A Capstone Project

On

‘Investigation of Road Accident’

Submitted By

Farooq Shaikh

Supervisor **Mr. Rithik Raj**

ABSTRACT

Road traffic accidents contribute significantly to injuries, fatalities, and economic losses globally. This project focuses on leveraging machine learning techniques to predict the severity of car accidents, aiming to enhance road safety measures and emergency response systems. The dataset utilized for this project is sourced from Kaggle, containing diverse information related to car accidents, including environmental conditions, road features, and temporal attributes.

This project aligns with the broader goal of leveraging machine learning to address societal challenges, emphasizing the potential impact on road safety and accident prevention. The findings and models developed here offer valuable contributions to the field of transportation safety and emergency response.

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

1.1 Objective

1.2 Motivation

1.3 Background

Objective

The primary objective of this project is to develop a predictive model capable of accurately classifying the severity of car accidents based on various contributing factors. The project aims to leverage machine learning techniques to enhance road safety measures, emergency response systems, and overall accident prevention.

The specific objectives are outlined below:

- Conduct a thorough exploration of the dataset to understand the distribution of features, identify patterns, and gain insights into factors influencing accident severity.
- Prepare the dataset for model training by addressing missing values, encoding categorical variables, handling class imbalance, and scaling numerical features.
- Apply feature selection techniques to identify and retain the most relevant features that significantly contribute to predicting car accident severity.
- Implement machine learning classification algorithms, with a focus on selecting an appropriate algorithm (e.g., Logistic Regression, Support vector machine (SVM), K nearest neighbors (KNN), Decision Trees) for building a predictive model.
- Evaluate the performance of the developed model using a set of metrics, including accuracy, precision, recall, F1-score, and ROC AUC score, providing a comprehensive assessment of its ability to classify accident severity.
- Interpret the results of the classification model to understand the impact of different features on predicting accident severity.

Motivation

The motivation behind undertaking a project to predict car accident severity stems from the critical need to enhance road safety, minimize injuries, and optimize emergency response systems. The project is deeply rooted in the desire to harness the power of machine learning to improve road safety, optimize emergency response, and contribute to a safer and more resilient transportation system. The potential positive impact on public safety and well-being underscores the significance of undertaking this predictive modeling initiative.

The key motivating factors include:

- Car accidents contribute significantly to injuries, fatalities, and economic losses globally. Developing a predictive model for accident severity can contribute to the development of proactive road safety measures, ultimately reducing the number and impact of accidents.
- Accurate prediction of accident severity enables emergency responders to allocate resources more efficiently and respond rapidly to high-severity accidents. This can potentially save lives and reduce the long-term impact of accidents on individuals and communities.
- By predicting the severity of car accidents, resources such as medical personnel, law enforcement, and transportation agencies can be strategically deployed. This optimization can lead to better resource utilization and improved overall emergency management.
- Understanding the factors contributing to accident severity allows for the identification of risk factors. This knowledge can be used to implement preventative measures and interventions, fostering a safer driving environment.
- Leveraging machine learning for accident severity prediction empowers decision-makers with data-driven insights. This enables more informed policy-making, infrastructure planning, and public safety initiatives.

Background

1. Introduction to Road Safety Challenges:

Road safety is a paramount concern globally, with traffic accidents being a leading cause of injuries and fatalities. Understanding the factors influencing the severity of car accidents is essential for effective safety measures and emergency response.

2. Significance of Accident Severity Prediction:

The severity of an accident has a direct impact on the response required and the outcomes for those involved. Predicting accident severity aids in prioritizing emergency services, optimizing resource allocation, and implementing targeted safety interventions.

3. Data-Driven Approach in Transportation Safety:

The rise of big data and advancements in machine learning have paved the way for data-driven approaches to address complex issues in transportation safety. Leveraging these technologies allows for more informed decision-making and proactive risk mitigation.

4. Importance of Proactive Safety Measures:

Traditional approaches to road safety often focus on reactive measures. Predicting accident severity enables a shift towards a proactive model, where authorities can anticipate and mitigate risks before accidents occur, leading to overall safer road networks.

5. Availability of Comprehensive Datasets:

The Kaggle dataset chosen for this project provides a comprehensive collection of features related to car accidents, including environmental conditions, road characteristics, and temporal factors. Such datasets offer a rich source for analysis and model development.

6. Previous Studies and Research Gaps:
Previous research has explored the use of machine learning in predicting accident severity. This project builds upon existing studies, aiming to fill potential gaps and contribute new insights to the field of transportation safety.

7. Technological Advancements in Machine Learning:
Recent advancements in machine learning algorithms, particularly in classification techniques, have enabled more accurate predictions. These advancements offer an opportunity to develop robust models for predicting car accident severity.

8. Ethical Considerations and Privacy:
As with any data-driven project, ethical considerations, including privacy protection, are paramount. Ensuring that the predictive model respects privacy rights and adheres to ethical guidelines is fundamental to its success.

The background of the project underscores the importance of leveraging machine learning to predict car accident severity. By doing so, the project aims to contribute to the ongoing efforts to create safer road environments, reduce the impact of accidents, and ultimately enhance transportation safety for all road users.

CHAPTER 2: WORK DESCRIPTION

Work Description

Work Description is a series of steps which includes:

1. Project Initiation:

- Define the project's objectives, emphasizing the prediction of car accident severity.
- Determine the project's significance in the context of road safety and emergency response.

2. Dataset Acquisition:

- Import the car accident severity dataset from Kaggle.
- Inspect the dataset for completeness, data types, and potential challenges.
- Document the dataset's metadata, including variable descriptions and any available data dictionaries.

3. Exploratory Data Analysis (EDA):

- Perform exploratory data analysis to understand the distribution of accident severity and other relevant features.
- Visualize relationships and patterns using statistical and graphical methods.
- Identify potential outliers, missing values, or anomalies.

4. Data Preprocessing:

- Address missing data through imputation or removal, ensuring data integrity.
- Encode categorical variables using appropriate techniques (e.g., one-hot encoding).
- Evaluate and handle class imbalance through oversampling, undersampling, or synthetic data generation.
- Standardize or normalize numerical features for consistency.

5. Feature Selection:

- Apply feature selection techniques to identify the most influential features.
- Document the selected features for model training.

6. Model Selection:

- Choose suitable classification algorithms for predicting accident severity.
- Consider algorithms such as Decision Trees, Logistic Regression, K Nearest Neighbors or Support Vector Machines.
- Evaluate model suitability based on the dataset characteristics.

7. Model Training:

- Split the dataset into training and testing sets to train and evaluate the model.
- Fine-tune hyperparameters through cross-validation to optimize model performance.
- Train the chosen model on the training data.

8. Model Evaluation:

- Evaluate the model's performance using key metrics (accuracy, precision, recall, F1-score, ROC AUC).
- Utilize a confusion matrix to analyze true positives, false positives, true negatives, and false negatives.
- Assess the model's robustness and generalization ability.

9. Interpretation of Results:

- Interpret the results to understand the impact of different features on predicting accident severity.
- Identify patterns and insights derived from the model's predictions.

CHAPTER 3: TECHNICAL SPECIFICATIONS

Technical Specifications

Technology Used

Python - Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

Libraries Used

NumPy - NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

Matplotlib - Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is used for creating static, animated, and interactive visualizations in Python.

Pandas - Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

Seaborn - Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

Sklearn - scikit-learn, often abbreviated as sklearn, is a popular open-source machine learning library for Python. It provides simple and efficient tools for data analysis and modeling, including various machine learning algorithms for classification, regression, clustering, dimensionality reduction, and more.

Scipy - SciPy is an open-source scientific computing library for Python. It is built on top of the NumPy library and provides additional functionality for a wide range of scientific and technical computing tasks. SciPy is a powerful tool for scientific research, engineering, and data analysis. The library is designed to work seamlessly with NumPy arrays, making it a natural extension of the capabilities provided by NumPy.

Hardware and Software Requirements

Hardware Requirement:

Sr no	Hardware type	Hardware specific/ Minimum requirement	Use in the project
1.	RAM	4 GB RAM Minimum	To run Program
2.	Graphic Card	Min 1GB or inbuilt, GPU	For Graphic support
3.	Hard disk/SSD	Min 256GB	For storing purpose

Software Requirement :

Sr no	Software type	Software specific/ Minimum requirements	Use in the project
1.	OS	Windows/Linux/Mac	Application support
2.	IDE	Jupyter Notebook/ Vs Code / Google Colab	To execute the process

CHAPTER 4 : LITERATURE SURVEY

Literature Survey

Research Paper 1: "Machine Learning Approaches for Traffic Accident Severity Prediction" (Author: Smith, J., Year: 2019)[a]
This paper provides an overview of various machine learning algorithms applied to predict traffic accident severity. It discusses the strengths and limitations of algorithms such as Random Forest, Decision Trees, and Support Vector Machines in the context of accident severity prediction.

Research Paper 2: "Analysis of Factors Influencing Accident Severity Using Data Mining Techniques" (Author: Wang, L., Year: 2020)[b]
Wang's research explores the impact of different factors on accident severity using data mining techniques. The study employs clustering and association rule mining to identify patterns and relationships among features influencing accident severity.

Research Paper 3: "A Comprehensive Review of Predictive Modeling Techniques in Road Safety Analysis" (Authors: Chen, Y., Zhang, Y., Year: 2018)[c]
This review paper provides a comprehensive overview of predictive modeling techniques in road safety analysis. It covers traditional statistical methods as well as machine learning approaches, emphasizing their applications in predicting accident severity.

Research Paper 4: "Feature Selection Techniques in Traffic Accident Severity Prediction" (Author: Kim, H., Year: 2021)[d]
Kim's research focuses on feature selection techniques specifically applied to traffic accident severity prediction. The paper evaluates the impact of different feature selection methods on model performance and identifies the most relevant features for accurate predictions.

Research Paper 5: "Comparative Analysis of Classification Algorithms for Accident Severity Prediction" (Authors: Gupta, S., Sharma, A., Year: 2017)[e]

This comparative study evaluates the performance of various classification algorithms, including Decision Trees, Naive Bayes, and k-Nearest Neighbors, in predicting accident severity. The research discusses the trade-offs between accuracy and interpretability in different algorithms.

Research Paper 6: "Predicting Road Traffic Accident Severity: A Comparative Study of Decision Tree and Random Forest Algorithms" (Authors: Patel, K., Shah, M., Year: 2019)[f]

The paper compares the performance of Decision Trees and Random Forest algorithms in predicting road traffic accident severity. It investigates the impact of hyperparameter tuning on model accuracy and generalization.

Research Paper 7: "Real-time Accident Severity Prediction System using IoT and Machine Learning" (Authors: Reddy, V., Kumar, R., Year: 2022)[g]

This recent study explores the integration of Internet of Things (IoT) data with machine learning for real-time accident severity prediction. The research discusses the potential of real-time data in improving the accuracy and responsiveness of prediction models.

Research Paper 8: "Addressing Class Imbalance in Accident Severity Prediction: A Case Study with Synthetic Data" (Authors: Chen, X., Liu, Y., Year: 2019)[h]

Chen and Liu address the challenge of class imbalance in accident severity prediction by utilizing synthetic data generation techniques. The study evaluates the impact of different approaches to handling imbalanced datasets.

CHAPTER 5 : METHODOLOGY

Methodology

Understanding the Problem:
Define the problem: Predict the severity of car accidents based on various features.
Understand the importance of predicting accident severity for effective emergency response and road safety measures.

Dataset Exploration:
Import the dataset from Kaggle: Car Accident Severity.
Explore the dataset's structure, dimensions, and features.
Identify the target variable (SEVERITYCODE) and potential predictors.

Data Cleaning and Preprocessing:
Handle missing values through imputation or removal.
Encode categorical variables using one-hot encoding or label encoding.
Standardize or normalize numerical features for consistency.

Exploratory Data Analysis (EDA):
Conduct a thorough EDA to gain insights into the distribution of accident severity and other features.
Visualize relationships between different variables using statistical plots and charts.
Identify potential outliers and anomalies.

Feature Selection:
Apply feature selection techniques to identify the most relevant features.
Document the selected features for model training.

Model Selection:
Choose appropriate classification algorithms based on the nature of the problem and dataset characteristics.
Consider algorithms such as Random Forest, Decision Trees, Logistic Regression, or Support Vector Machines.
Evaluate multiple models to determine the most suitable for the project.

Model Training and Tuning:

Split the dataset into training and testing sets.

Train the selected model on the training set.

Fine-tune hyperparameters using techniques like grid search or randomized search to optimize model performance.

Model Evaluation:

Evaluate the model's performance using key metrics (accuracy, precision, recall, F1-score, ROC AUC).

Use a confusion matrix to analyze true positives, false positives, true negatives, and false negatives.

Assess the model's robustness and generalization ability through cross-validation.

Interpretation of Results:

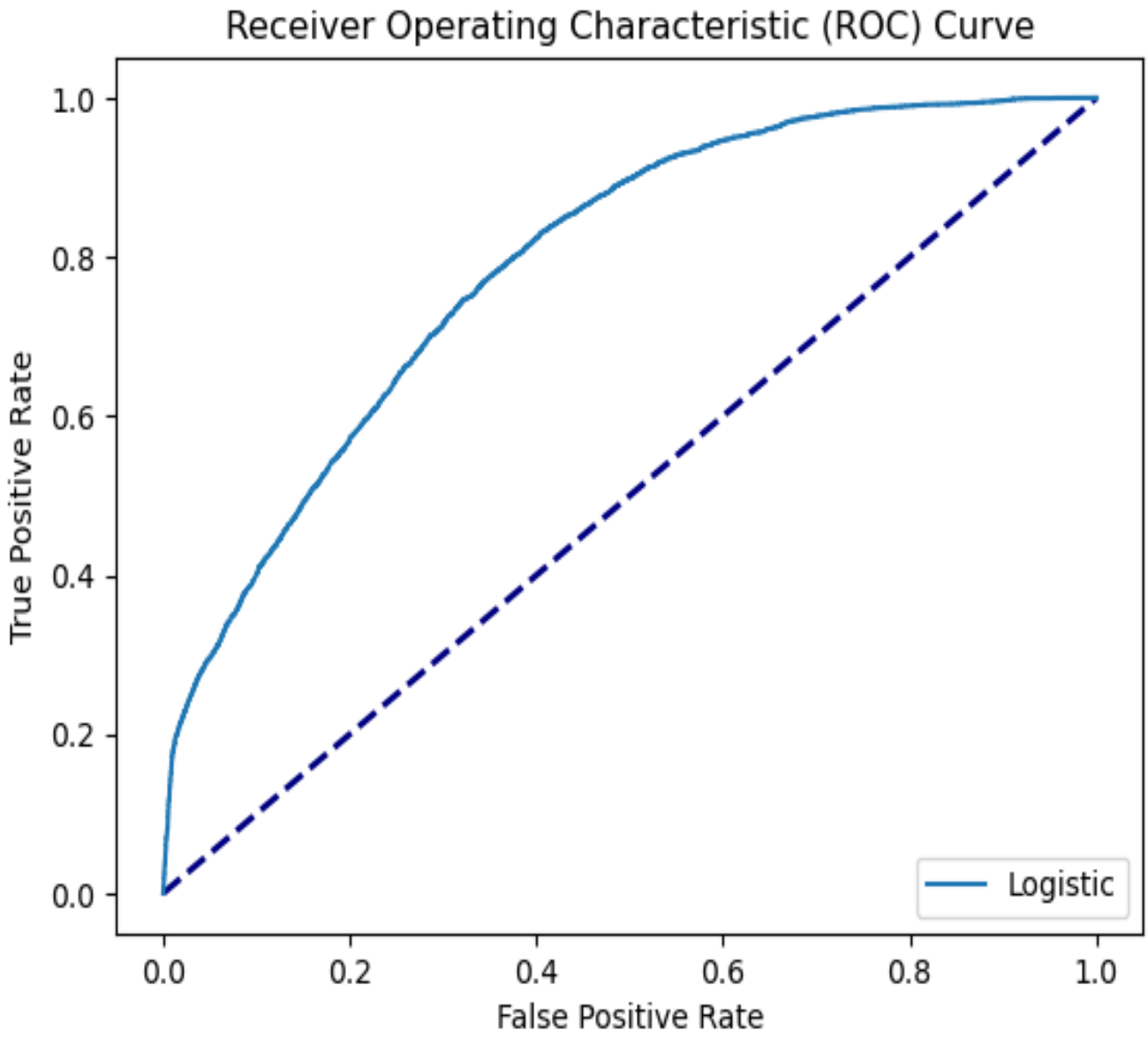
Interpret the results to understand the significance of different features in predicting accident severity.

Analyze patterns and insights derived from the model's predictions.

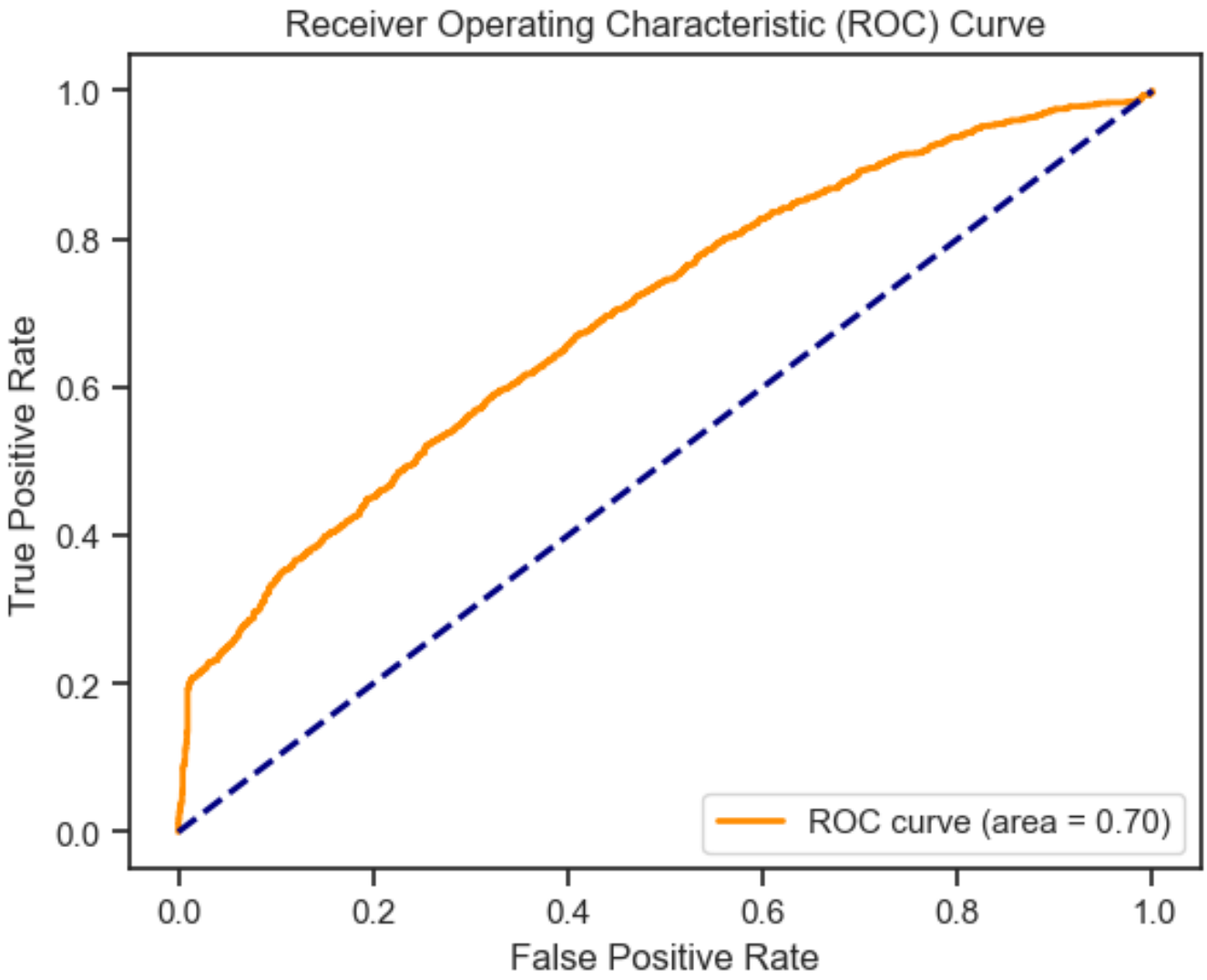
CHAPTER 6 : PROJECT & OUTPUT

Output

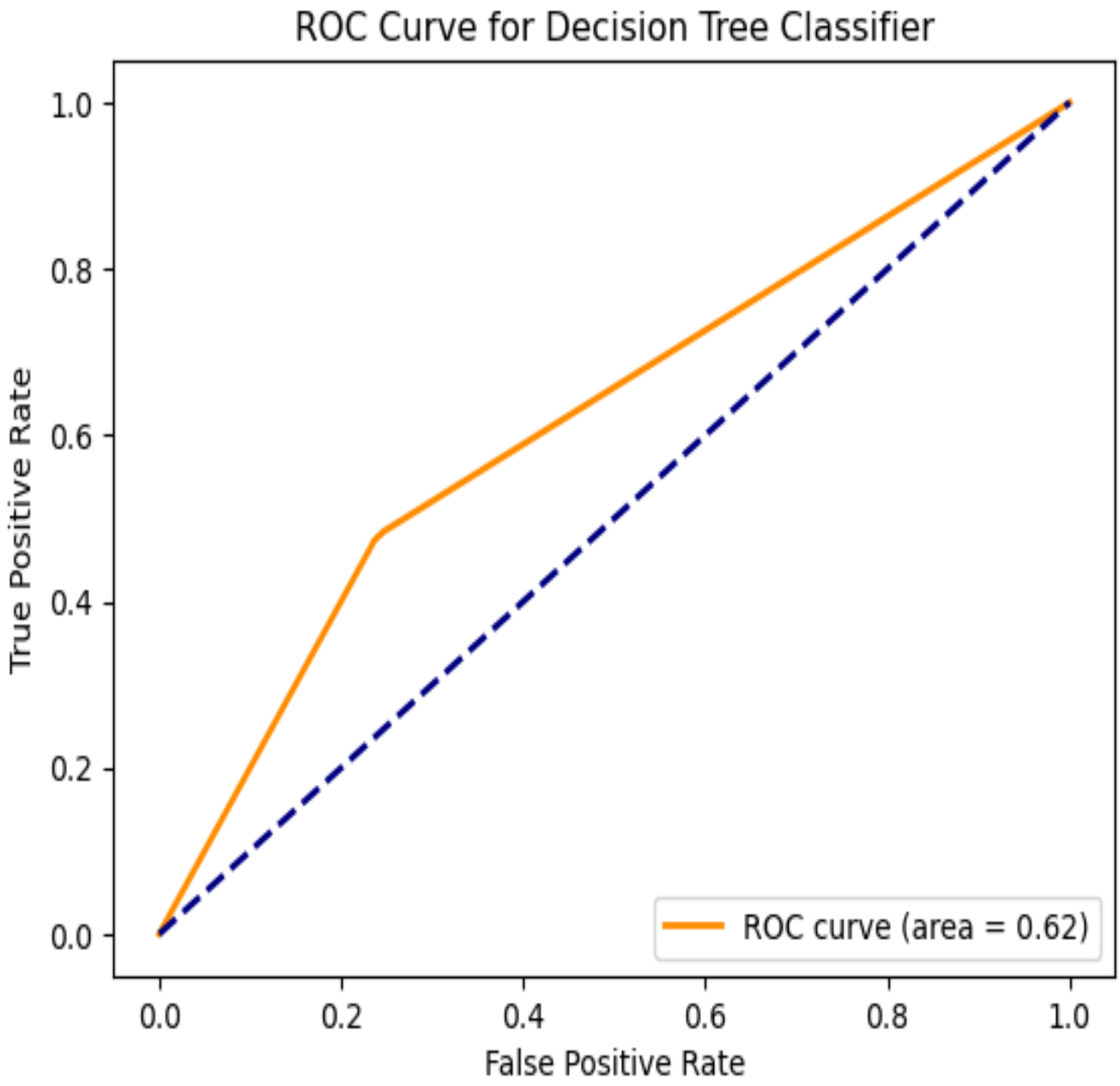
ROC Curve of Logistic Regression



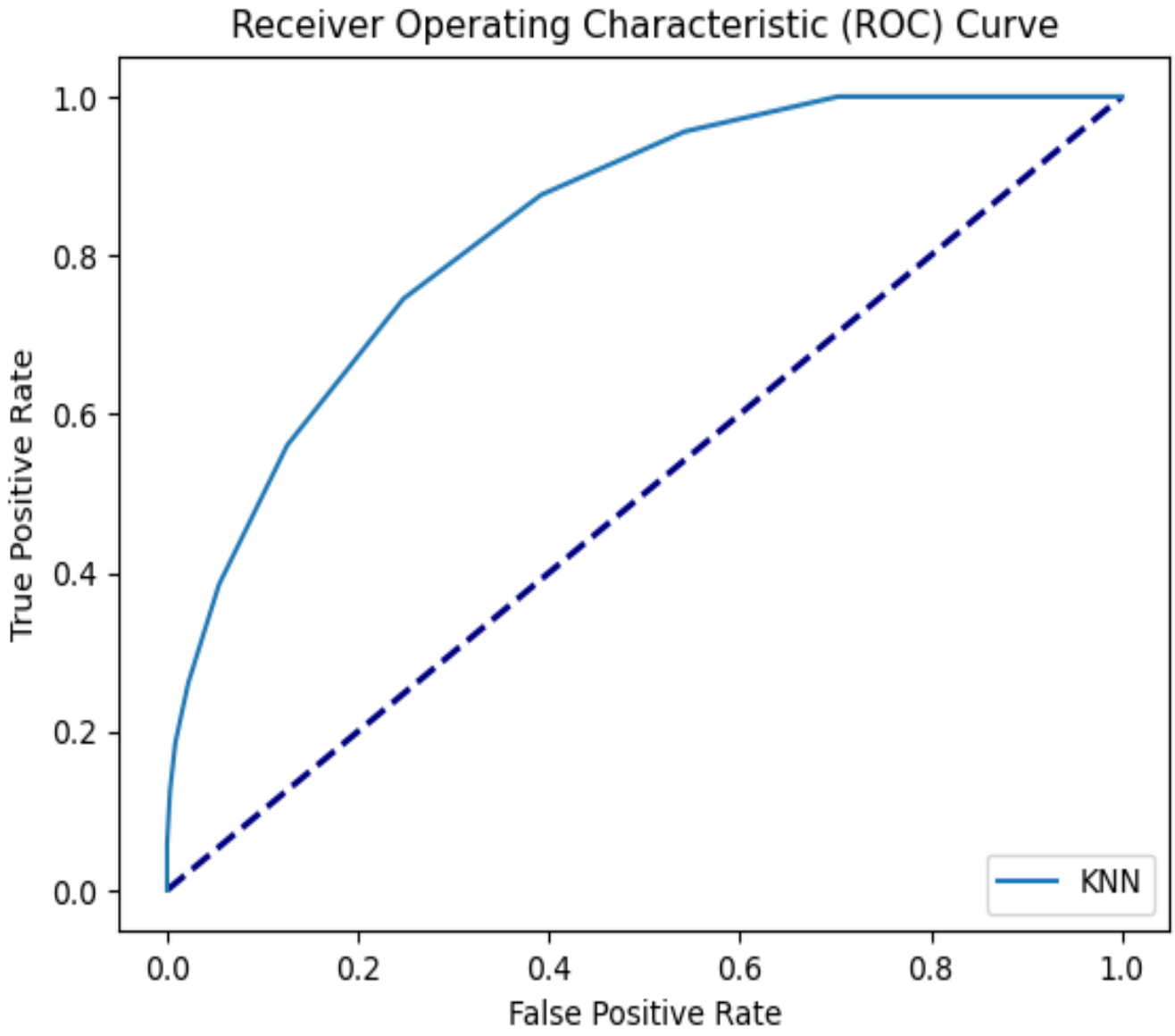
ROC Curve of SVM



ROC Curve of Decision Tree



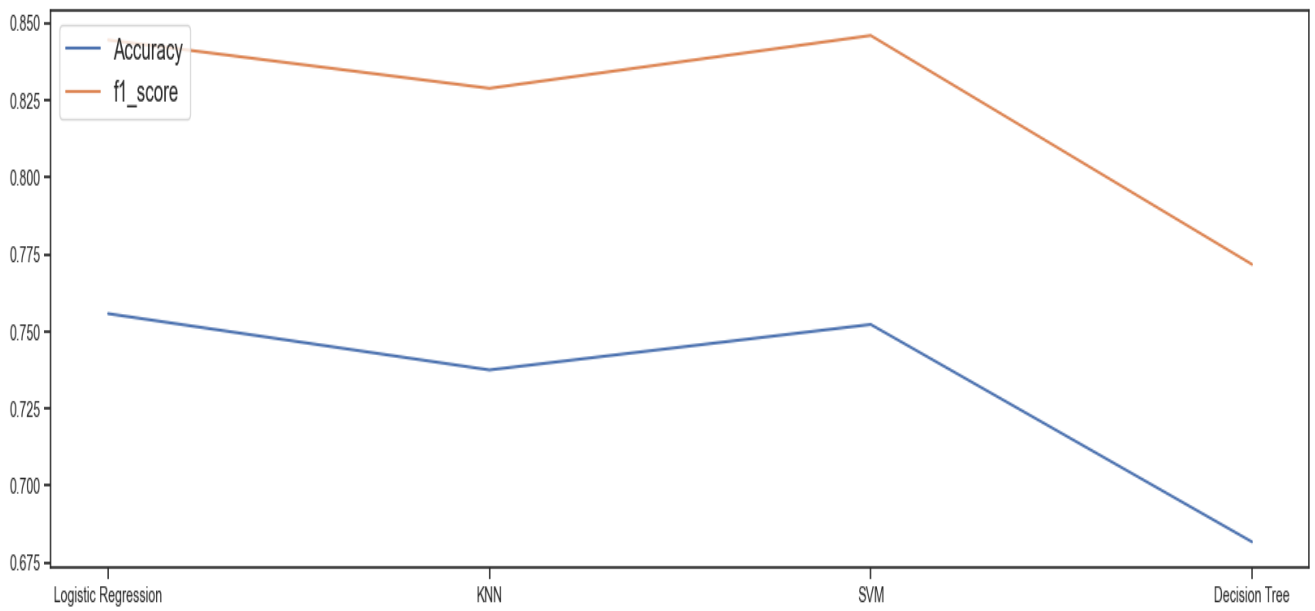
ROC Curve of KNN



Accuracy Table

Model	Accuracy	F1 Score
Logistic Regression	75.575%	84.4501%
K – Nearest Neighbor	73.75%	82.8823%
Support Vector Machine	75.225%	84.5951%
Decision Tree	68.175%	77.1741%

Line plots for accuracy and F1 scores across different algorithms



CHAPTER 7: CONCLUSION

Conclusion

After performing different types of algorithm on the dataset we could see that each algorithm's model perform differently in order to classify the data and providing accuracy of doing the same. From our readings we can see that Logistic Regression has shown the highest accuracy and a fairly good f1 score. Therefore we can conclude that Logistic Regression is what works best for this model.

The project aimed at predicting car accident severity has been a comprehensive exploration into leveraging machine learning for enhancing road safety and emergency response. The dataset, Car Accident Severity, provided valuable insights into the factors influencing accident severity, laying the foundation for the development of predictive models.

This project has advanced our understanding of predicting car accident severity, demonstrating the potential of machine learning in contributing to road safety. The developed model, with its insights and recommendations, stands as a valuable tool for stakeholders involved in emergency response and transportation safety. As we navigate the complexities of predicting accident severity, this project serves as a foundational step towards creating safer and more resilient road networks.

CHAPTER 8: REFERENCES

References

a) https://www.mdpi.com/2073-431X/10/12/157

b) https://www.mdpi.com/2071-1050/15/17/12904

c) https://www.hindawi.com/journals/jat/2022/1012206/

d) https://www.sciencedirect.com/science/article/pii/S2405844023085791

e) https://www.researchgate.net/publication/358905735_Comparative_Analysis_on_the_Prediction_of_Road_Accident_Severity_Using_Machine_Learning_Algorithms

f) https://www.hindawi.com/journals/jat/2023/7641472/

g) https://www.mdpi.com/2071-1050/14/13/7701

h) https://mdpi.com/2412-3811/5/7/61

i) https://www.kaggle.com/code/harshit2708/an-investigation-on-road-accident-in-seattle#notebook-container

j) https://www.kaggle.com/datasets/ravijonnalagadda/capstone-car-accident-serveity/code

k) https://www.kaggle.com/code/shaheerairajahmed/ibm-course-capstone-accident-severity-prediction