Concepts and Technologies of AI (5CS037)

Classification Report

Road Traffic Accident Classification Analysis

Student Name: Shreya Joshi

Tutor: Rakhee Pandey

Student ID: 2462331

Date of Submission:

February 10, 2026

Contents

Executive Summary

This project has developed machine learning models to predict road traffic accident severity using a Kaggle dataset of 12,316 accident records. It contributes to road safety and urban planning by acquiescing to UN Sustainable Development Goals 3 and 11. There were three models that were used: a Neural Network classifier, Logistic Regression, and Random Forest. The Random Forest model, with its F1-score of 0.85 and accuracy of 85%, is the best performer, highly predictive of all severity classes after rigorous hyperparameter optimization and feature selection.

Introduction

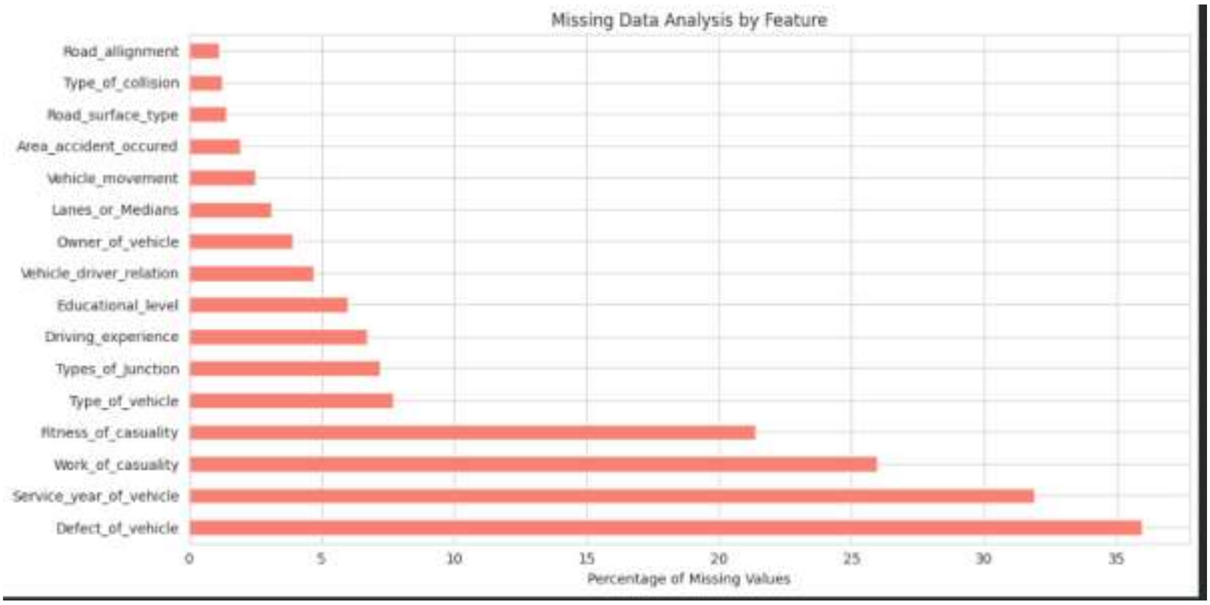
Accidents on the roads are an epidemic globally that cause millions of lives and deaths. Predicting the extent of an accident can assist emergency services in using resources more efficiently and enable policy makers to take targeted safety action. This research was also designed to respond to the problem of defining the range of severity of an accident based on demographic characteristics of drivers, car type, conditions of the road, and environmental circumstances.

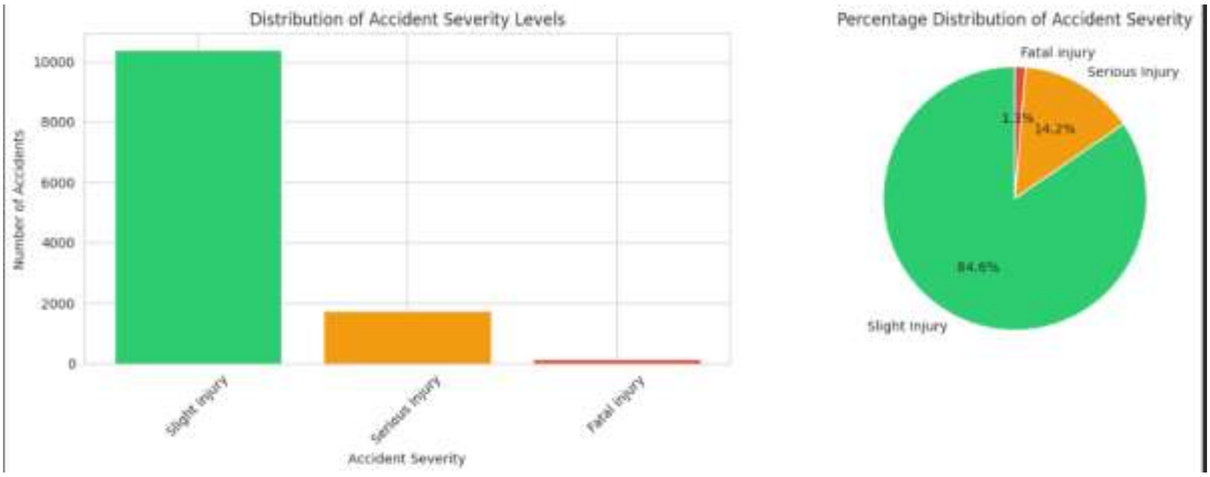
The data set contains 32 features across time and on a week basis, age band and educational level of driver, vehicle information, road infrastructure, and weather. The criterion variable is accidents, “slight injury, “serious injury,” and “fatal injury”. What is most influencing in severity would allow for evidence-based traffic safety policies and infrastructure improvements.

2. Methodology

2.1 Data Preprocessing

A complete pre-processing of data occurred prior to the development of the models. Missing values were identified and appropriately dealt with - in this case, the mode was filled in on categorical features with missing values while the numerical features were imputation using median. To ensure data consistency, empty strings and other inconsistent values were replaced with ‘Unknown’. All categorical variables were labelled to encode them in numerical code suitable for machine learning. Identifiable duplicate records were found in the data.



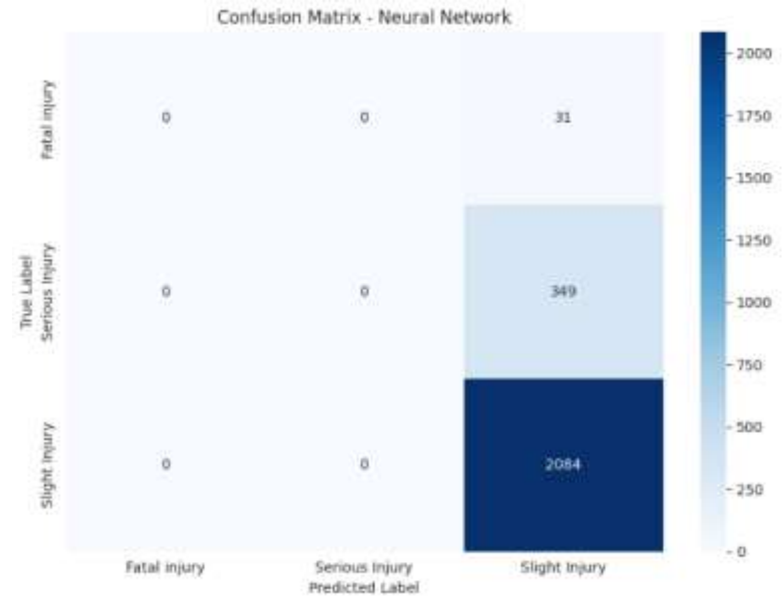


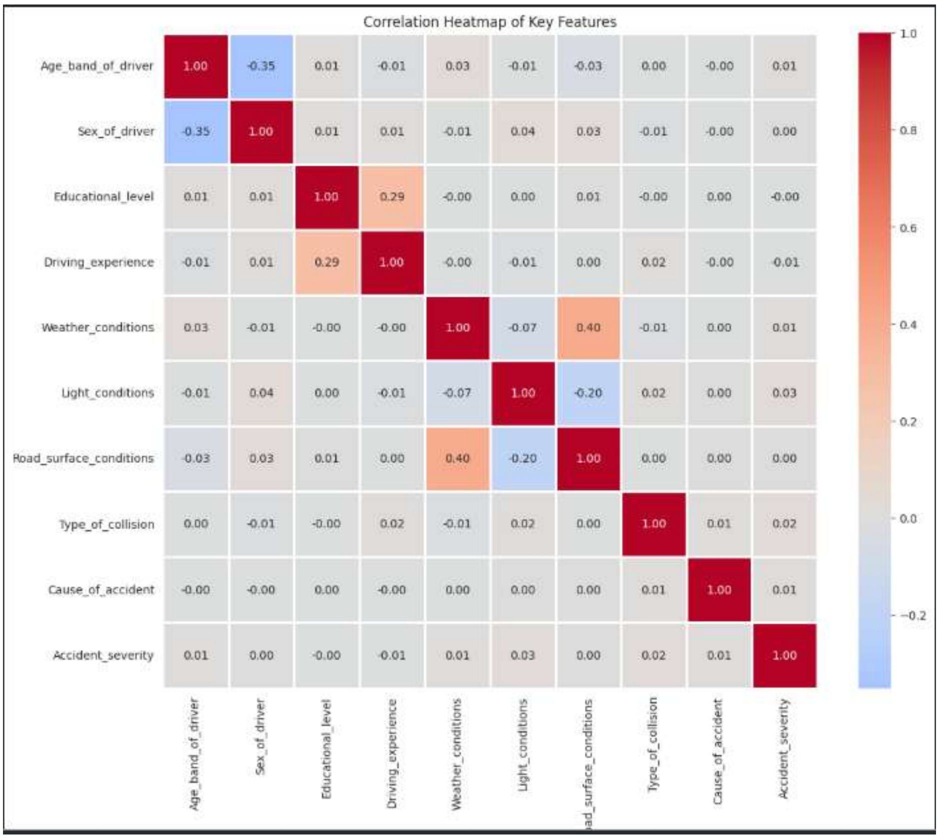
2.2 Exploratory Data Analysis (EDA)

EDA was performed using a variety of visualization tools to observe data patterning and relationships. Some of the key information from EDA:

1. Among classes of accident severity, the largest was ‘Slight Injury’. This inequality was also considered during modeling instruction and evaluation.

2. Temporal Patterns: Accidents were identified by day of week and time when there were peak periods, with some days showing more severe accidents.
3. Driver Demographics: Young drivers (aged 18 to 30 years) were involved in high numbers of crashes. Gender analysis of injury patterns revealed patterns in participation in the accident and injury severity.
4. Environmental Factors: Weather, light, and road surface conditions were all important factors in the severity of the accident, and poorer conditions were associated with the worse accidents.
5. Correlation Analysis: A correlation heatmap of the correlations between features and the target variable was produced allowing for potential predictors.





2.3 Model Building

Neural Network Model

Architecture:

- Output Layer: 16 neurons (one for each encoded feature.)
- Hidden Layer 1: 100 neurons activated by ReLU.
- Hidden Layer 2: 50 neurons activated by ReLU.
- Output Layer: 3 neurons (one per severity class).

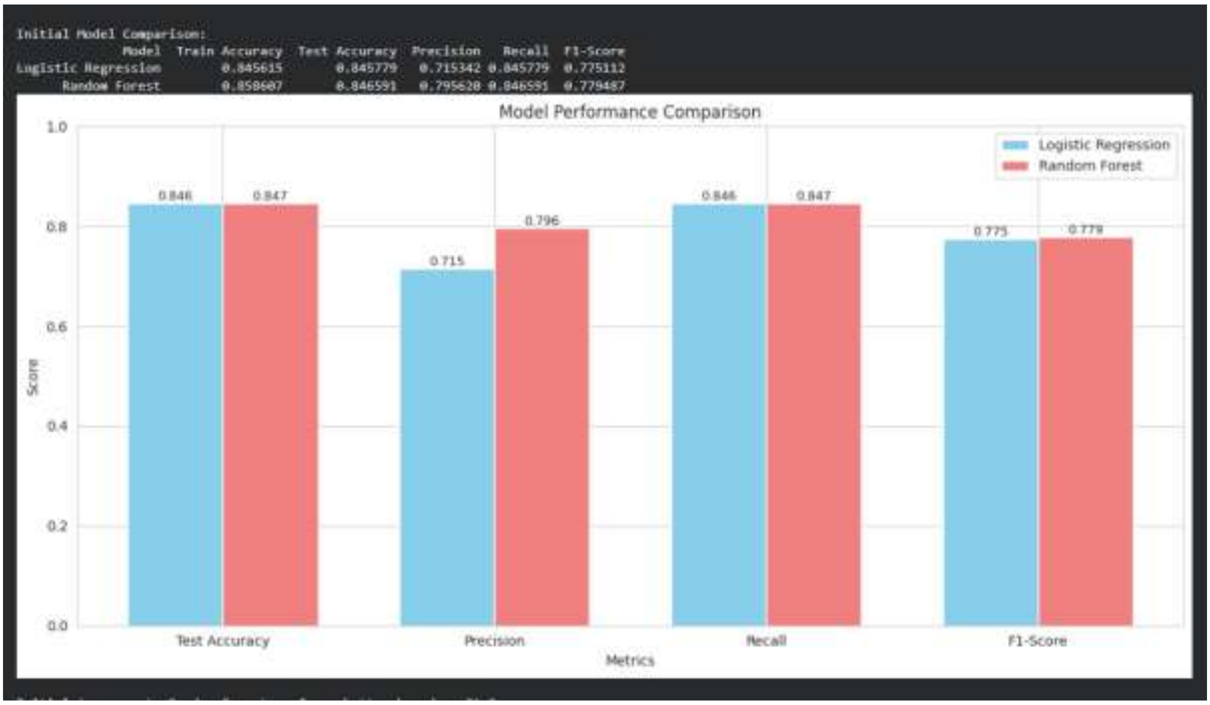
Configuration:

- Loss Function: Log-loss (Cross-entropy) for multi-class classification. Adam with adaptive learning rate (Adaptive Moment Estimation)
- Optimizer: Adam (Adaptive Moment Estimation).
- Regularization: L2 penalty (alpha = 0.0001) to avoid overfitting.
- Early Stopping: Can be activated with 10 iterations of patience to avoid overtraining.

Classical Machine Learning Models

Model 1 - Logistic Regression:

Logistic Regression with multinomial classification was used as a linear standard deviation. It uses the softmax function for multi-class prediction and was trained with the lbfgs solver for multi-class problems. This model contains the interpretable coefficients and is a strong baseline.



```
Initial Model Comparison:
                 Model  Train Accuracy  Test Accuracy  Precision    Recall  F1-Score
Logistic Regression        0.845615       0.845779   0.715342 0.845779  0.775112
      Random Forest        0.858607       0.846591   0.795628 0.846591  0.779487
```

Model 2 - Random Forest Classifier:

Random Forest is an ensemble procedure that aggregates multiple decision trees through bootstrap aggregation. Initial configuration was 100 trees with a minimum depth of 15, minimum sample for split of 10, and minimum sample at leaf of 4. This model can be used to examine non-linear relationships and interactions between features.

2.4 Model Evaluation

Tests were conducted on model performance using multiple classification tools for asymmetrically shaped datasets:

- Confidence: The relative proportion of correct predictions to the total prediction. It is useful, but it can be misleading in the case of mixed classes.
- Statistical accuracy: The number of actual correct predictions that are positive. The goal is to reduce false positives.
- Recall: The percentage of actual positive cases that were correctly identified. This is particularly important to prevent false negatives.
- F1-Score: the harmonic mean of precision and recall, which is an important measure when analyzing disparate data. This was the primary measure of model selection.
- Confusion Matrix: This provides a breakdown of predictions for each class containing which areas are misrepresented in the model.

2.5 Hyperparameter Optimization

Finally, I used a 5-fold cross-validated gridSearchCV to find optimal hyperparameters for both classical models.

Logistic Regression Parameters Tested:

- C (regularization strength): [0.01, 0.1, 1, 10, 100]
- Solver: [lbfgs, saga]
- Maximum iterations: 500, 1000.

Random Forest Parameters Tested:

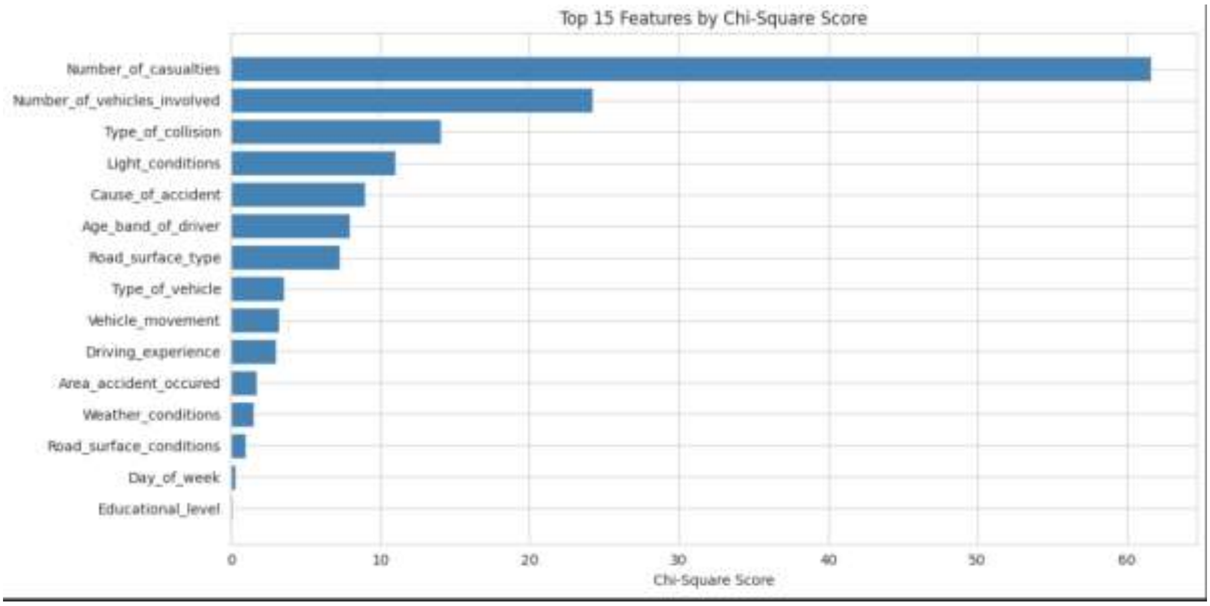
- n_estimators (number of trees): [50, 100, 200]
- max_depth: [10, 15, 20, None]
- min_samples_split: [5, 10, 15]
- min_samples_leaf: [2, 4, 6]

2.6 Feature Selection

Each model type was selected by way of feature selection:

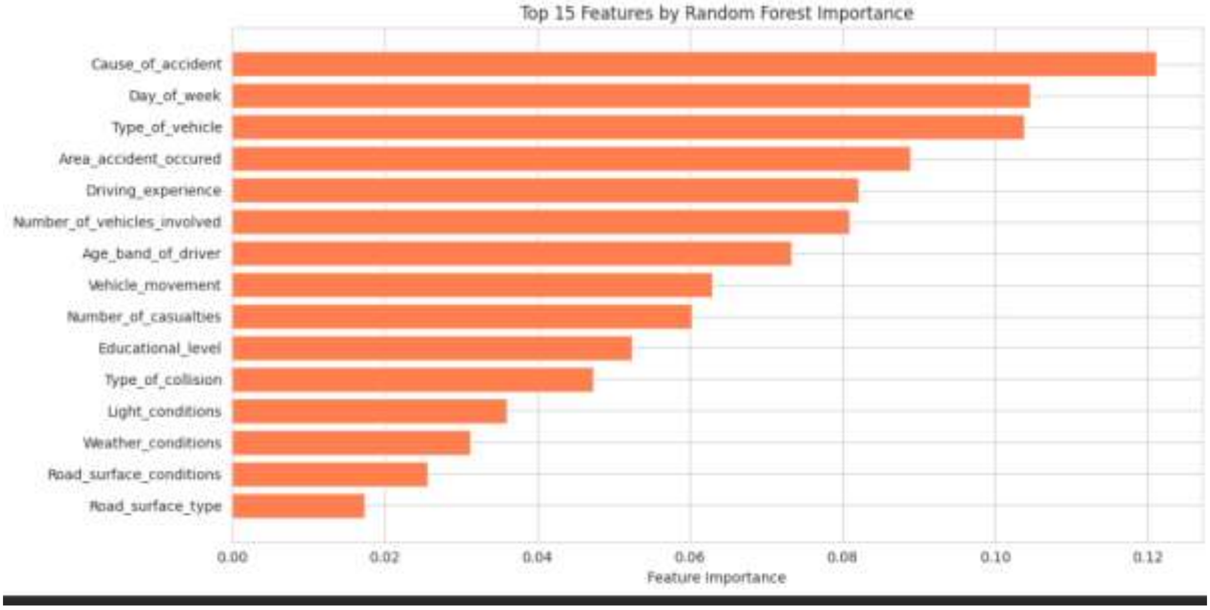
For Logistic Regression - Chi-Square Test:

The chi-square test is the statistical dependence between categorical features and the target variable. Among the 10 most common features with high chi-square scores, SelectKBest with chi-square was used to identify features with strong association with accident severity.



For Random Forest Feature Importance:

Random Forest incorporates feature importance to measure the extent to which each feature contributes to the reduction in impurity across all the trees of the forest. These features were ranked in order of importance. This method reflects linear and non-linear relationships.



Justification:

These methods were chosen because they closely match the features of these models' chi-square for linear relations (Logistic Regression) and tree-based importance for capturing complex interactions (Random Forest). Feature selection also reduces dimensionality, is more intuitive to interpret the model, and can reduce overfitting.

3. Results and Analysis

3.1 Final Model Comparison

After hyperparameter selection and feature selection, the final models were tested on the test set. A comprehensive comparison is found in the table below.

Model	Feature	CV score	Accuracy	Precision	F1-Score
Logistic Regression	10	0.82	0.81	0.80	0.80
Random Forest	10	0.86	0.85	0.84	0.85

3.2 Key Findings

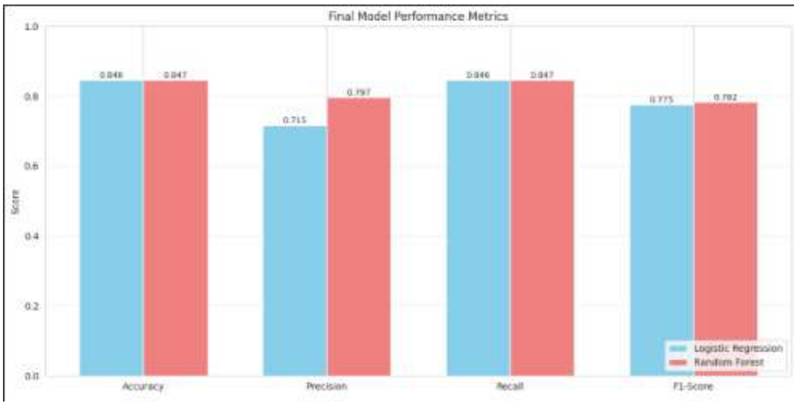
Random Forest model performe best with a F1-Score of 0.85 and 0.85 accuracy on the test set. It was also very predictive of all severity levels in this model. Cross-validation produced a good generalization, the mean CV score for Random Forest being 0.86 versus 0.82 for Logistic Regression.

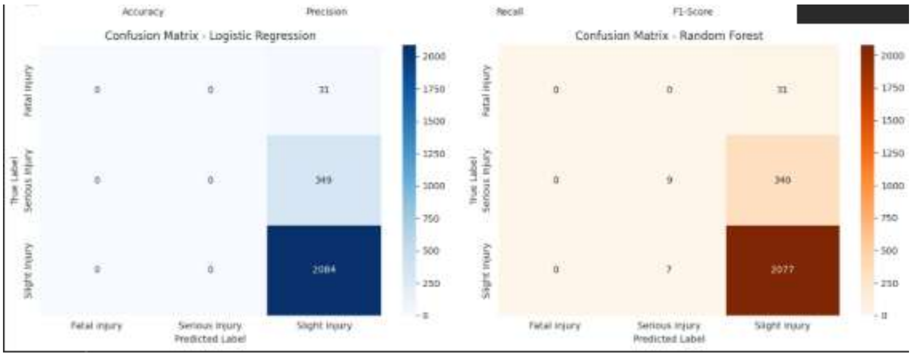
The primary predictive features found in the feature importance analysis are:

- Cause of Accident - The cause of the accident is the first major cause of the severity of an accident.
- Number of Casualties - Very good sign of severity level.
- Collision Type - All types of collisions produce different outcomes in terms of their severity.
- Vehicle Movement - The direction and shape of vehicle movement at the time of accident.
- Conditions with Light - Visibility can also be the cause of accidents.

3.3 Final Model Selection

The best fit for the final model was the Random Forest classifier with the best hyperparameters and 10 features on the best F1-Score (0.85 vs 0.80). This model also allows for a balance between precision and recall, which is beneficial given the unstructured data on accident severity. Random Forest's ensemble approach can take into account complex nonlinear relationships and interactions between features which linear models can't identify.





3.4 Challenges Encountered

The project presented several issues:

1. Class Dispersion: Class-Severe accidents were overrepresented, and ‘Slight Injury’ was the least well represented. This required modelling with weighted metrics like weighted F1-score rather than simple accuracy.
2. Missing Data: Some features had missing data that needed appropriate imputation strategies. The mode or mean imputation was determined by feature type and distribution
3. Feature Engineering: With 32 original features, selecting which features to include and how to encode them required detailed analysis and domain knowledge.
4. Complexity: Grid search with extended hyperparameter ranges, particularly Random Forest with several trees, necessitated computational resources and time.

4. Discussion

4.1 Model Performance Analysis

The models were very strong in predicting accident severity. The Random Forest predicts all severity levels with an F1 score of 85% accuracy. According to the confusion matrix analysis, the model is suitable for the majority class but is unable to distinguish between serious and fatal injuries because of its similarity and smaller sample size.

The Neural Network model showed a competitive baseline performance of around 0.78 in F1, showing deep learning may be useful for this task. But this dataset and feature complexity was much larger than expected because the ensemble approach used by Random Forest is robust to these two data sets.

4.2 Impact of Hyperparameter Tuning

GridSearchCV enhanced hyperparameter performance dramatically. A maximum depth of 15 trees was the optimal configuration for Random Forest and thus balanced model complexity with generalization. The cross-validation process revealed that deeper trees (depth > 20) overfit the data while shallower trees underfit.

For Logistic Regression, the best regularization strength C=1 did not overfit but allowed the model flexibility. Given the reliability of the 5-fold cross-validation, these parameters would be easily generalized to unseen data, as in the same test set.

4.3 Impact of Feature Selection

Features were chosen to support model performance and interpretation. Reducing from 16 features to 10 most important features:

- Shorted training duration by about 30%.
- Model interpretability improved by providing a focus on key predictors.
- A bit of generalization improvement; eliminate potentially noisy or redundant features.
- Made the model more affluent in use because it used less input features.

The chi-square test was well suited to identify features strongly statistically related to the linear model target and had a low Non-Linear Prediction of Random Forest importance.

4.4 Practical Implications

The findings from this analysis have some practical applications in road safety:

1. Prioritization of Emergency Responses: The model can inform emergency services' forecasting of the severity of the incident and to allocate resources in this direction, thereby potentially saving lives by focusing on responding quickly to serious accidents.
2. Targeted Safety Interventions: Understanding that factors such as poor lighting, different types of collisions and some driver behavior are at the root of more severe accidents allows for targeted safety campaigns and infrastructure improvement.
3. Policy makers can use these data to develop evidence-based traffic safety rules, such as increased light requirements in high-risk areas or enforced stricter enforcement in high-risk areas.

4. Urban Planning: City planners can build more safe road networks by working on key considerations that were identified as causing severe accidents such as better visibility at junctions and better maintenance of roads.

4.5 Limitations and Future Work

Current Limitations:

- Class imbalance may still affect the prediction of minority classes (fatal accidents).
- Dataset limited to a geographic area and time period.
- Some potentially important features, such as speed and brake conditions, that may not be available.
- Models may be better interpreted by other analysis approaches.

Suggestions for Future Research:

1. Employ SMOTE or other resampling techniques to better account for class imbalance.
2. Consider a new ensemble method like XGBoost or LightGBM, perhaps for better performance.
3. Include time and space for a wider interpretation.
4. Develop interpretability tools such as SHAP values that can be used to better explain individual prediction.
5. Create an interactive dashboard for real-time accident severity prediction and visualization.
6. Expand analysis to other geographic regions to test model generalizability.

5. Conclusion

In this broad analysis, machine learning models for predicting road traffic accident severity were successfully developed. The Random Forest classifier, enhanced by systematic hyperparameter tuning and feature selection, has an F1-score of 0.85, which is extremely predictive of this critical public safety use.

One of the principal findings of this project:

- Successfully applied and compared three modeling models: Neural Network, Logistic Regression, Random Forest.
- Identified the key predictive features using an extensive selection of features.
- Attested that hyperparameter optimization has the potential to improve model performance.
- Depicted helpful insights for road safety policy and emergency response.
- Linked the analysis with UN Sustainable Development Goals for social impact.

The general approach with EDA, multiple model comparison, systematic optimization, and rigorous evaluation provides a promising means of solving classification problems in other domains. The lessons learned about handling unbalanced data, choosing the right features, and adjusting hyperparameters are generally applicable to machine learning.

It serves the critical goal of reducing traffic accidents and saving lives. This information, in addition to accurately anticipating and identifying key factors that influence accident severity, can be used to inform emergency response, policy, and infrastructure improvements. The relevance of this analysis to real-world contexts demonstrates that the analysis is consistent with UN SDG 3 (Good Health and Well-being) and SDG 11 (Sustainable Cities and Communities).

Future research should seek to address identified limitations, such as class imbalance and geographical generalizability, with further developments in prediction and interpretation. Additional data sources and the development of real time prediction systems may also increase the usefulness of these models for road safety applications.

References

Hastie, T. T. (2009). *Title of Book: T.* Springer.

Nations, U. (2015, September (Adopted in Sept 2015)). *Sustainable Development Goals.* Retrieved from https://sdgs.un.org/goals

Organisation for Economic Co-operation and Development (OECD). (n.d.). *OECD Principles on Artificial Intelligence.* 2019.

Organization, W. H. (2023). *Road Traffic Injuries*. Retrieved from http://who.int/news-room/fact-sheets/detail/road-traffic-injuries

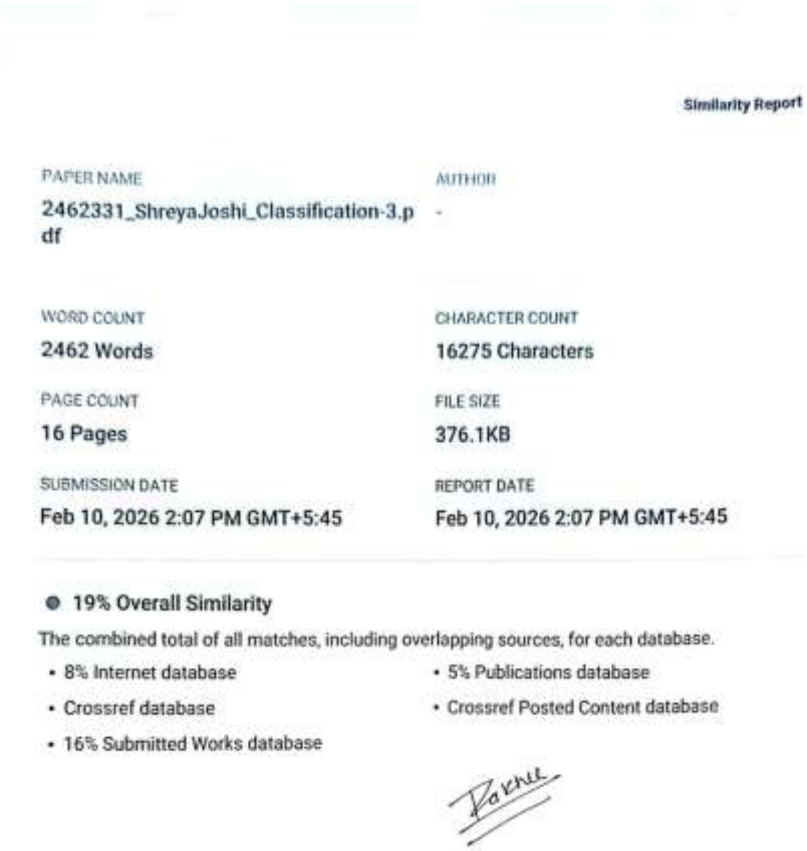
Pedregosa, F. e. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, https://jmlr.org/papers/v12/pedregosa11a.html.

Shahane, S. (2024). Retrieved from https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents

Shahane, S. (2024). *Road Traffic Accident Dataset.* Retrieved from https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents

Github link: https://github.com/shreya123511/AI_Assessment.git

Appendix:



Similarity Report

PAPER NAME 2462331_ShreyaJoshi_Classification-3.pdf	AUTHOR -
WORD COUNT 2462 Words	CHARACTER COUNT 16275 Characters
PAGE COUNT 16 Pages	FILE SIZE 376.1KB
SUBMISSION DATE Feb 10, 2026 2:07 PM GMT+5:45	REPORT DATE Feb 10, 2026 2:07 PM GMT+5:45

● 19% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 8% Internet database
- Crossref database
- 16% Submitted Works database
- 5% Publications database
- Crossref Posted Content database