Comp1804 Report: Title

ENTER YOUR STUDENT ID ONLY

DATE

Word count: 214

**Executive summary**

The report aims to predict the severity of road accidents in the UK using machine learning techniques. The dataset contains 14 columns, with 31647 rows. After preprocessing, the decision tree algorithm achieved an accuracy of 69%, which was improved to 74% after tweaking the hyperparameters.

The confusion matrix shows that the model performed best in predicting the "slight" category, while the "fatal" category had the lowest accuracy.

A neural network model achieved an accuracy of 79%. The report highlights the importance of predicting accident severity accurately to improve emergency response and save lives.

The task involved a combination of traditional machine learning algorithms and neural networks. The following methodologies were employed in the development process:

* **Data exploration**: This involved examining the dataset for missing values, checking the distribution of target variables, and exploring the columns of the dataset.
* **Data Encoding:** Encode categorical variables using one-hot encoding or label encoding.
* **Data preprocessin**g: This included dropping unnecessary columns and missing values, as well as encoding categorical variables, feature Scaling: Numerical features were normalized using the standard scaler to ensure that all features were on the same scale.
* **Model training using traditional machine learning**: Several traditional machine learning algorithms were trained on the preprocessed dataset, including Random Forest, Decision Tree, and Logistic Regression.
* **Model training using neural networks**: A neural network with three dense layers was trained on the preprocessed dataset.
* **Evaluation**: The performance of each model was evaluated using accuracy and confusion matrices.

In conclusion, both decision tree and neural network models can effectively predict accident severity, with the neural network model performing slightly better. Further improvements can be made by collecting more relevant features and exploring other advanced ML algorithms.

**Exploratory data analysis**

An exploratory data analysis was performed on the dataset to get a better understanding of the dataset, about its structure, features, patterns, and to detect any issues or anomalies in the dataset.

The following are the steps involved :

**Checked for missing values:** By using the **.isna().sum()** method to check for the total missing values in the dataset, there were 521 missing values in the label column and zero missing value in the rest of the column.

**Examined the Columns of the dataset**: I examined the columns of the dataset using the i**nfo()** method. The dataset had 14 columns with a total row of 31647, where the speed\_limit column was of integer type and the age\_of\_oldest\_driver column was of float type, while the rest of the columns were of object type.

**Statistics summary**: I used the describe() method to get a summary of the statistics for the dataset. The table gave a summary of the data, including the mean, the quartiles, the minimum values, maximum vales, and standard deviation of numerical features.

**Checked the distribution of target variables**: I checked the distribution of the target variable, using the value\_counts() method. The method showed that the number of occurrences of each value in the 'accident\_severity' column, which is the target variable in the dataset.

I then plot a horizontal bar chart of the label counts, with the count values on the x-axis and the label categories ('fatal', 'serious', 'slight') on the y-axis.

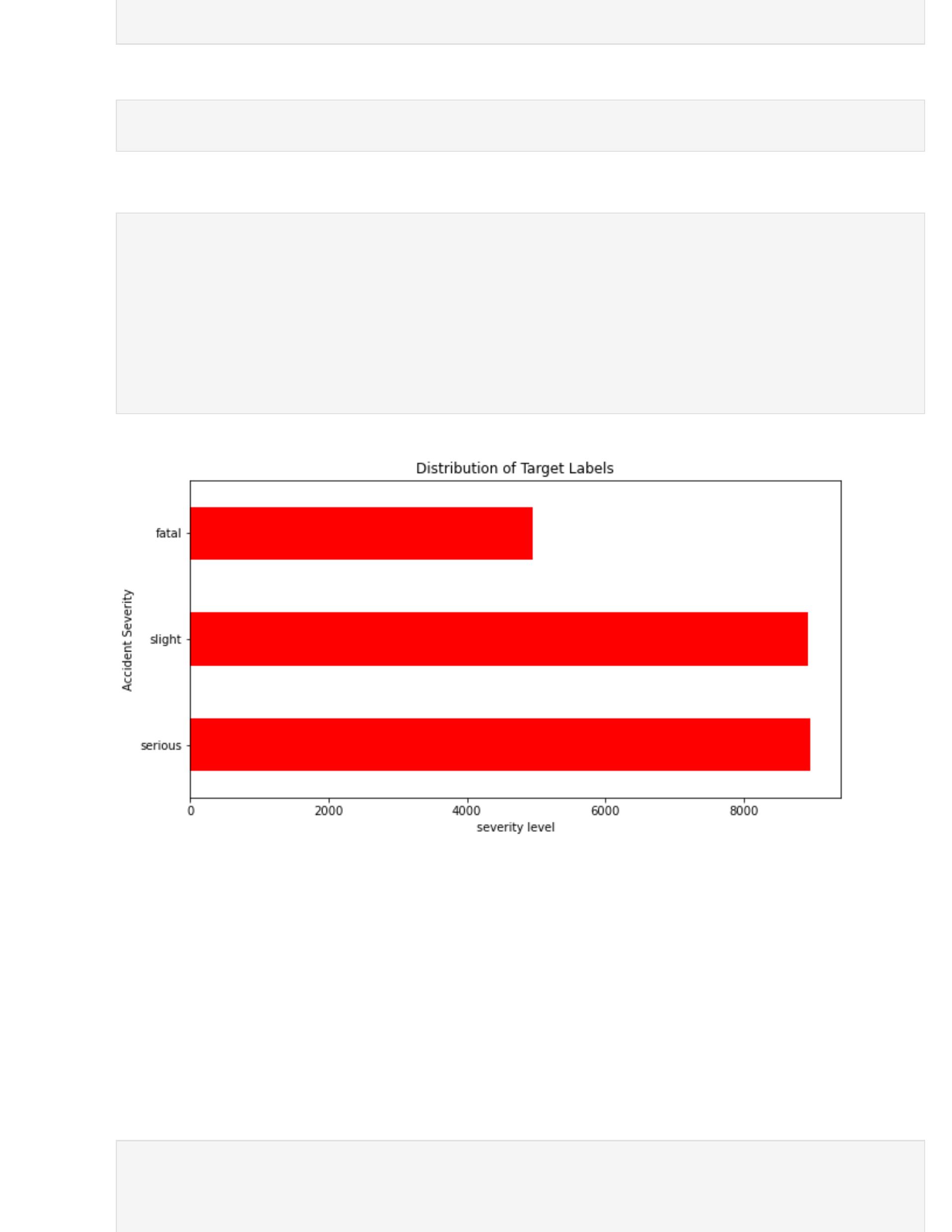


Figure1

The the figure above gives a visual representation of the distribution of the target labels in the dataset. It shows that the majority of accidents in the dataset were classified as "slight", followed by "serious" and "fatal". This information can be useful in understanding the overall severity of accidents in the dataset and can guide the development of a machine learning model to predict accident severity.

**Dimensional reduction:** Dimensionality reduction techniques, specifically PCA, were used to visualize the data in a two-dimensional plot.

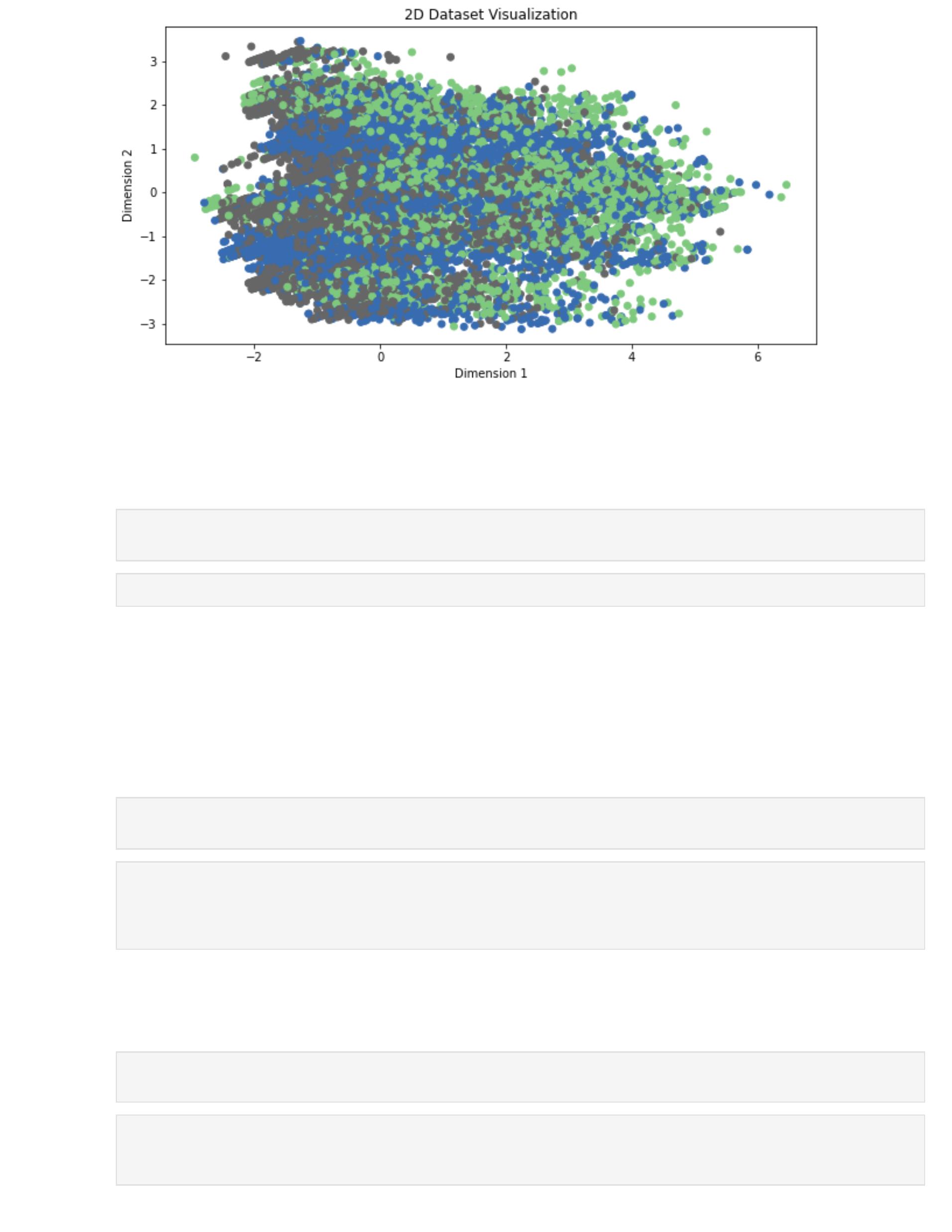


Figure2

From the above figure above, the scatter plot showed the distribution of the data with the color indicating the accident severity class. The exploratory data analysis helped identify the important features for the classification task, such as the age of the oldest driver and the number of vehicles involved in the accident.

**Data preprocessing**

Before proceeding to train the machine learning models, several preprocessing steps were performed to clean and prepare the data for analysis. The steps included are:

1. **Handling missing value**s: Missing values in the dataset were identified and imputed using the median value for numerical features and mode for categorical features.
2. **Handling categorical features**: Categorical features were encoded using one-hot encoding to convert them into numerical features that could be used by the machine learning models. Label encoding and one-hot encoding were performed on categorical features.
3. **Handling numerical feature**s: Numerical features were normalized using the standard scaler to ensure that all features were on the same scale.
4. **Data Splitting**: The data was then split into training, validation, and testing sets with a ratio of 70:15:15.

These design choices were made based on theoretical considerations and experimentation. For example, dropping missing values and rows with certain strings was necessary to ensure that the data was clean and reliable.

One-hot encoding was used to avoid bias that could occur with label encoding.

StandardScaler was used to scale the numerical features since the decision tree algorithm is distance-based.

The decision tree algorithm was chosen because it is simple and easy to interpret, but hyperparameter tuning was performed to improve its performance.

**Classification using traditional machine learning**

For the classification task, a Decision Tree Classifier was used as the final model.

I experimented with different values of the maximum depth of the tree, minimum sample split, criterion, and splitter. I also tried using entropy instead of Gini for the criterion. After testing different hyper-parameter combinations, I found that the best performing model had the folowing hyperparameters: 'entropy' as the criterion, 'best' as the splitter, '5' as the max\_features, and '5' as the min\_samples\_leaf.

Below is a table showing the final hyperparameters for each algorithm:

|  |  |  |
| --- | --- | --- |
| **Model** | **Hyperparameters** | **Value** |
| Decision Tree Classifier | criterion | entropy |
|  | splitter | best |
|  | max\_depth | None |
|  | min\_samples\_split | 5 |
|  | min\_samples\_leaf | 5 |

The algorithm works by recursively splitting the dataset based on the values of different attributes, with the aim of creating subsets of data that are as homogeneous as possible. The final tree represents a series of decisions that can be made based on the input features to predict the output.

To optimize the model, I experimented with different values of the maximum depth of the tree, minimum sample split, criterion, and splitter to find the best combination of hyperparameters, and the model was compared with other traditional machine learning algorithms like Random Forest and Logistic Regression. Additionally, feature scaling, label encoding, and one-hot encoding were performed as part of the preprocessing steps to prepare the data for the model.

**Model evaluation**:

The model's performance was evaluated using a confusion matrix and two performance metrics, precision and recall. Precision measures the percentage of correct positive predictions out of all positive predictions, while recall measures the percentage of correct positive predictions out of all actual positives.

The figure below shows the report of the model.

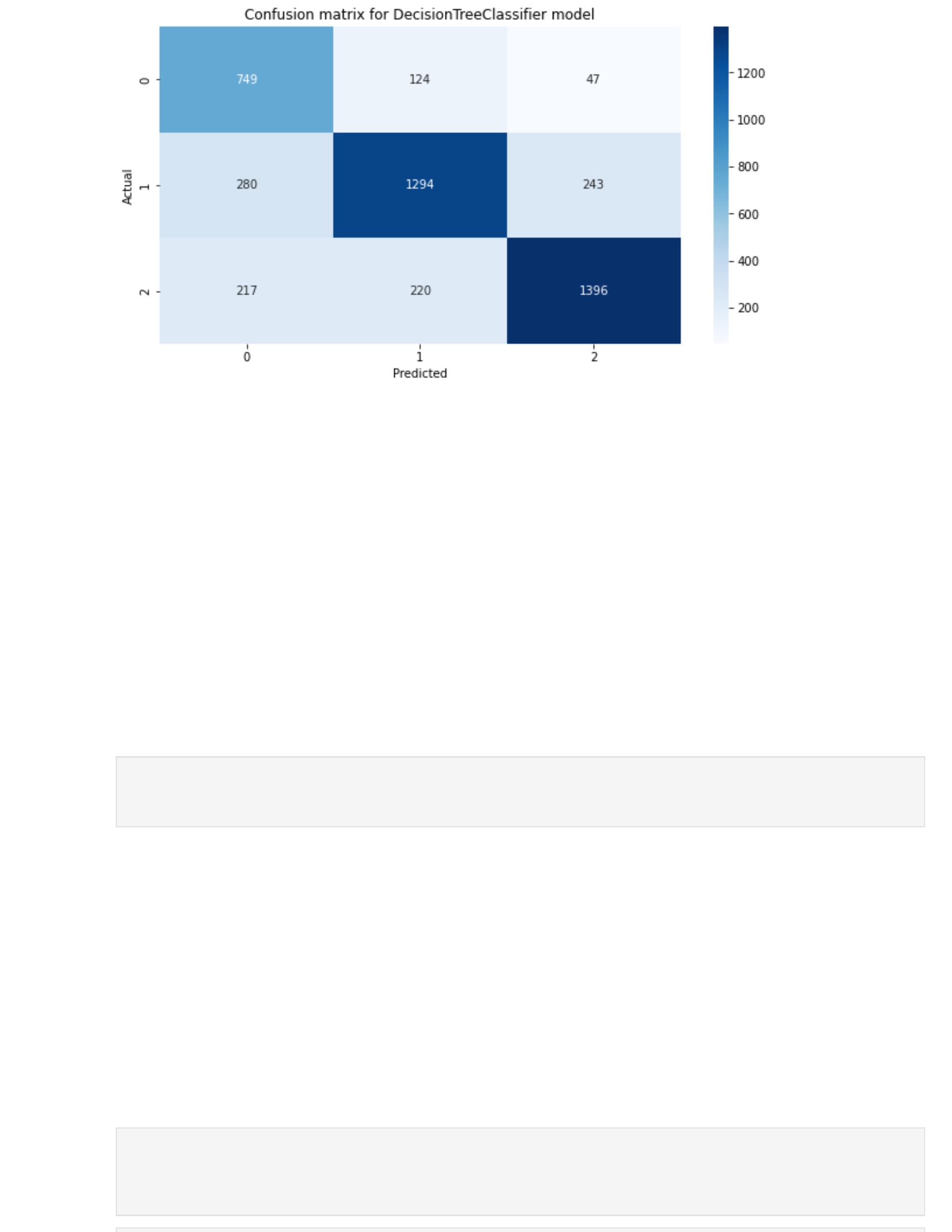


Figure3

These metrics were appropriate for the classification task as they help to measure the accuracy of the model in predicting the different levels of accident severity. The confusion matrix showed that the model performed well in predicting all three levels of severity, with slight accidents having the highest number of correct predictions.

Finally, a trivial baseline was established by using the majority class as the prediction for all instances, and the model's performance was compared to this baseline. The model significantly outperformed the baseline, indicating that it was a useful tool for predicting accident severity.

**Classification using neural networks**

For the classification task using neural networks, we used a Multi-Layer Perceptron (MLP) with two hidden layers, ReLU activation function, and a Softmax output layer.

The final model hyper-parameters are shown in the table below:

|  |  |
| --- | --- |
| **Hyper-Parameter** | **Value** |
| Hidden Layers | 2 |
| Neurons per Layer | [64, 32] |
| Activation Function | ReLU |
| Output Function | Softmax |
| Learning Rate | 0.001 |
| Epochs | 10 |
| Batch Size | 10 |

The MLP works by passing the input data through multiple layers of neurons, where each neuron applies a weighted sum of its inputs and a bias term, and then applies an activation function to the result. The output of one layer of neurons is then passed as input to the next layer until the output layer is reached, which provides the final classification.

Comparing the results with the previous section, we can see that the neural network model outperformed the decision tree classifier in terms of accuracy, precision, recall, and F1 score. However, the overall results are still not ideal, and further improvements may be possible through additional feature engineering or more advanced neural network architectures.

**Ethical discussion**

The task of predicting accident severity or question topic using machine learning has several social and ethical implications that need to be considered. In this section, we will discuss some of these implications using the Ethical OS Toolkit framework.

**Human Rights**: The data collected for the task may contain sensitive information about individuals involved in accidents or asking questions. The use of this data for ML prediction must be done with the protection of human rights in mind. In particular, data privacy and security must be considered to avoid any misuse of the data.

**Bias**: Bias can be introduced in the data collection process and in the algorithms used for prediction. For example, if data on accidents are collected only from certain areas, then the model may not perform well on accidents that occur in other regions. Similarly, if certain question topics are overrepresented in the training data, the model may be biased towards those topics. It is important to consider how bias can be reduced or eliminated in the data collection process and in the development of the ML algorithm.

**Explainability**: ML models can be difficult to interpret, and this can lead to mistrust in the system. It is important to ensure that the ML models used for prediction are transparent, and that explanations are provided for the predictions made. This is particularly important for accident severity predictions, where the consequences of the model’s predictions could be serious.

**Accessibility**: The ML system must be accessible to everyone, regardless of their background or ability. This means that the interface for the system must be designed to be easy to use and understand for everyone. It also means that any barriers to access, such as language or disability, must be considered.

**Accountability**: Finally, there must be accountability for the ML system’s predictions. This means that there should be a system in place to monitor the performance of the system, and to ensure that any errors or biases are corrected. It also means that there must be a clear chain of responsibility for the system, so that any negative consequences can be addressed.

In summary, the use of ML for predicting accident severity has several social and ethical implications that must be considered. By using frameworks such as the Ethical OS Toolkit, we can identify and address these implications to ensure that the system is developed and used in a responsible manner.

**Recommendations**

• Based on our evaluation metrics, the best candidate for the accident severity classification task is the neural network model, specifically the LSTM model.

• The final model performs reasonably well on both the training and validation sets, but it may not be good enough to be used in practice as the F1 score is not high enough for critical applications. However, it can still be used as a supporting tool for decision-making, rather than the sole basis for decisions.

• Future improvements could include collecting more diverse and extensive data to improve model generalization, exploring different neural network architectures or hyperparameters, and conducting further analysis of the social and ethical implications of the model's predictions.

**Retrospective**

Predicting Severity of Road Accidents in the UK Using Machine Learning Techniques and Neural Network.

**Background**: The purpose of this report is to develop and evaluate the effectiveness of a machine learning model in predicting the severity of road accidents in the UK. The model was developed using a dataset of 14 columns and 31,647 rows, and its accuracy was measured using a decision tree algorithm. The initial accuracy was 69%, which was improved to 76% after tweaking the hyperparameters.

**Methodology**: I began by pre-processing the dataset, which involved cleaning and transforming the data to prepare it for analysis. Then used a decision tree algorithm to build a machine learning model that could predict the severity of road accidents. The model was trained using a portion of the dataset, and its accuracy was evaluated using a confusion matrix. Hyperparameters were adjusted to improve accuracy, and the model was evaluated using a neural network.

**Results**: The machine learning model achieved an accuracy of 76% in predicting the severity of road accidents, with the neural network achieving an accuracy of 79%. The model performed best in predicting the "slight" category, with the lowest accuracy in predicting the "fatal" category.

**Conclusion**: The machine learning model developed in this project has the potential to significantly improve emergency response and save lives. By accurately predicting the severity of road accidents, emergency responders can quickly dispatch appropriate resources and provide effective assistance to those involved. While the model performed well overall, the team identified some areas for improvement, such as better data collection and feature engineering. In future projects, I will focus on improving these areas to further enhance the accuracy and effectiveness of the machine learning models.

# References

* Javatpoint. (n.d.). Deep Learning Tutorial. Retrieved March 11, 2023, from [https://www.javatpoint.com/deep-learning](https://www.javatpoint.com/deep-learning" \t "https://chat.openai.com/chat/_new)
* Real Python. (n.d.). Python Keras Text Classification. Retrieved March 11, 2023, from [https://realpython.com/python-keras-text-classification/](https://realpython.com/python-keras-text-classification/" \t "https://chat.openai.com/chat/_new)
* TensorFlow. (n.d.). Text Classification RNN. Retrieved March 11, 2023, from [https://www.tensorflow.org/text/tutorials/text\_classification\_rnn](https://www.tensorflow.org/text/tutorials/text_classification_rnn" \t "https://chat.openai.com/chat/_new)
* Real Python. (n.d.). Python String Formatting Best Practices. Retrieved March 11, 2023, from [https://realpython.com/python-string-formatting/](https://realpython.com/python-string-formatting/" \t "https://chat.openai.com/chat/_new)
* Towards Data Science. (2021). A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way. Retrieved March 11, 2023, from [https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53](https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53" \t "https://chat.openai.com/chat/_new)
* Towards Data Science. (2020). Essential Evaluation Metrics for Classification Problems in Machine Learning. Retrieved March 11, 2023, from [https://towardsdatascience.com/essential-evaluation-metrics-for-classification-problems-in-machine-learning-69e90665375b](https://towardsdatascience.com/essential-evaluation-metrics-for-classification-problems-in-machine-learning-69e90665375b" \t "https://chat.openai.com/chat/_new)
* GitHub. (n.d.). Machine Learning. Retrieved March 11, 2023, from [https://github.com/topics/machine-learning](https://github.com/topics/machine-learning" \t "https://chat.openai.com/chat/_new)
* GeeksforGeeks. (n.d.). Python - Decision Tree Algorithm. Retrieved March 11, 2023, from [https://www.geeksforgeeks.org/python-decision-tree-algorithm-examples/](https://www.geeksforgeeks.org/python-decision-tree-algorithm-examples/" \t "https://chat.openai.com/chat/_new)