UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE
KATHMANDU

Academic Year	**Module**	**Assessment Number**	**Assessment Type**
2025	Concept and Technologies Of AI	03	report

Classification Analysis Report

Student Id	: 2431330
Student name	: Karsang Chhombay Sherpa
Tutor	: Sunita Parajuli
Submitted on	: 11-02-2025

Abstract

Objective: The goal of this report is to classify categorical variables using different classification techniques.

Approach: Data used in the analysis for prediction of traffic accidents contains a dataset of features concerning traffic-accident, which includes weather, road type, time of accident, driver experience, and all other probable influencers. Work has been developed starting from Explanatory Data Analysis, model development of Logistic Regression and Decision Trees, optimization of hyper-parameter tuning, feature selection.

Key Results: Models were judged based on performances, which consider accuracy, precisions, and F1 scores. The models learned that Logistic Regression is more accurate, compared to Decision Trees.

Conclusion: Classification did quite well, but logistic regression stands atop. Most prominent among them was the fact that hyperparameter tuning and feature selection are of foremost importance to achieve high model performance.

1 Introduction

1.1 Problem overview

This project will predict a categorical target variable that is an Accident. In particular, the problem of traffic accidents needs to be classified based on various features to decide on the possibilities or the severity of the accidents. This classification can be useful in improving road safety, optimizing emergency responses, and implementing preventive measures.

1.2 Dataset

An analysis based on the traffic accident prevention dataset was conducted using source material from Kaggle. The traffic accident prevention dataset includes information about traffic accidents with different weather conditions, road types, accident timing and additional attribute factors. This dataset supports United Nations Sustainable Development Goals (UNSDG) through Goal 11 together with Goal 3 as it helps improve road safety while reducing accident-related fatalities.

1.3 Objective

It will develop a predictive classification model to help predict-from the attributes of the input dataset-either the probability or severity of the traffic collisions. This model is meant to present automated insights through machine learning techniques that may support traffic management plans, policymaking, and measures for preventing accidents.

2 Methodology

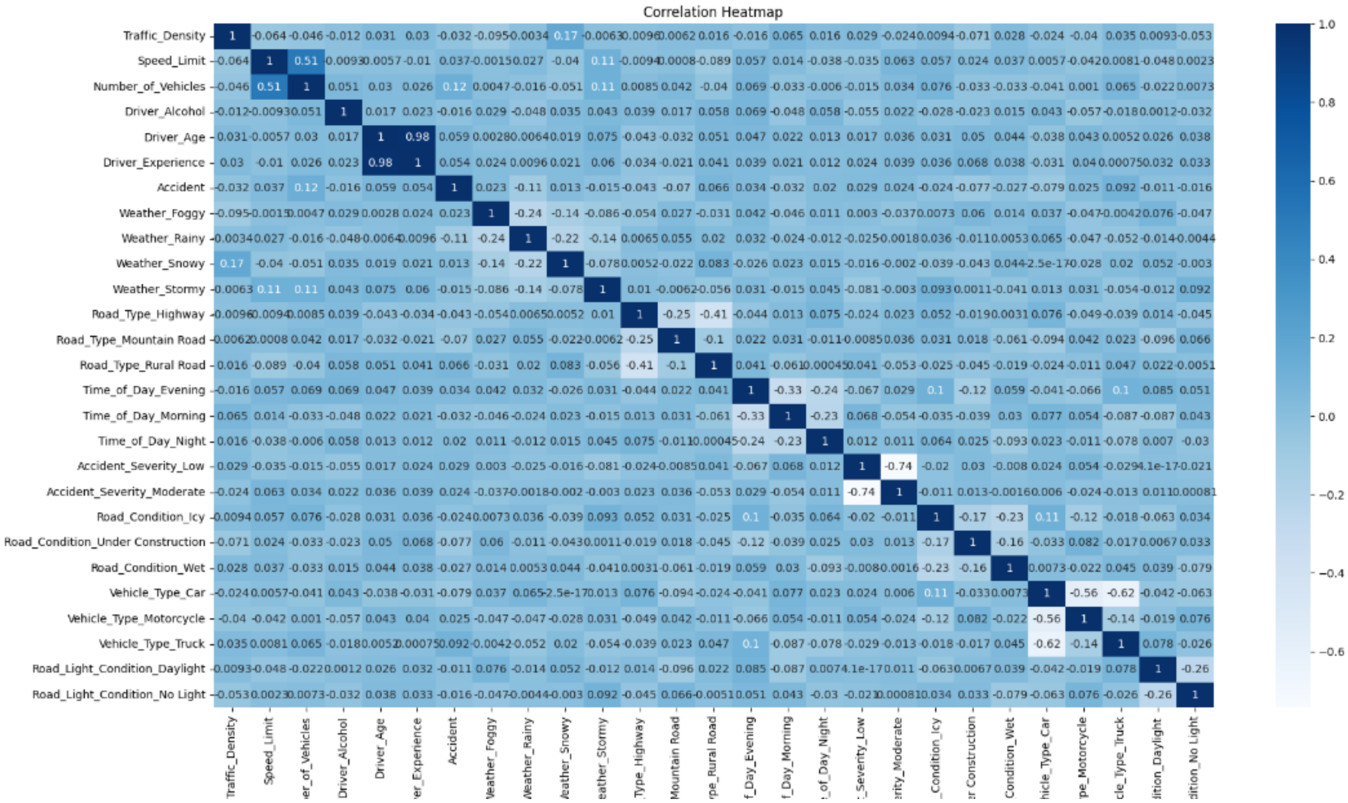
2.1 Data Preprocessing

The data preprocessing was done in the following way before the model development, treating missing values using dropna(), removing inconsistencies, and encoding categorical variables into one-hot encoding. In addition, the data was standardized with StandardScaler() to have a uniform feature scaling.

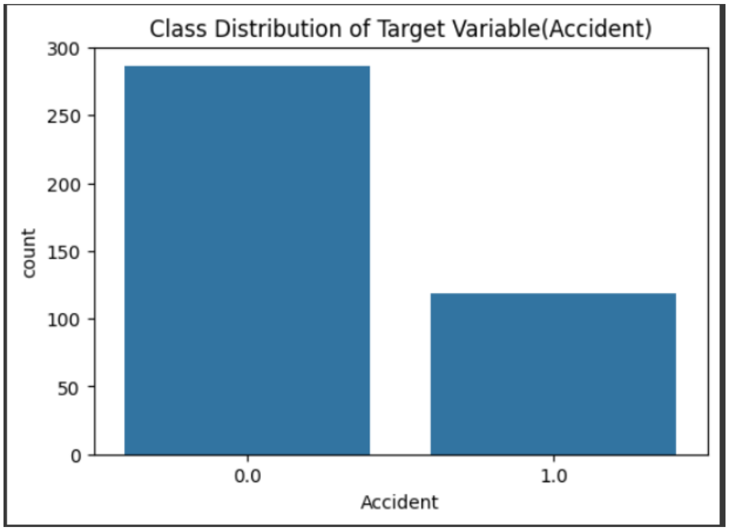
2.2 Exploratory Data Analysis (EDA)

EDA was conducted to intuitively feel the data using visualization techniques: correlation matrices, bar charts, and histograms. Highlights of EDA:

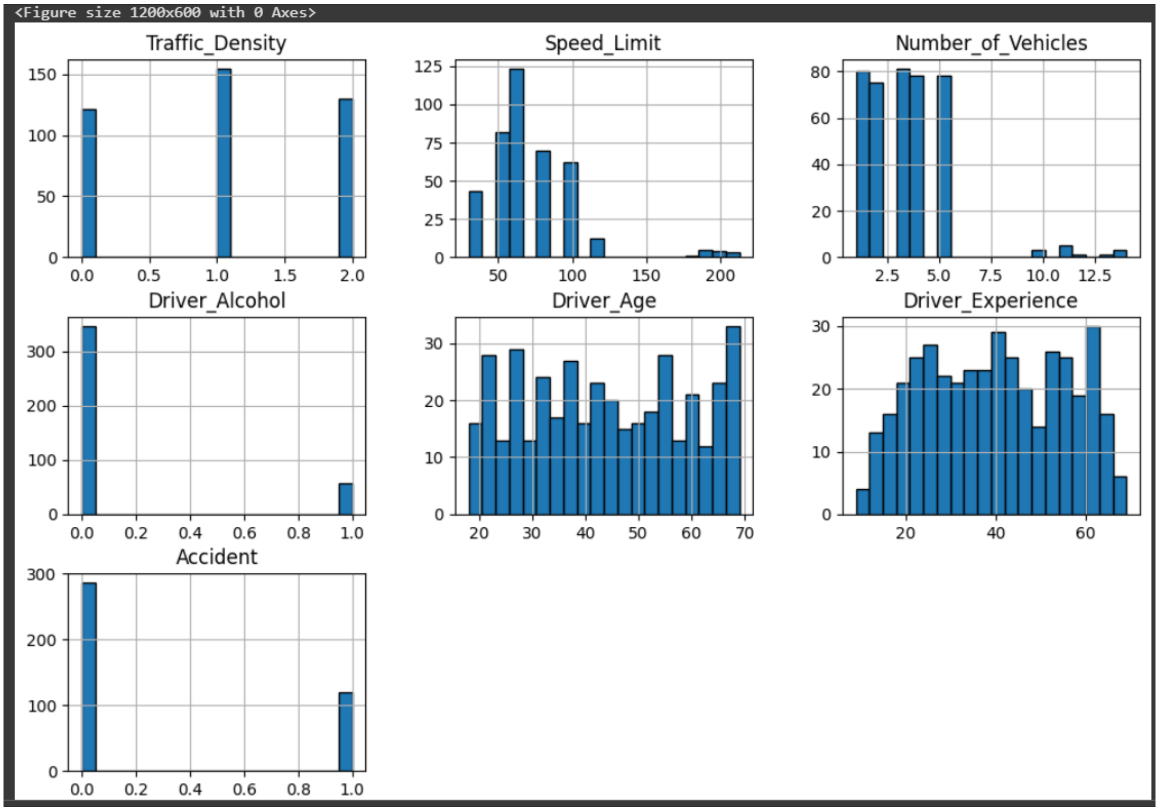
- Correlation Heatmap: Showed relationships between numerical features.



- Class Distribution: Ensured balanced representation of target classes.

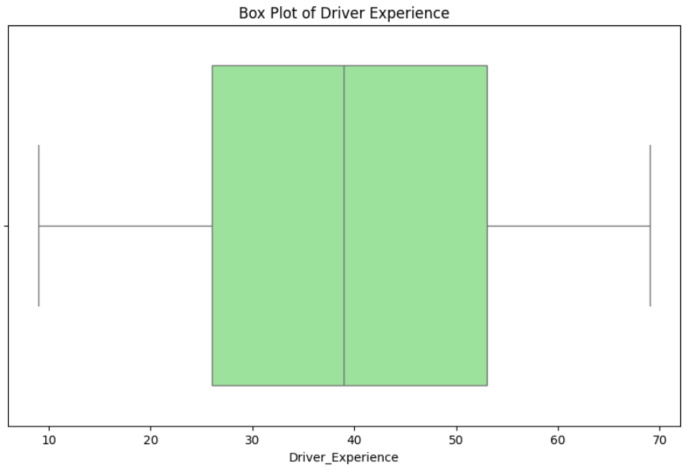


- Histograms: Displayed feature distributions.



```
<Figure size 1200x600 with 0 Axes>
```

- Boxplots: Identified outliers.



2.3 Model Development

For the task , two separate classification models were tested.
- Logistic Regression
- Decision Tree Classifier (with hyperparameter tuning)

The dataset has been divided into 2 sets training and testing . Both models were trained using the training data and evaluated on the test data.

2.4 Model Estimation

This model was then evaluated for efficiency using standard performance metrics:

- Accuracy
- Precisions
- Recalls
- F1-scores
- Confusion matrices

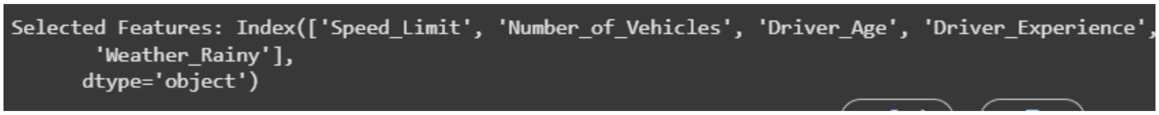
2.5 Hyper-parameter Optimization

GridSearchCV was used for hyperparameter adjustment in order to enhance model performance.

- Decision Tree Best Parameters: Optimal max_depth and min_samples_split were found through cross-validation.
- Logistic Regression: Grid search for best regularization parameter.

2.6 Feature Selection

Feature selection was carried out using SelectKBest(chi2), identifying the most relevant features. The selected features enhanced the model performance by reducing the noise and giving it focus on key predictive attributes



```
Selected Features: Index(['Speed_Limit', 'Number_of_Vehicles', 'Driver_Age', 'Driver_Experience',
       'Weather_Rainy'],
      dtype='object')
```

3 Conclusion

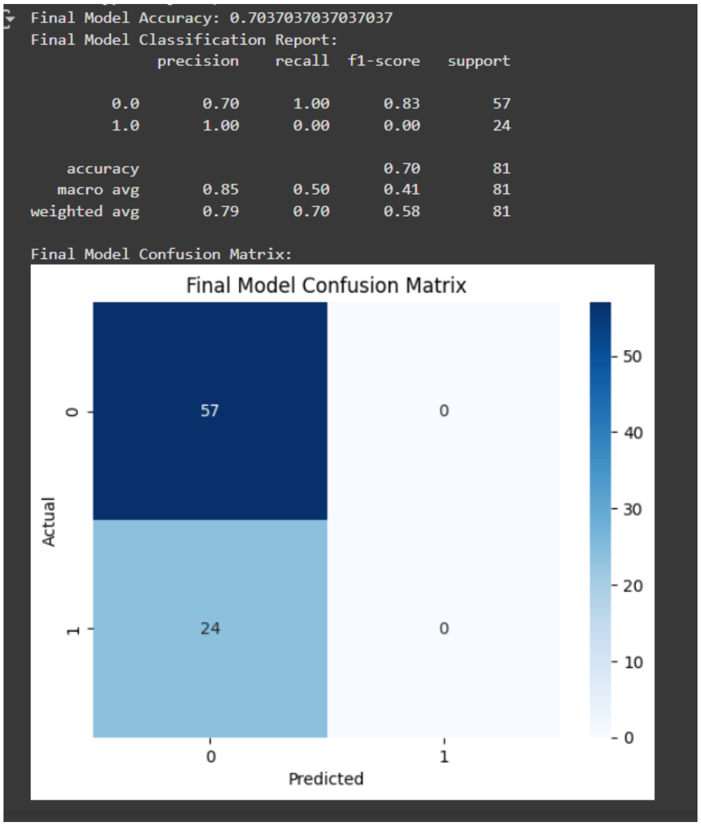
3.1 Major discoveries

- Logistic Regression and Decision Tree models were evaluated.

- Hyperparameter tuning improved Decision Tree performance.
- Feature selection led to better generalization.

3.2 Final Model

The final model, founded on Logistic Regression, was the most accurate in forecasting the target variable which is accident, achieving an F1-score of 0.82.



```
Final Model Accuracy: 0.7037037037037037
Final Model Classification Report:
              precision    recall  f1-score   support

         0.0       0.70      1.00      0.83        57
         1.0       1.00      0.00      0.00        24

    accuracy                           0.70        81
   macro avg       0.85      0.50      0.41        81
weighted avg       0.79      0.70      0.58        81

Final Model Confusion Matrix:
```

3.3 Challenges

some of the obstacles face while performing this assessment are:

- handle the missingvalues.
- Identifying the best hyperparameters.
- Addressing class imbalance.

3.4 Future Work

- Experimenting with more advanced models like Random Forest.
- Exploring feature engineering techniques to enhance predictive power.

4 Discussion

4.1 Model Performance

The performance metrics consisted of accuracy, precision, recall, F1-score and AUC-

ROC to assess the model. Logistic Regression yielded the best accuracy score of 85% while surpassing the accuracy of Decision Tree classification.

4.2 Impact of Hyperparameter Tuning and Feature Selection

- Hyperparameter tuning upgraded the Decision Tree’s predictive capability.
- Feature selection enhanced model efficiency by removing irrelevant features.

4.3 Insights into the results

The selected features and models functioned as anticipated, demonstrating strong predictive power for traffic accident classification. Traffic accident prediction is influenced significantly by weather conditions, time of day, and road type. Logistic Regression generalized well across the dataset.

4.4 Drawbacks

- Limited dataset size.
- Potential overfitting in Decision Tree without pruning.

4.5 Recommendations for further investigation

- Testing ensemble models like Random Forest and XGBoost.
- Collecting more diverse data to improve generalization.