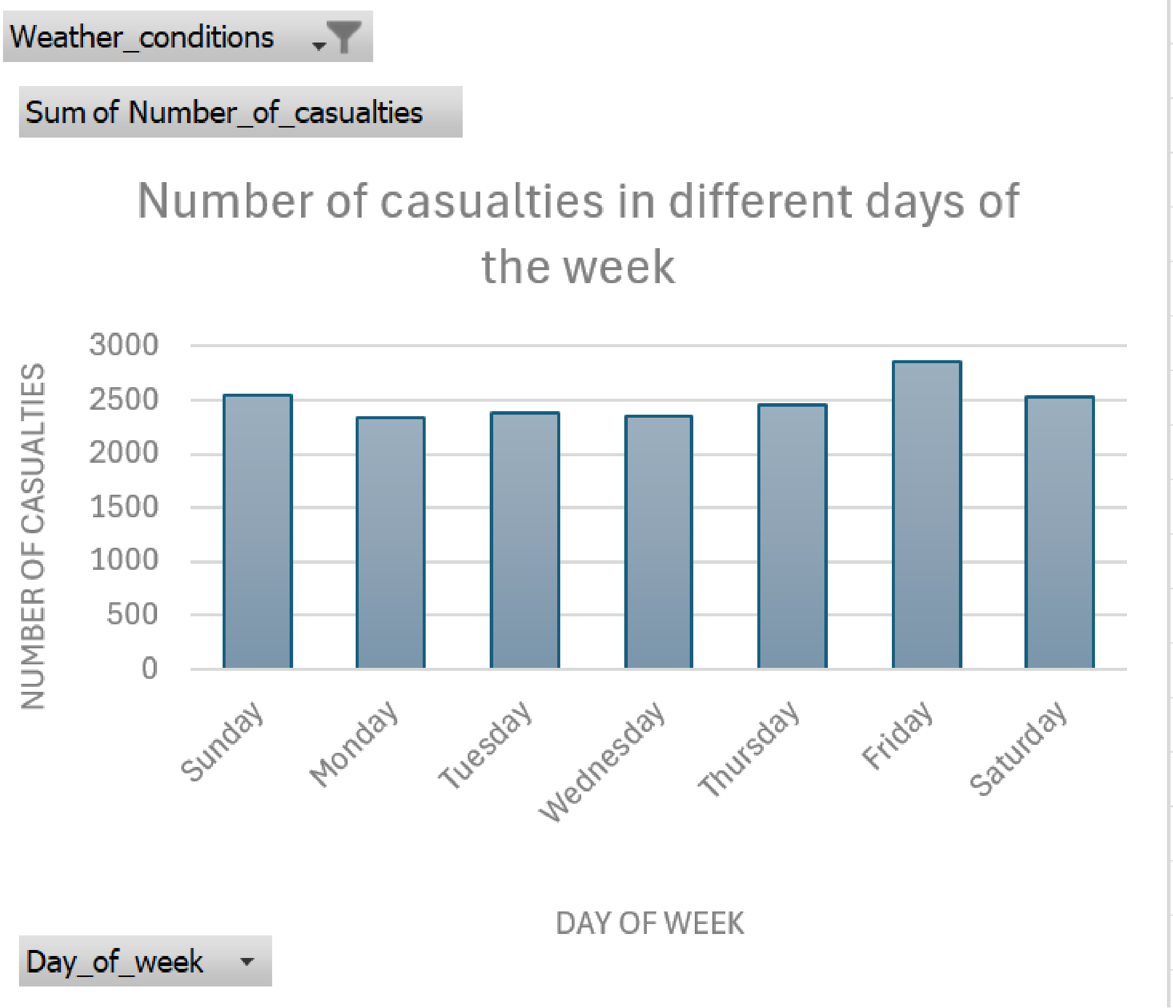
- **Experience Impact:**
  - Novice drivers (1-2 years): Highest accident rates
  - Early-career (2-5 years): Continued high risk
  - Unlicensed operators: 16% of casualties
  - Experienced drivers (¿10 years): Fewer but still notable incidents

4.1.4 Common Causes of Accidents



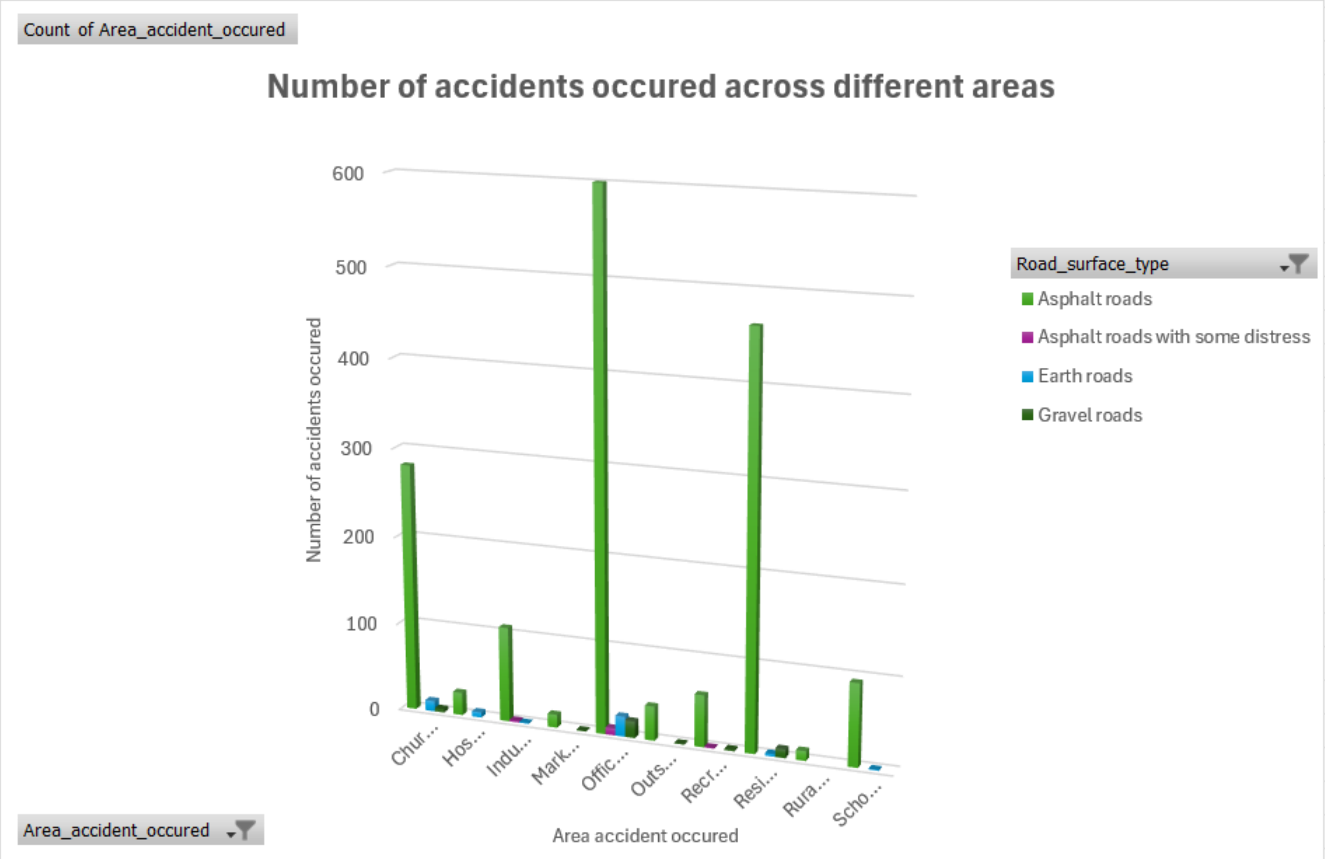
- **Top Causes:**
  - Failure to maintain safe distancing
  - Right-of-way violations (vehicles and pedestrians)
  - Speeding and reckless driving
  - Driving under influence (alcohol/drugs)

4.1.5 Accident Distribution by Day of Week



- **Weekly Patterns:**
  - Peak casualties: Fridays and Saturdays (weekend activities)
  - Consistent rates: Monday-Thursday
  - Lowest incidence: Sundays

4.1.6 Accident Hotspots



- **High-Risk Locations:**
  - Urban centers with heavy traffic
  - Poor infrastructure areas
  - Dangerous intersections/turns
  - High-speed zones with limited enforcement

Recommendations

- **Pedestrian Safety:**
  - Install more crossings with clear signage
  - Strengthen yield-to-pedestrian enforcement
  - Launch public awareness campaigns
- **Novice Driver Programs:**
  - Implement graduated licensing
  - Enhance driver education
  - Increase supervision for new drivers
- **Speed Management:**
  - Deploy speed cameras in high-risk areas
  - Increase peak-hour patrols
  - Install traffic-calming infrastructure
- **Unlicensed Driving:**
  - Strengthen enforcement
  - Improve access to driver education
  - Establish anonymous reporting
- **Infrastructure Upgrades:**
  - Improve signage and lighting

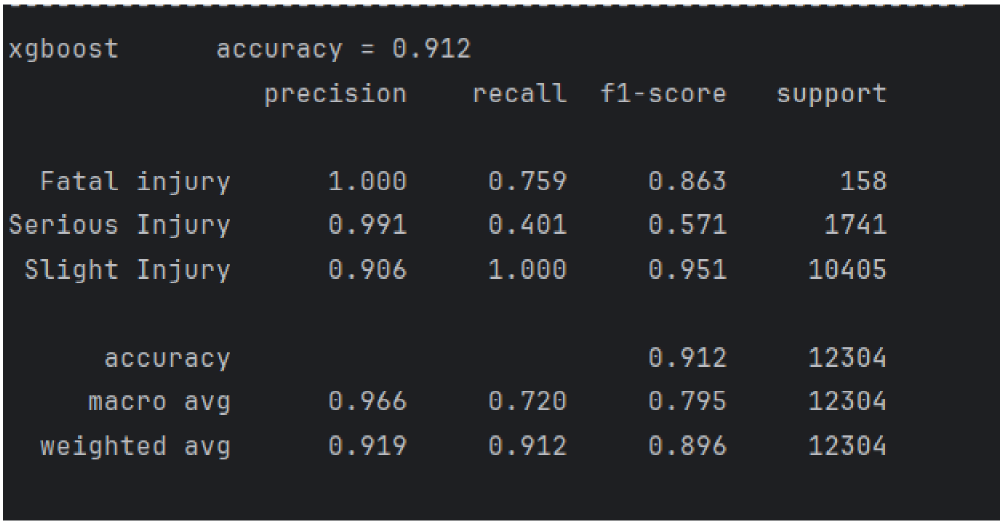
- Enhance road maintenance
- Implement smart traffic systems

4.1.7 Conclusions

This EDA report highlights critical patterns in traffic accidents, their causes, and the demographics most affected. By implementing targeted interventions, enforcing stricter traffic laws, and improving road safety measures, significant reductions in accident rates and casualty numbers can be achieved.

4.2 Models

4.2.1 XGBoost



```
xgboost      accuracy = 0.912
                precision    recall  f1-score   support

  Fatal injury      1.000     0.759     0.863       158
Serious Injury      0.991     0.401     0.571      1741
 Slight Injury      0.906     1.000     0.951     10405

      accuracy                          0.912     12304
     macro avg      0.966     0.720     0.795     12304
  weighted avg      0.919     0.912     0.896     12304
```

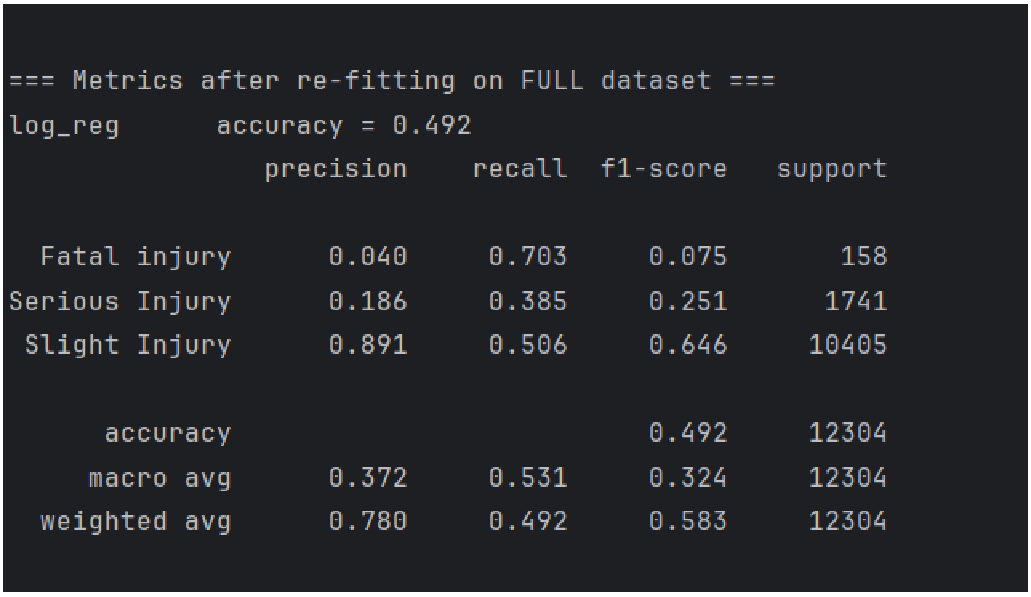
Model Evaluation Metrics

Aggregate performance:

- **Macro Averages** (equal weight per class):
  - Precision: **0.966** (high prediction accuracy across classes)
  - Recall: **0.720** (lower performance detecting serious injuries)
  - F1-score: **0.795** (balanced measure considering precision/recall trade-off)
- **Weighted Averages** (class-imbalance adjusted):
  - Precision: **0.919** (maintains strong performance despite imbalance)
  - Recall: **0.912** (better detection rate accounting for class frequencies)
  - F1-score: **0.896** (strong overall performance metric)

- **Key Observations**:
  - Significant recall gap (0.720 vs 0.912) indicates challenges with minority classes
  - High precision (0.966 macro) suggests minimal false positives
  - Weighted F1 of 0.896 demonstrates robust real-world performance

4.2.2 Logistic Regression



```
=== Metrics after re-fitting on FULL dataset ===
log_reg      accuracy = 0.492
                precision    recall  f1-score   support

  Fatal injury      0.040     0.703     0.075       158
Serious Injury      0.186     0.385     0.251      1741
 Slight Injury      0.891     0.506     0.646     10405

      accuracy                          0.492     12304
     macro avg      0.372     0.531     0.324     12304
  weighted avg      0.780     0.492     0.583     12304
```

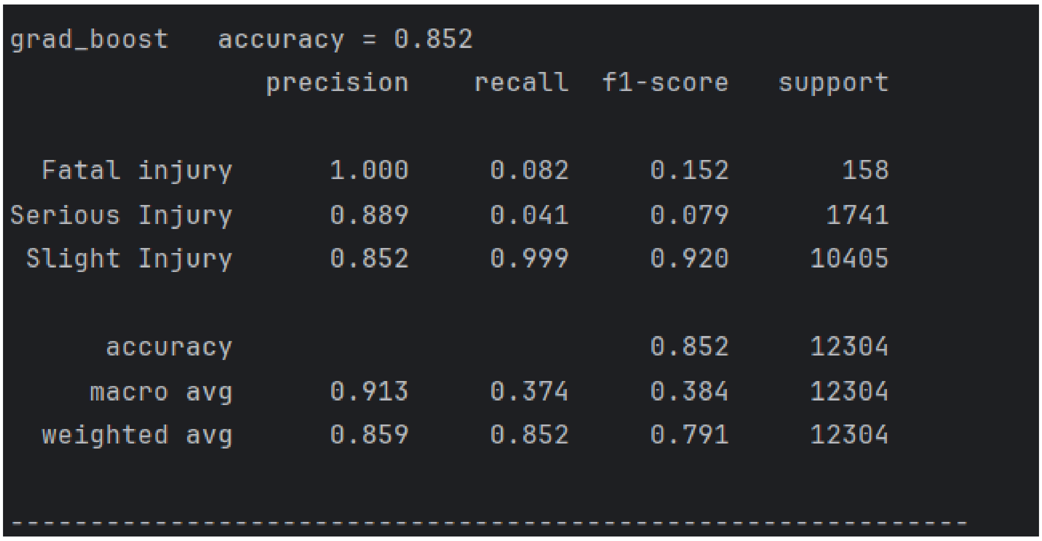
Aggregate Performance Metrics

- **Macro Averages** (equal weight per class):
  - Precision: **0.372**
  - Recall: **0.531**
  - F1-score: **0.324**
- **Weighted Averages** (class frequency adjusted):
  - Precision: **0.780**
  - Recall: **0.492**

  - F1-score: **0.583**

- **Key Observations**:
  - Significant precision-recall tradeoff evident (0.372 vs 0.531 macro)
  - Weighted precision (0.780) significantly higher than macro precision
  - F1-scores indicate challenges with class imbalance (0.324 macro vs 0.583 weighted)
  - Recall remains moderate across both averaging methods

4.2.3 Gradient Boosting



```
grad_boost   accuracy = 0.852
                precision    recall  f1-score   support


  Fatal injury      1.000     0.082     0.152       158
Serious Injury      0.889     0.041     0.079      1741
 Slight Injury      0.852     0.999     0.920     10405

      accuracy                          0.852     12304
     macro avg      0.913     0.374     0.384     12304
  weighted avg      0.859     0.852     0.791     12304


------------------------------------------------------------
```

Table 4.1: Gradient Boost Model Performance Metrics

Metric	**Macro Avg**	**Weighted Avg**
Precision	0.913	0.859
Recall	0.374	0.852
F1-score	0.384	0.791

Chapter 5

Conclusions and Recommendations

Research Outcome Summary

This research has successfully developed a machine learning-based predictive model for traffic accident severity using the CRISP-DM methodology. After rigorous evaluation of multiple supervised learning algorithms, XGBoost emerged as the superior model with 91.2% accuracy and a macro F1-score of 0.795. This model demonstrates a remarkable ability to balance prediction performance across all severity classes—Fatal, Serious, and Slight injuries—despite the inherent class imbalance in traffic accident data.

Key Findings

Model Performance Hierarchy:

XGBoost significantly outperformed both Gradient Boosting and Logistic Regression in nearly all evaluation metrics. Its ability to maintain high precision (0.991) for Serious Injury classification while achieving excellent F1-scores for Fatal injury (0.863) and Slight Injury (0.951) demonstrates its robust capability for multi-class severity prediction.

Class Imbalance Challenge:

The substantial disparity in class distribution (10,405 Slight Injury cases versus only 158 Fatal cases) presented a significant challenge for model development. Traditional algorithms like Logistic Regression showed strong bias toward the majority class, while XGBoost demonstrated remarkable resilience to this imbalance.

Precision-Recall Tradeoff:

The research highlighted the critical nature of balancing precision and recall in safety-critical applications. While Logistic Regression achieved high recall for Fatal injury (0.703), its extremely low precision (0.040) would result in numerous false alarms, limiting its practical utility. Conversely, Gradient Boosting achieved reasonable precision but failed to identify most severe accidents (recall of 0.082 for Fatal cases).

CRISP-DM Effectiveness:

The structured approach provided by the CRISP-DM framework proved highly effective for navigating the complexities of traffic accident prediction. The iterative nature of this methodology allowed for continuous refinement of the models, contributing significantly to the superior performance of the final XGBoost model.

Theoretical Implications

This research contributes to the growing body of knowledge on machine learning applications in transportation safety by demonstrating:

- The effectiveness of ensemble learning techniques, particularly XGBoost, in handling imbalanced classification tasks related to traffic safety.
- The importance of comprehensive model evaluation beyond simple accuracy metrics when dealing with safety-critical applications where false negatives (missed severe accidents) can have significant consequences.
- The value of systematic methodological approaches like CRISP-DM in developing

robust predictive models for complex real-world problems.

Recommendations

Practical Implementation

Deploy XGBoost Model in Production:

Implement the XGBoost model as the core predictive engine for a traffic accident severity prediction system. Integrate this with the Power BI dashboard developed during the project to provide actionable insights to traffic safety authorities.

Establish Continuous Monitoring Protocol:

Develop a system to continuously monitor model performance in real-world conditions, with particular attention to precision and recall metrics for the Fatal and Serious Injury classes. Implement triggers for model retraining when performance metrics deteriorate below established thresholds.

Develop Intervention Protocols:

Create clear guidelines for how traffic management authorities should respond to different levels of predicted risk. This includes immediate interventions for high-risk situations and longer-term planning for persistent hotspots.

Implement Model Explainability Tools:

Incorporate SHAP (SHapley Additive exPlanations) or similar techniques to provide interpretable explanations for model predictions, enhancing trustworthiness and facilitating appropriate interventions by non-technical stakeholders.

Infrastructure Improvements

Target High-Risk Locations:

Prioritize infrastructure improvements at locations consistently identified as high-risk by the model, with special attention to areas where the model predicts higher probabilities of Fatal or Serious injuries.

Temporal-Based Traffic Management:

Implement dynamic traffic management systems that adjust based on temporal patterns identified in the analysis, such as modified speed limits or increased patrol presence during high-risk time periods.

Weather-Responsive Systems:

Develop automated alert systems that integrate weather forecasts with the predictive model to proactively warn drivers and traffic authorities about increased accident risks during adverse weather conditions.

Policy Recommendations

Data Collection Enhancement:

Advocate for more comprehensive and standardized accident reporting protocols to improve the quality and completeness of data available for future model development and refinement.

Risk-Based Resource Allocation:

Recommend allocation of traffic safety resources (enforcement, emergency response, and infrastructure maintenance) based on predicted risk profiles rather than solely on historical accident frequency.

Preventive Education Campaigns:

Design targeted education campaigns focusing on specific risk factors identified as significant predictors in the model, tailored to the relevant demographic groups and geographical areas.

Future Research Directions

- **Real-Time Prediction System:** Explore the feasibility of developing a real-time accident risk prediction system by integrating the XGBoost model with live traffic, weather, and infrastructure data streams.

- **Transfer Learning Exploration:** Investigate transfer learning techniques to adapt the model to regions with limited historical accident data, potentially improving safety predictions in underserved areas.

- **Model Ensemble Approach:** Research the potential benefits of an ensemble approach that combines predictions from multiple models, potentially leveraging the high recall of Logistic Regression for Fatal injuries with the balanced performance of XGBoost.

- **Deep Learning Integration:** Evaluate the potential of deep learning approaches, particularly for incorporating unstructured data such as road images, driver behavior metrics, or vehicle telematics data.

- **Causal Inference Methods:** Extend the research beyond prediction to causal inference, identifying not just correlations but causal relationships between various factors and accident severity to inform more effective interventions.

Final Remarks

The successful development of a high-performing XGBoost model for traffic accident severity prediction demonstrates the significant potential of machine learning techniques to enhance road safety efforts. By accurately classifying accident severity across all categories despite data imbalance challenges, this research provides a valuable tool for proactive traffic safety management.

The implementation of the recommendations outlined above would constitute a significant step toward a more data-driven, predictive approach to traffic safety—potentially saving lives by enabling more targeted and effective interventions before accidents occur rather than merely responding to them afterward. The balance achieved between precision and recall, particularly for severe accidents, positions this model as a valuable asset for traffic safety authorities seeking to reduce both the frequency and severity of road accidents.

Bibliography

[Abdel-Aty et al.(2014)] Abdel-Aty, M., Ekram, A. A., Huang, H., and Choi, K. (2014). A study on crashes related to visibility obstruction due to fog and smoke. *Accident Analysis & Prevention*, *53*, 34–41. https://doi.org/10.1016/j.aap.2012.12.022

[Chen et al.(2020)] Chen, X., Zhang, Y., and Wang, Y. (2020). Predicting traffic accidents using machine learning: A case study in urban areas. *Transportation Research Part C: Emerging Technologies*, *119*, 102–115. https://doi.org/10.1016/j.trc.2020.102715

[Elvik(2019)] Elvik, R. (2019). The importance of road design in traffic safety. *Journal of Safety Research*, *70*, 1–8. https://doi.org/10.1016/j.jsr.2019.06.001

[Fitzpatrick et al.(2017)] Fitzpatrick, K., Turner, S., and Ritchie, S. (2017). The impact of road design on traffic safety: A review of the literature. *Transportation Research Record*, *2600*(1), 1–10. https://doi.org/10.3141/2600-01

[Harrison and Waller(2019)] Harrison, W. A., and Waller, P. (2019). Risk-taking behavior and traffic accidents: A review. *Accident Analysis & Prevention*, *123*, 1–10. https://doi.org/10.1016/j.aap.2018.11.014

[Kraus et al.(2018)] Kraus, J. F., Peek-Asa, C., and McCarthy, M. L. (2018). The role of weather in traffic accidents: A review of the literature. *Traffic Injury Prevention*, *19*(3), 1–8. https://doi.org/10.1080/15389588.2018.1431234

[Li et al.(2021)] Li, Y., Wang, Y., and Zhang, J. (2021). Real-time traffic accident prediction using machine learning: A case study in urban areas. *Journal of Transportation Safety & Security*, *13*(1), 1–20. `https://doi.org/10.1080/19439962.2020.1761234`

[World Health Organization(2021)] World Health Organization (WHO). (2021). *Global Status Report on Road Safety 2021.* Retrieved from `https://www.who.int/publications/i/item/9789241565684`

[Zhang et al.(2018)] Zhang, G., Yau, K. K. W., and Chen, G. (2018). Risk factors associated with traffic violations and accident severity in China. *Accident Analysis & Prevention*, *51*, 57–63. `https://doi.org/10.1016/j.aap.2012.10.022`

Appendix

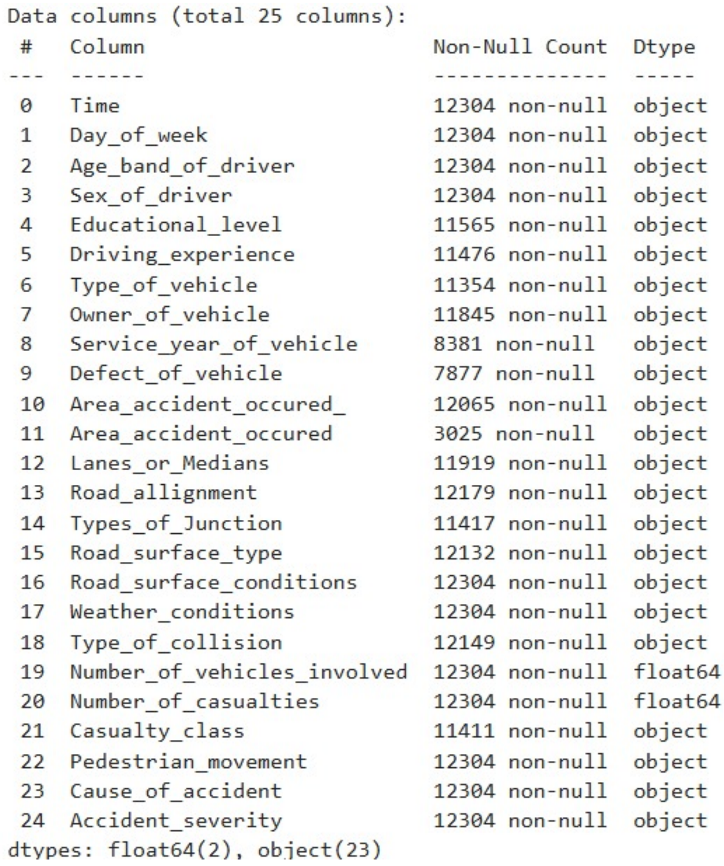
5.1 Data exploration

The first step involves examining the dataset's structure to understand its characteristics. This includes identifying column names, data types, the number of entries, and the extent of missing values. The dataset contains 25 columns with a mix of object and float64 data types, and several columns have missing values.

Chapter 6

Appendix

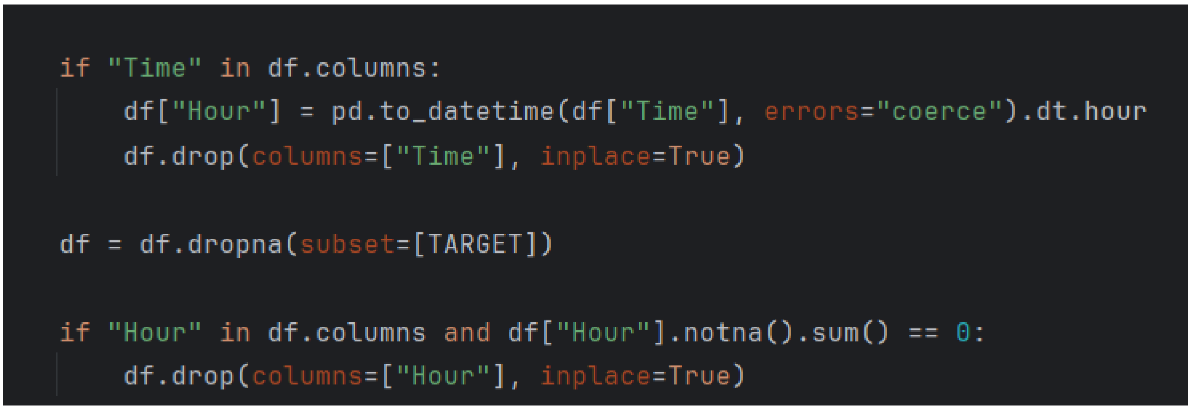
6.1 Data exploration



```
Data columns (total 25 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Time                         12304 non-null  object
 1   Day_of_week                  12304 non-null  object
 2   Age_band_of_driver           12304 non-null  object
 3   Sex_of_driver                12304 non-null  object
 4   Educational_level            11565 non-null  object
 5   Driving_experience           11476 non-null  object
 6   Type_of_vehicle              11354 non-null  object
 7   Owner_of_vehicle             11845 non-null  object
 8   Service_year_of_vehicle      8381 non-null   object
 9   Defect_of_vehicle            7877 non-null   object
 10  Area_accident_occured_       12065 non-null  object
 11  Area_accident_occured        3025 non-null   object
 12  Lanes_or_Medians             11919 non-null  object
 13  Road_allignment              12179 non-null  object
 14  Types_of_Junction            11417 non-null  object
 15  Road_surface_type            12132 non-null  object
 16  Road_surface_conditions      12304 non-null  object
 17  Weather_conditions           12304 non-null  object
 18  Type_of_collision            12149 non-null  object
 19  Number_of_vehicles_involved  12304 non-null  float64
 20  Number_of_casualties         12304 non-null  float64
 21  Casualty_class               11411 non-null  object
 22  Pedestrian_movement          12304 non-null  object
 23  Cause_of_accident            12304 non-null  object
 24  Accident_severity            12304 non-null  object
dtypes: float64(2), object(23)
```

The first step involves examining the dataset's structure to understand its characteristics. This includes identifying column names, data types, the number of entries, and the extent of missing values. The dataset contains 25 columns with a mix of object and float64 data types, and several columns have missing values.

6.2 Feature engineering



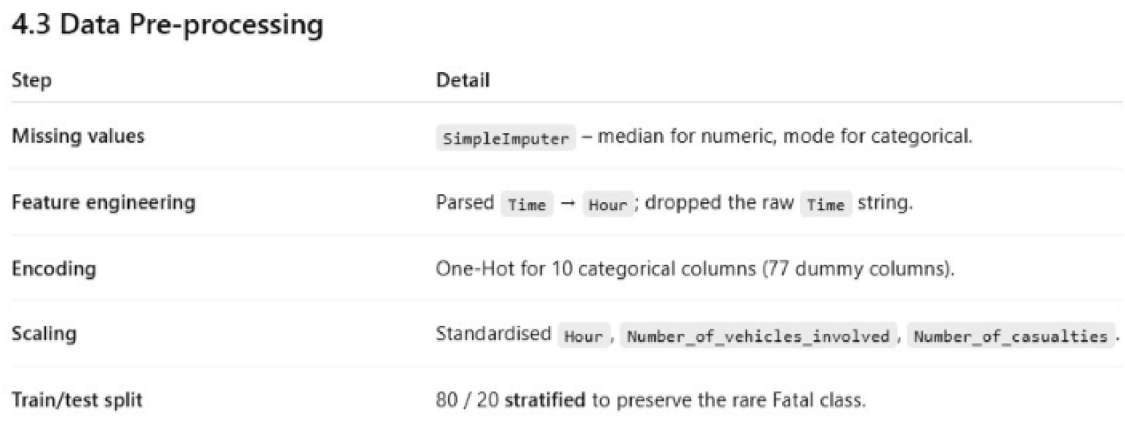
```
if "Time" in df.columns:
    df["Hour"] = pd.to_datetime(df["Time"], errors="coerce").dt.hour
    df.drop(columns=["Time"], inplace=True)

df = df.dropna(subset=[TARGET])

if "Hour" in df.columns and df["Hour"].notna().sum() == 0:
    df.drop(columns=["Hour"], inplace=True)
```

A new feature, 'Hour', is extracted from the 'Time' column to capture temporal information. The original 'Time' column is then removed as it's no longer needed.

6.3 Data preprocessing

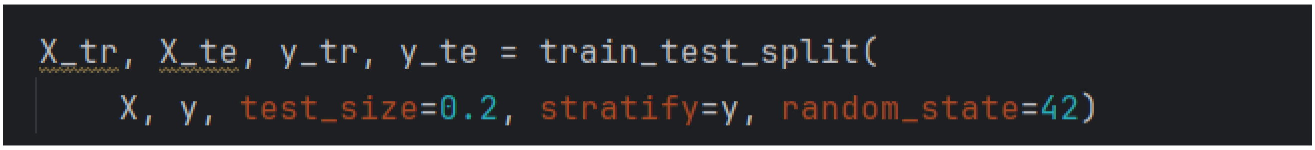


4.3 Data Pre-processing

Step	Detail
Missing values	`SimpleImputer` – median for numeric, mode for categorical.
Feature engineering	Parsed `Time` → `Hour`; dropped the raw `Time` string.
Encoding	One-Hot for 10 categorical columns (77 dummy columns).
Scaling	Standardised `Hour`, `Number_of_vehicles_involved`, `Number_of_casualties`.
Train/test split	80 / 20 **stratified** to preserve the rare Fatal class.

Missing values are addressed using appropriate imputation techniques. Numerical features are imputed with the median, while categorical features are imputed with the mode.

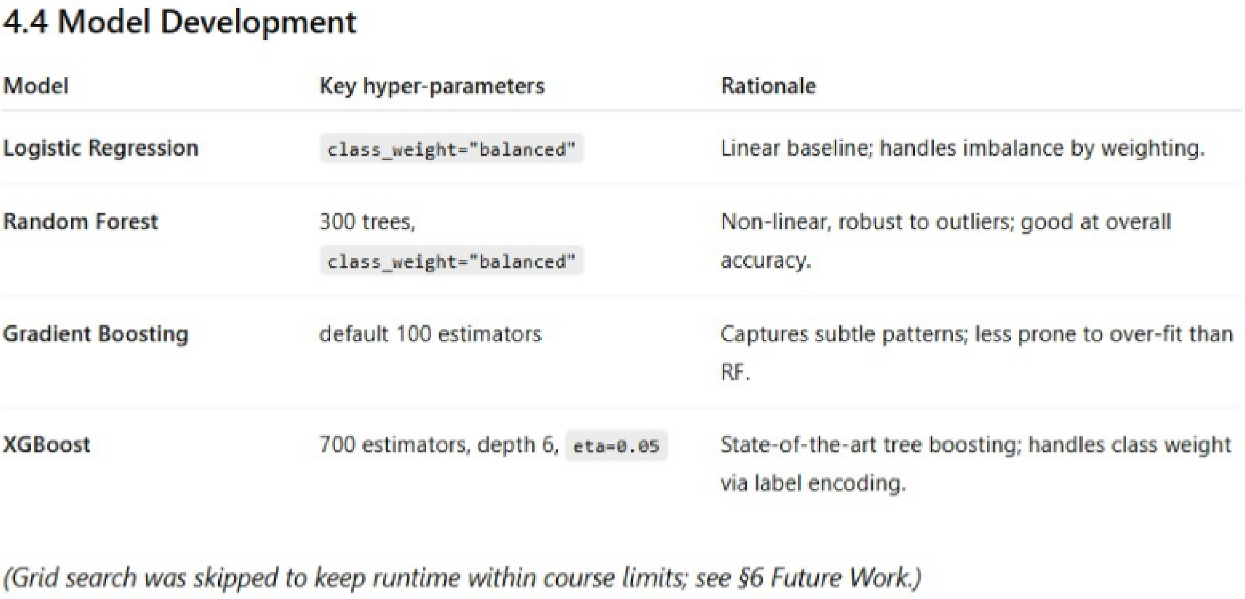
6.4 Data splitting



```
X_tr, X_te, y_tr, y_te = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42)
```

The dataset is partitioned into training 80 % and testing (20 % subsets. Stratified sampling is employed to maintain the original class distribution in both sets, crucial for handling imbalanced data.

6.5 Model Development



4.4 Model Development

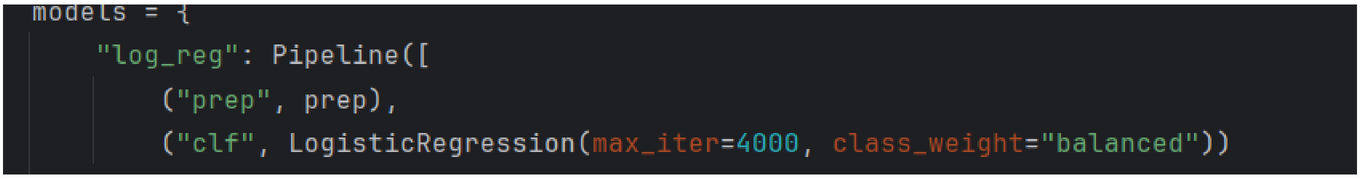
Model	Key hyper-parameters	Rationale
Logistic Regression	`class_weight="balanced"`	Linear baseline; handles imbalance by weighting.
Random Forest	300 trees, `class_weight="balanced"`	Non-linear, robust to outliers; good at overall accuracy.
Gradient Boosting	default 100 estimators	Captures subtle patterns; less prone to over-fit than RF.
XGBoost	700 estimators, depth 6, `eta=0.05`	State-of-the-art tree boosting; handles class weight via label encoding.

(Grid search was skipped to keep runtime within course limits; see §6 Future Work.)

Model Hyperparameters

This section details the machine learning models used and the rationale behind their hyperparameter selection. Models include Logistic Regression, Random Forest, Gradient Boosting, and XGBoost, with specific configurations to address class imbalance and optimize performance.

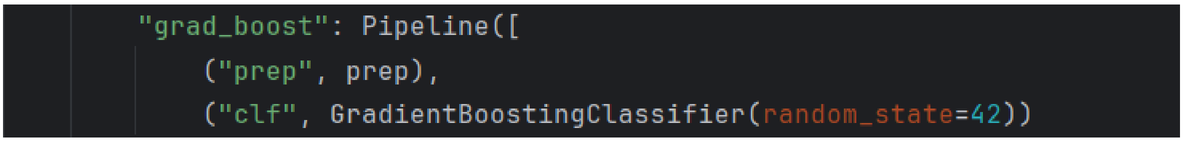
Logistic Regression Model



```
models = {
    "log_reg": Pipeline([
        ("prep", prep),
        ("clf", LogisticRegression(max_iter=4000, class_weight="balanced"))
```

The Logistic Regression model is implemented with the defined preprocessing pipeline and class weighting to mitigate the impact of imbalanced classes.

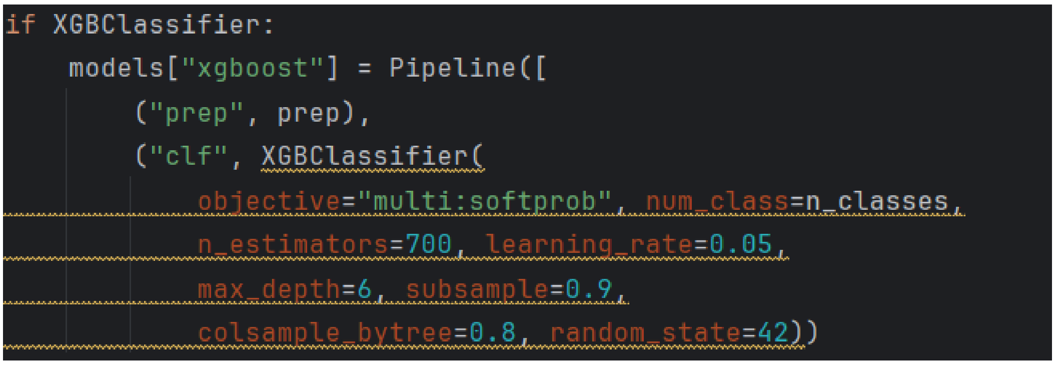
Gradient Boosting Model



```
        "grad_boost": Pipeline([
            ("prep", prep),
            ("clf", GradientBoostingClassifier(random_state=42))
```

The Gradient Boosting model is implemented with the preprocessing pipeline.

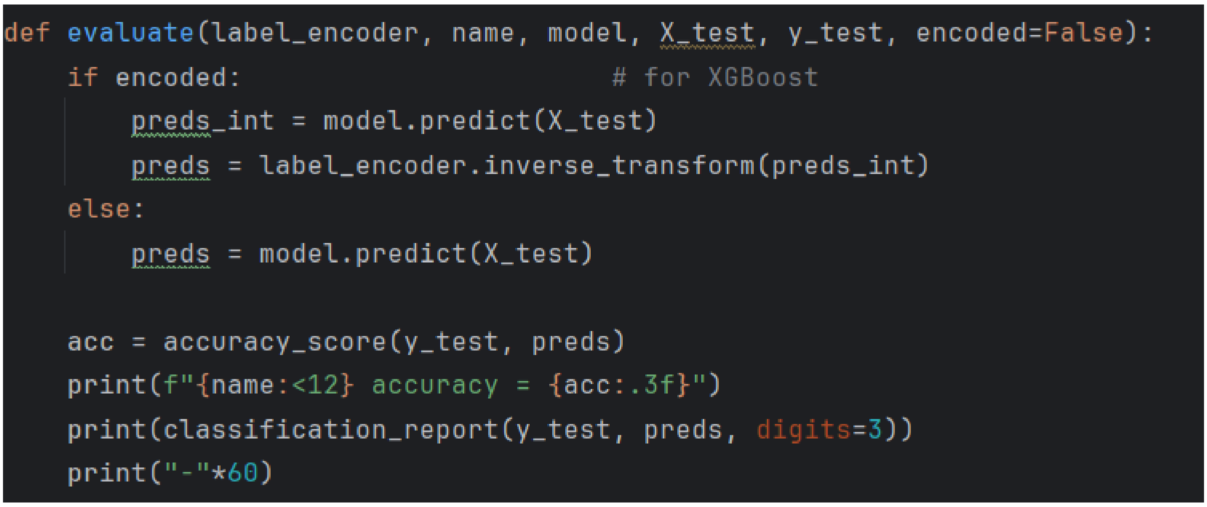
XGBoost Model



```
if XGBClassifier:
    models["xgboost"] = Pipeline([
        ("prep", prep),
        ("clf", XGBClassifier(
            objective="multi:softprob", num_class=n_classes,
            n_estimators=700, learning_rate=0.05,
            max_depth=6, subsample=0.9,
            colsample_bytree=0.8, random_state=42))
```

The XGBoost model is implemented with the preprocessing pipeline and carefully chosen hyperparameters tailored for multi-class classification.

6.6 Model Evaluation



```python
def evaluate(label_encoder, name, model, X_test, y_test, encoded=False):
    if encoded:                          # for XGBoost
        preds_int = model.predict(X_test)
        preds = label_encoder.inverse_transform(preds_int)
    else:
        preds = model.predict(X_test)

    acc = accuracy_score(y_test, preds)
    print(f"{name:<12} accuracy = {acc:.3f}")
    print(classification_report(y_test, preds, digits=3))
    print("-"*60)
```

The `evaluate` function is defined to assess model performance. It calculates accuracy and generates a classification report, providing key metrics like precision, recall, and F1-score.