# Comparative Study of Machine Learning and Deep Learning Approaches for Predicting Irish Road Accidents

MSc Research Project
Data Analytics

Karthika Rajan
Student ID: X21122920

School of Computing
National College of Ireland

Supervisor: Muslim Jameel Syed

## National College of Ireland
## Project Submission Sheet
## School of Computing

| | |
|---|---|
| **Student Name:** | Karthika Rajan |
| **Student ID:** | X21122920 |
| **Programme:** | Data Analytics |
| **Year:** | 2023 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Muslim Jameel Syed |
| **Submission Due Date:** | 18/09/2023 |
| **Project Title:** | Comparative Study of Machine Learning and Deep Learning Approaches for Predicting Irish Road Accidents |
| **Word Count:** | 9000 |
| **Page Count:** | 33 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Karthika Rajan |
| **Date:** | 17th September 2023 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | □ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Comparative Study of Machine Learning and Deep Learning Approaches for Predicting Irish Road Accidents

Karthika Rajan
X21122920

**Abstract**

In past few years, the crashes on roads in Ireland have become an increasing concern. According to Road Safety Authority of Ireland (RSA), number of crashes in Ireland has increased by 13 percentage, demonstrating the need for further measures for lowering the crash rate. The purpose of this research is to perform a comparison between several machine learning and deep learning algorithms on accident dataset of Ireland in order to determine the ideal technique for accident forecasting. This project also intends to determine if hyperparameter optimisation can improve the performance of the ML techniques. The evaluated machine learning algorithms are Random Forest, Decision Tree, XGBoost, Ridge Regression and KNN, whereas the deep learning models are Long Short-Term Memory (LSTM) and Feedforward Neural Network (FNN). Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-squared score are used to measure the effectiveness of the models. The results indicate that the hyperparameter optimisation enhanced the performance of the machine learning models significantly. After tuning their hyperparameters, XGBoost outperformed other models followed by Random Forest and Ridge Regression. XGBoost has got an R-squared score of 0.96. KNN algorithm underperformed when compared to other models. For deep learning models, FNN model performed better than LSTM algorithm. This study offers important insights into the application of machine learning and deep learning models for prediction of traffic accidents in Ireland. The results can definitely help the policymakers and the road safety authorities to allocate resources more efficiently. By this it is possible to reduce the accidents and can enhance road safety in the nation.

# 1 Introduction

## 1.1 Background and Motivation

In many nations, development of economy and the society is facilitated by transportation which is made possible by motor vehicles. Nevertheless, millions of casualties and fatalities are caused by vehicular crashes every year.*Global Road Safety* (2023) has released an analysis of vehicular collisions happening each and every year. According to this report collisions involving vehicle including cars, bus, motorbikes, or pedestrians result in the deaths of up to 3,700 individuals every day globally. Alarmingly, about half of the killed were either a motorcyclists, pedestrians, or a cyclists. As per the estimates, road accidents are considered the seventh most common cause of death in the world across all age categories. It is predicted that from 2015 to 2030, both the fatal and the nonfatal collision damages will cost global economy 1.8 trillion dollars. Also, the report indicates that deaths from traffic accidents now outnumber the mortality from HIV/AIDS, highlighting the seriousness of the problem with road safety.

As of 2021, nearly 2.9 million vehicles were registered in Ireland. This figure has increased consistently throughout the past few decades due to the expansion of population, economy, and the urbanisation. Because of rising traffic congestion on Ireland's roads in urban areas brought on by an increase in the amount of automobiles, issues over road safety and the requirement for better infrastructure has been arisen. Based on early reports from the Irish Police, RSA (2023) has issued a report regarding Irish road accidents that is up-to-date as of June 2023. As of December 31, 2022, there had been 150 fatal collisions on Ireland roads, claiming approximately 156 lives. In comparison to provisional Garda figures for the year 2021, this equates to 26 more fatal collisions and about 20 more fatalities. Some of the key reasons for  rise in the traffic accidents includes speeding, drunk driving, and non-compliance with thetraffic safety regulations. In order to address the issue of traffic accidents, the Irish government and road safety authorities have adopted a number of efforts, including awareness campaigns, greater enforcement of traffic regulations, and modifications in the infrastructure of the road. Despite efforts to improve the traffic safety and to reduce the fatality rates, there are still many concern regarding the traffic accidents in Ireland.

Various research studies have been done for the forecasting of road crashes in many countries. The application of machine learning methods to forecast traffic accidents and alert people about danger in new environments is discussed in the study by Alagarsamy et al. (2021). A combination of random forest and Gaussian distribution algorithms is employed on the research paper. To forecast traffic accidents, the model makes use of user behaviour data, which includes traffic characteristics and breaches of traffic laws. The study offers crucial information and accurate predictions of accidents as part of a proactive and data-driven solution to improve safety on the roads.

Even though a lot of studies have been conducted for the prediction of road accidents in many other countries, a study exclusively focused on the forecast of traffic accidents in Ireland is required. Because Ireland contains both urban and rural locations, and each poses different traffic problems.In the urban areas of Ireland, traffic density, congestion, and count of vulnerable road users such as the pedestrians and cyclists are frequently increased. On the other hand, the rural areas may have wider roads and distinct road conditions, which can influence the accident patterns.Because of these differences, it is very important to understand Irish road system and how complicated it is in order to make accurate predictive models that fit country's needs. Predictive models can help forecasting the traffic accidents, which will enable authorities to allocate resources more effectively to lower accidents.

## 1.2 Research Question

**What is the comparative predictive effectiveness of machine learning models versus deep learning models in forecasting traffic collisions in Ireland, and can the performance of machine learning algorithms for this purpose be enhanced through hyperparameter optimization?**

## 1.3 Research Objectives

The objective of the research paper are as follows:

- To conduct a comparative analysis of both the machine learning and the deep learning models to evaluate the forecasting performance for the traffic collision in Ireland. This involves performing experiments and evaluating the findings to figure out how effectively each approach predicts traffic accidents.
- To perform experiments to find out if hyperparameter optimization can improve the performance of the ML algorithms.

By accomplishing these research objectives, the study can improve efforts to prevent accidents and improve road safety by offering insightful information.

## 1.4 Structure of the Paper

- **Section 2 :** The research findings that have already been published on the topic of predicting traffic accidents are thoroughly summarised in this section.
- **Section 3 :** This section contains the methodological strategy used in this project.
- **Section 4 :** This section discuss about the design specification.
- **Section 5 :** Thorough explanation of the implementation of various machine learning and deep learning models are discussed in this section.
- **Section 6 :** Detailed discussion of findings from all the experiments are mentioned in this section.
- **Section 7:** Conclusion and recommendations for further research are mentioned here.

# 2 Related Work

This section offers a thorough summary of the research studies that have already been published on the topic of predicting traffic accidents using different machine learning and deep learning approaches. The accumulated knowledge from related work will help to develop a better predictive model for vehicle collision forecasting in Ireland.

## 2.1 Predictive Analysis for Traffic Accidents Using Machine Learning Approaches

The study by Dia et al. (2022) concentrates on crashes on roads in Senegal and seeks to create several models that predict traffic accidents using various machine learning algorithms. Models including Random Forest, KNN, SVM, Logistic Regression are employed in this study. These models have been developed based on driver behaviour, vehicle condition, and road environment characteristics. In order to train and evaluate the models, a dataset containing various accident-related attributes, such as road conditions, weather, and time of day, were utilised by the researchers. The importance of this study lies in its ability to enhance road safety in Senegal, given the rapid increase in the number of automobiles and the concerning rise in crashes on the roads. By determining the significant variables that cause the accident, forecasting algorithms can help in the reduction of accidents, injuries, and financial losses. The findings indicated that the RF and SVM methods demonstrated the highest level of performance, obtaining a remarkable R squared score of 0.85 for both the models. The study highlighted the importance of the RF model, which consistently produced the best results across a variety of evaluation metrics. This demonstrates its applicability for predicting accident and suggests that it should be favoured in future studies of traffic accidents in Senegal.

The study by Lee et al. (2020) investigates the use of machine learning techniques to predict the number of traffic accidents during rainy seasons in Seoul. The research paper examines the requirement for precise forecasting of accidents, particularly in severe weather that can have major consequences on road safety. In previous studies, conventional statistical methods were employed, but this study evaluates their efficacy using methods of machine learning such as random forest, artificial neural network, and decision tree. A total of three datasets were employed, including Seoul, Korea's Naebu Motorway road geometry, traffic accident and  precipitation data. It was evaluated using three distinct metrics: out-of-bag estimate of error rate (OOB), MSE and RMSE. With low OOB, MSE, and RMSE values, the RF model exhibited greater performance. The purpose of the study by Biswas et al. (2019) is to predict the number of road accidents and fatalities in Bangladesh. The researchers utilised Random Forest Regression to deal with relationships that are nonlinear and high-dimensional data. To train and evaluate the algorithm, they utilised historical Bangladeshi crash data. Various factors, such as road conditions, weather, and traffic patterns, were taken into account in order to identify the variables that contribute to road crashes. R squared value and RMSE are used to assess the efficacy of the regression model. Comparing the actual outcomes to the predicted results reveals that the RF algorithm works pretty well with the dataset.

The research paper by Malik et al. (2021) concentrates on creating a framework for predicting the crashes using machine learning algorithms. By offering precise and on-time accident predictions, the study seeks to improve rescue response times, increase safety on the roads, and decrease traffic congestion. A UK crash dataset was used for the implementation and assessment of six distinct machine learning methods, including Logistic Regression, Decision Tree, Naive Bayes, RF, Bagging, and AdaBoost. The outcomes demonstrated that the Random Forest algorithm outperformed the other models. Utilising the SMOTE method, the framework was able to successfully address class imbalance. In addition, the research identified important collision patterns from the vast traffic data, giving transportation officials with insights for implementing proactive accident prevention strategies. This study highlights the importance of machine learning techniques in road safety analysis and illustrates the possible application of forecasting algorithms to identify and rank high-risk collision possibilities. The research paper by Yamparala et al. (2022) discusses the problem of road traffic accidents (RTAs), specifically cyclist accidents, which remain a significant issue in India. Using a variety of machine learning algorithms, the study seeks to create a precise forecast system for bicycle accidents. The analysis relies on the Indian dataset STATS2021 for the year 2020, which contains variables pertaining to accident details, roadway conditions, atmospheric conditions, accident severity, fatalities, and automobile data. Predictions are made using Logistic Regression, Random Forest, and Decision Tree. The purpose of this research is to determine the optimal algorithm for predicting bicycle accidents. In addition, the study identifies significant factors that contribute to road accidents. By analysing the dataset and identifying important crash patterns, the research offers helpful insight into the variables that contribute to cyclist accidents. Random Over-Sampling Examples (ROSE) is applied to the unbalanced dataset to avoid biased forecasts and assure precise outcomes. The Random Forest method demonstrates to be the most efficient of all models.

## 2.2 Predictive Analysis for Traffic Accidents Using Deep Learning Approaches

The paper by Rahim and Hassan (2021) discusses the pressing problem of traffic jams and accidents in motorway work areas, highlighting the significance of timely and precise forecasting of crashes to speed up emergency response times and improve traffic safety. Previous research has utilised machine learning and statistical techniques to forecast collisions, but their performance was unsatisfactory. To overcome this problem, the study proposes a novel method for predicting accidents that utilises deep learning, namely a convolutional neural network (CNN). A deep learning framework comprises a feature extractor that automatically draws information from the data using CNN layers and a classifier that classifies the data according to the extracted features. Using a generalised numeric-to-image transformation method, the initial numeric accident data is converted to images in order to facilitate CNN usage. Transfer learning is also used to enhance the performance of the model by employing the knowledge of the previously-trained model. Utilising the numeric to image transformation method, establishing a deep learning-based approach with a customised f1-loss function for forecasting accidents, and contrasting the outcomes with an SVM model are the paper's primary contributions.

Investigating the efficacy of a deep learning method for predicting short-term accident risk in urban areas using multiple datasets from Manhattan, New York City, is the objective of the study by Bao et al. (2019). The researchers gathered numerous datasets, like incident data, GPS data from taxis on a large scale, road network characteristics, demographic information, and information on the weather. They propose a spatiotemporal convolutional long short-term memory network (STCL-Net) to forecast the short-term accident risk in a citywide context. The proposed architecture for spatiotemporal deep learning can efficiently investigate the temporal and spatial relationships in high-dimensional explanatory variables. Incorporating various explanatory variables into an end-to-end deep learning architecture, the model improves the accuracy of collision risk predictions for complicated and nonlinear phenomena. The research by Jiang et al. (2020) examines traffic collision detection and aims to investigate the connections between roadway conditions and accident risk in order to prevent possible collisions and improve road safety. Previous collision identification techniques have constraints, such as the inability to simulate changing traffic conditions prior to crash occurrences. In order to overcome these drawbacks, the study proposes LSTMDTR, a framework based on Long Short-Term Memory (LSTM) that takes into account traffic data at various temporal resolutions for crash detection. The LSTMDTR model consists of 3 LSTM networks, each of which considers traffic information at distinct temporal resolutions in order to fully depict traffic variations at various time intervals. A fully-connected layer is utilised to aggregate the outcomes of these LSTM networks, while a dropout layer is employed to minimise overfitting and enhance the accuracy of predictions.

Using deep learning technique and advanced driver assistance system (ADAS) technology, this study by Formosa et al. (2020) concentrates on forecasting accidents in real time. Historically, machine learning algorithms have been utilised for forecasting traffic conflicts using a singular safety surrogate measure. This approach, however, disregards other significant factors which impact road accidents, including speed variation, traffic density, and weather. Combining and mining heterodox data, as well as managing unbalanced data, represent further difficulties. The paper recommends a centralised digital architecture and a technique based on deep learning for predicting accidents. On a section of the M1 in the United Kingdom, an instrumented vehicle collects highly separated traffic information as well as vehicle sensor data. A Regional-Convolution Neural Network (R-CNN) model is used to identify traffic conflicts. Using the digital architecture, the information gathered is combined with traffic characteristics and estimated safety surrogate measures (SSMs) to construct a series of Deep Neural Network (DNN) algorithms for the prediction purpose. The results indicate that the best DNN model achieves an R squared value of 0.94. This paper by Li et al. (2020) examines the use of LSTM-CNN model to predict collision risk on urban main roads. Previous research efforts have concentrated mainly on crash prediction on motorways, but urban main roads show a more complex environment due to junctions and the requirement to consider signal timing. The objective is to create a model that can explicitly learn from a variety of features, including traffic flow characteristics, signal timing, and meteorological conditions. The LSTM-CNN model incorporates the advantages of LSTM and CNN. LSTM has the ability of recording dependencies that are long-term, whereas CNN can derive features that are time-invariant. By combining LSTM and CNN, the algorithm is capable of learning from both time-series and spatial data. SMOTE is used to deal with the imbalance issue in the dataset, where non-crash occurrences are more prevalent than crash events. Results show that the suggested LSTM-CNN algorithm beats other models.

## 2.3 Review of Papers Utilizing Hyperparameter Optimization in Machine Learning Algorithms

The paper by David (2020) discusses Hyperparameter optimisation techniques for enhancing the efficiency of ML algorithms. Hyperparameter optimisation is a crucial of component of machine learning that entails determining the optimal values for the variables that influence a model's learning process. Hyperparameters possess an important effect on the efficacy of a model. The algorithm can be trained more effectively by adjusting factors like learning rates, the number of estimators, and maximum depths. In order to maximise the efficiency on the data over a fair amount of time, hyperparameter optimisation seeks to determine the best combination of hyperparameter values. It is crucial since using the default values for hyperparameters may not always produce the best results for diverse machine learning applications. A frequent method for hyperparameter optimisation is grid search. To find out the ideal configuration, it entails methodically attempting all potential combinations of hyperparameter values. Grid search is a clear and simple method to comprehend. It offers a precise record of the employed hyperparameters and the associated performance data. This makes tests repeatable, allowing for the validation and expansion of the findings by other researchers.

The study by Aldhari et al. (2023) mainly focused on creating several forecasting algorithms for collisions that are happening in Saudi Arabia, notably in the Qassim Province. For the prediction, algorithms including logistic regression, random forest, and XGBoost were used. The significance of hyperparameter tuning and grid search for attaining optimal model performance is emphasised in this research paper. By determining the optimal parameter values to minimise model complexity, enhance generalisation performance, and prevent overfitting, hyperparameter optimisation plays an essential role in machine learning algorithms. Grid search was utilised to systematically examine the hyperparameter space and identify the optimal hyperparameter value combination for the algorithms. This procedure substantially enhanced the efficiency of the model. The outcomes showed that XGBoost outperformed the other ML models. The optimal hyperparameters for XGBoost were discovered with the aid of grid search, resulting in an R-squared value of 0.84. By applying the random forest algorithm, the research paper by Humera Khanum (2023) concentrates on building a model that predicts for traffic accident on Indian highways. The significance of the hyperparameter optimisation and grid search in optimising the model's performance is emphasised in this study. A multi-step methodology involving the collection of data, feature selection, model training, parameter optimisation, and evaluation was used to develop the predictive model. The hyperparameters of the random forest model included'max depth': 10,'max features':'sqrt', and 'n estimators': 100. Utilising grid search, the optimal hyperparameter values were determined, resulting in enhanced model performance. The findings highlight the significance of hyperparameter tuning and grid search in optimising the efficiency of the random forest model, thereby facilitating in forecasting of road accidents with precision. However, the study also recommends investigating alternative hyperparameter optimisation strategies to resolve computational challenges in high-dimensional parameter spaces.

# 3 Methodology

Figure1 shows the research methodology for this paper.The foundation of any research study is  thorough comprehension of existing knowledge. The literature review phase is comprised of two primary stages: superficial reading and detailed reading. Superficial reading can assist with recognising  subject's broad landscape, including its key concepts, major theories, and significant research works. In contrast, detailed reading delves deeper into selected papers and studies to extract essential insights, methodologies, and knowledge gaps. The next stage, following the thorough understanding of  current state of the field, is to formulate a clear and concise problem statement. This statement functions as the focal point of the research, outlining the issue which we intend to investigate. A well-formulated problem statement limits scope of the study and serves as a road map for the subsequent phases of research. Developing a thorough research proposal is essential for outlining the required methodology, resources, and timeline to address the identified issue. This phase entails elaborating on the research objectives, describing the significance of the study, delineating the research design, choosing appropriate methodologies, and describing the anticipated results. The initial experiment phase entails undertaking preliminary studies or experiments to collect the crucial data and insights pertaining to the identified issue. The design and implementation phase entails developing a structured model or strategy for addressing research problem, based on the insights obtained from the initial experiment. In this phase, model developed in the preceding step is rigorously tested and validated. This requires conducting experiments. The investigation process culminates in the creation of a comprehensive thesis report. This report summarises entire research process, from the initial literature review to the conclusion.

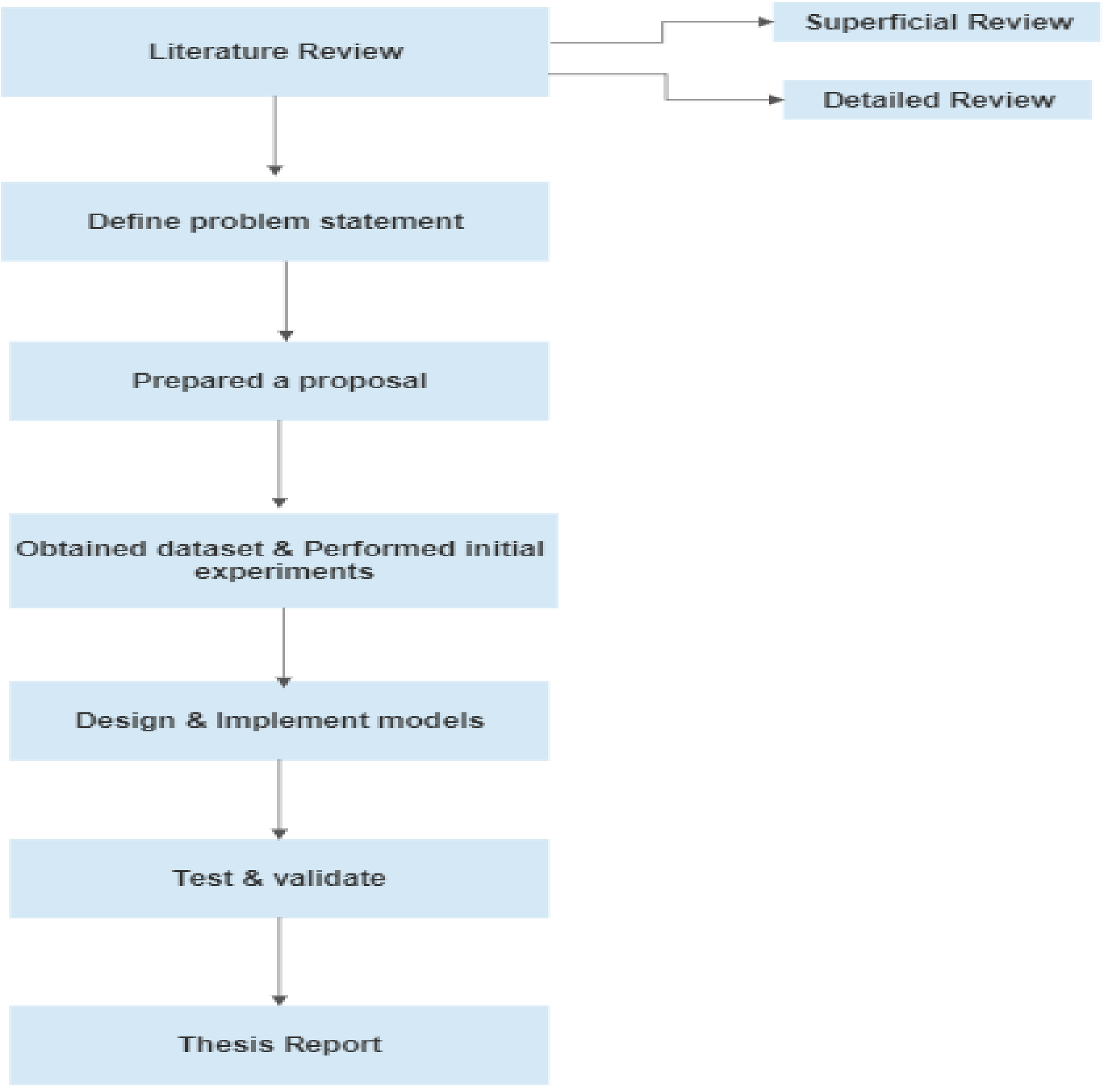


Figure 1: Research Methodology

## 3.1 Employing CRISP-DM for Structured Methodological Implementation

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a popular framework for data mining, ML and data science projects. It offers an organised and methodical approach for challenging problems. Utilising CRISP-DM as the framework of choice, the execution of machine learning and deep learning algorithms for crash forecast in Ireland is carried out successfully. The research paper by Purbasari et al. (2021) highlights the benefits of the CRISP-DM method. CRISP-DM is a framework that is adaptable to a variety of industries and domains, making it versatile and extensively applicable. The flexibility it offers enables data scientists and teams to modify the method according to particular project specifications, data features, and business objectives. By incorporating domain experts, data scientists, and business analysts all over the process, this framework creates a deeper comprehension of the problem domain and makes sure the resulting models reflect real-world requirements. The six phases of this framework are Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. Detailed descriptions of each stage are provided below.

### 3.1.1 Business Understanding

The first phase of CRISP-DM, which is the business understanding, entails gaining a comprehension of the goals of the project, identifying the business problem, and establishing the success criteria. This phase is crucial as it establishes the framework for the remainder of the work. A comprehensive Business Understanding phase can aid in identifying and mitigating potential threats prior to the occurrence of issues. Additionally, it assists in discovering the data sources pertinent to the issue, thereby saving time and resources in the future.

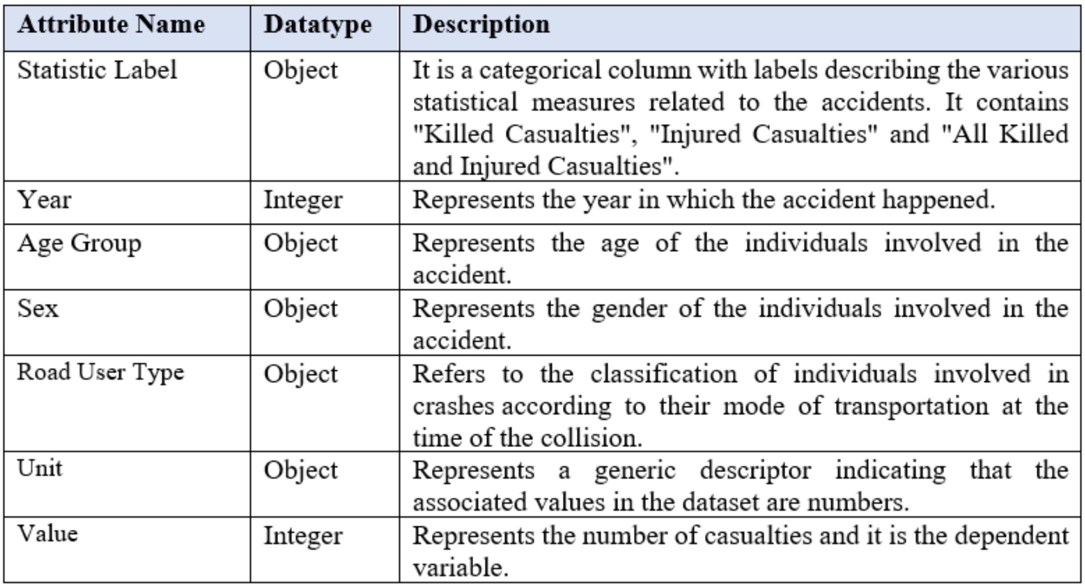
The major goal of this research project is to create precise and efficient predictive algorithms that can predict Irish traffic incidents. The models will examine data on Irish traffic accidents using deep learning and machine learning technologies. The ultimate objective is to deliver practical knowledge to various stakeholders, including government bodies, transportation departments, in order to improve traffic control techniques, lower the rates of crashes, and to increase road safety. The Business Understanding phase also includes an evaluation of the predictive models' prospective value and impact. Some of the benefits includes:

- By correctly performing the prediction of incidents, the RSA can provide methods to reduce the number of fatal crashes on Irish roads, which could potentially save lives.
- By reducing the frequency of severe accidents, the RSA can reduce the associated economic costs, such as rescue efforts and hospitalisation.
- Using machine learning and deep learning techniques, it is possible to identify trends in a vast quantity of accident data that may not be apparent to humans. This may allow for more accurate disaster predictions in the future.

### 3.1.2 Data Understanding

This phase attempts to familiarise the analyst with the data and ensure its suitability for the specified research. This can be accomplished by examining the data, identifying any issues, and making the necessary modifications to make sure the quality of the data and its dependability.

The dataset used to predict Irish road accidents in this study is obtained from the website of the Central Statistics Office (CSO)[1]. CSO is Ireland's national statistical agency, and its mission is to gather, analyse, and disseminate statistics about the Irish population, society, and economy.There are 9984 rows and 7 columns in the dataset. It contains information on accidents that happened in Ireland from the year 2005 to 2020. The description of each variable is given in the below table.



| Attribute Name | Datatype | Description |
|---|---|---|
| Statistic Label | Object | It is a categorical column with labels describing the various statistical measures related to the accidents. It contains "Killed Casualties", "Injured Casualties" and "All Killed and Injured Casualties". |
| Year | Integer | Represents the year in which the accident happened. |
| Age Group | Object | Represents the age of the individuals involved in the accident. |
| Sex | Object | Represents the gender of the individuals involved in the accident. |
| Road User Type | Object | Refers to the classification of individuals involved in crashes according to their mode of transportation at the time of the collision. |
| Unit | Object | Represents a generic descriptor indicating that the associated values in the dataset are numbers. |
| Value | Integer | Represents the number of casualties and it is the dependent variable. |

Figure 2: Description of variables

Now let's visually explore the data.

Figure 3 is a bar graph depicting the distribution of individuals involved in road accidents by age group. The chart provides insightful information regarding the prevalence of road accidents among various age groups. As indicated by the chart's tallest bar, age group of 25 to 34 years old had highest rate of involvement in the road accidents. Also, children are the least likely to be implicated in accidents.

Using a pie chart, Figure 4 depicts the gender distribution among those involved in accidents. The graph depicts the proportion of males and females among the total number of accident victims. Males account for 58.1% of the total number of individuals involved in accidents. This finding suggests that the males are more likely than the females to be involved in incidents, based on the available data.

[1] https://data.cso.ie/table/ROA16

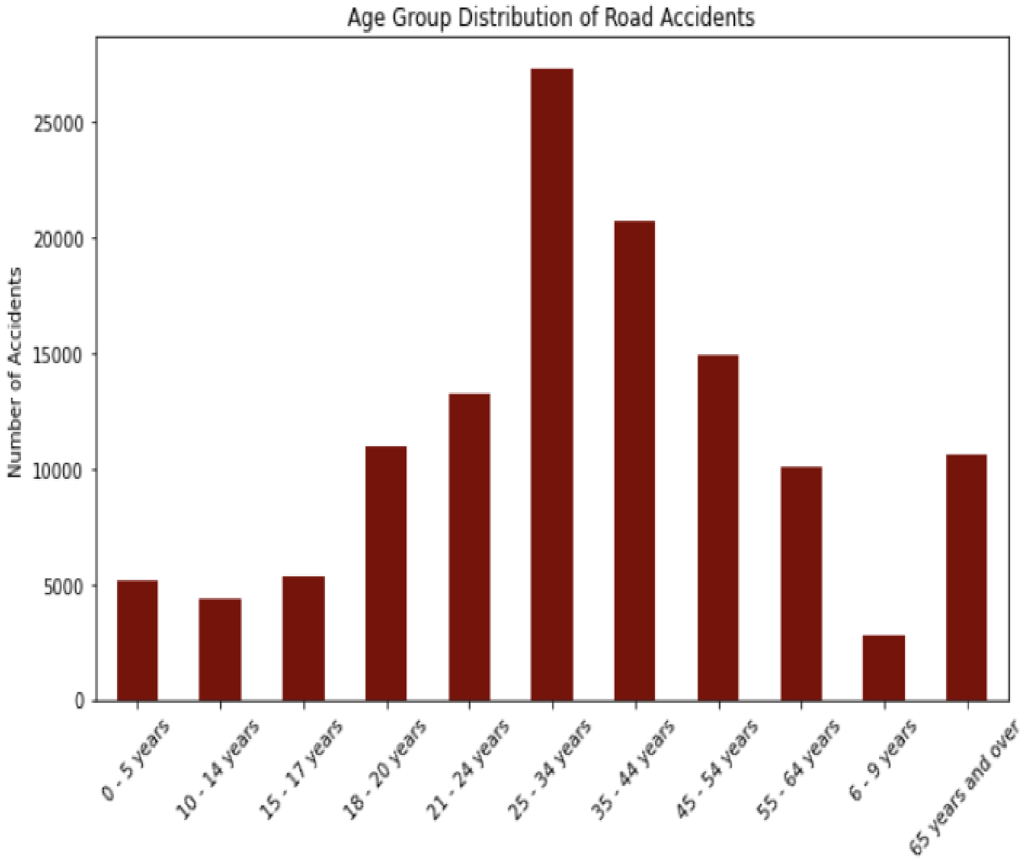


Figure 3: Age Group Distribution of People Involved in Road Accidents

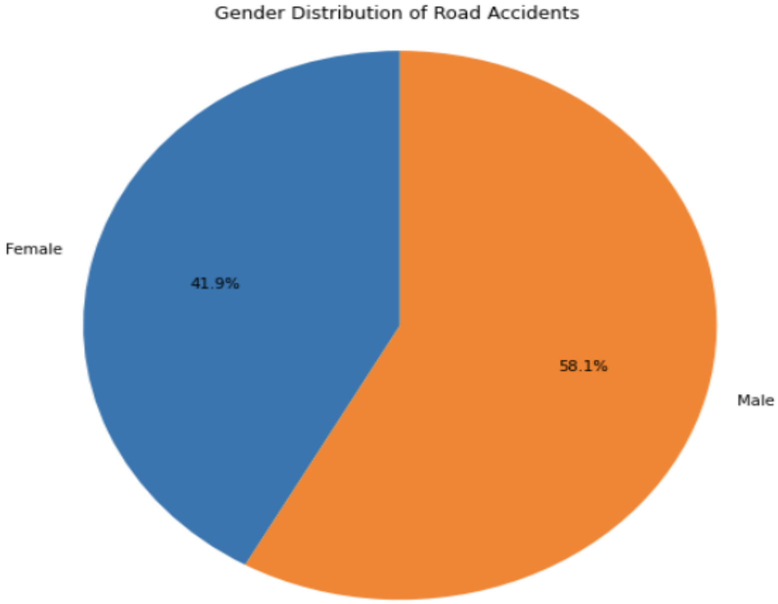


Figure 4: Gender Distribution of People Involved in Road Accidents

Figure 5 depicts the distribution of road Accidents by type and gender of road user. It

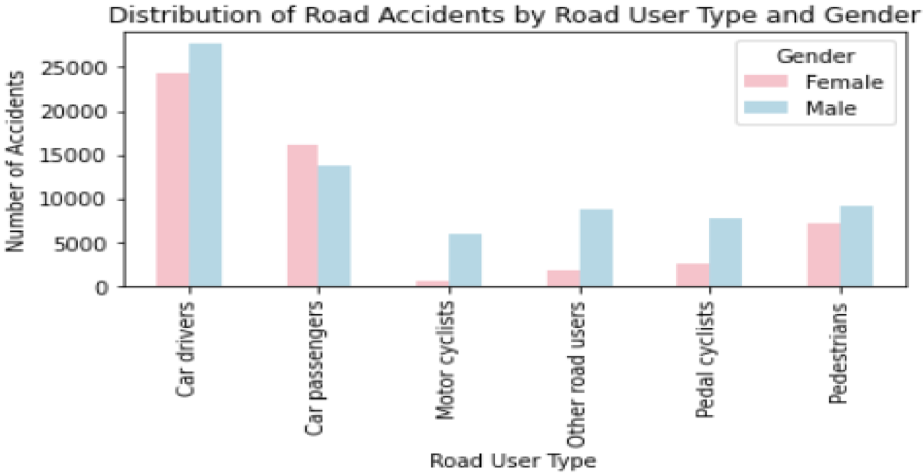


Figure 5: Distribution of Road Accidents by Road User Type and Gender

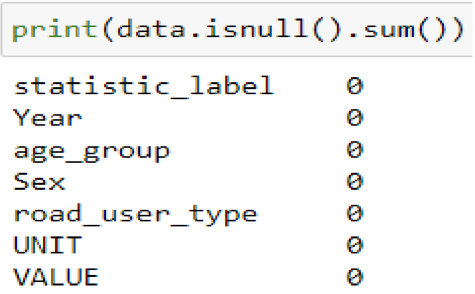
is evident from the graph that car drivers are the largest group implicated in accidents. In addition, male car drivers appear to be the most susceptible to accidents, as they account for a significant proportion of the total number of victims. In addition, the graph reveals that car passengers are the next most likely group to be involved in accidents, following drivers.

### 3.1.3 Data Preparation

In order to prepare the data for subsequent processing, the data preparation phase encompasses all activities necessary to construct the final dataset from the initial unprocessed data.

As part of data preparation, a few steps have been carried out in this project.

- Check for Null values: In the data preprocessing stage, handling missing values is crucial since disregarding them can result in the models producing biassed or erroneous results. With the help of the isnull function, the presence of null values was checked, and no null values were present in the dataset as shown in figure 14.



```
print(data.isnull().sum())
```

```
statistic_label    0
Year               0
age_group          0
Sex                0
road_user_type     0
UNIT               0
VALUE              0
```

Figure 6: No NA values

- Check for duplicated rows: Duplicate rows can compromise the dataset's integrity, reducing its suitability for analysis and modelling. It may also result in overfitting. Thus, the duplicated function was used to determine whether any rows were duplicated in the dataset. There were no duplicate rows in the dataset.

- Data Reduction: Eliminating rows that include summary or aggregate data is the goal of this process. The dataset becomes more focused and condensed by eliminating these aggregated rows, offering a better analytical basis.

  The statistical label column in the Irish road accident dataset has three values: "All Killed and Injured Casualties", "Killed Casualties," and "Injured Casualties." The total number of killed and Injured Casualties is represented by the value "All Killed and Injured Casualties". This aggregated value is omitted from the dataset because it doesn't specifically describe any killed or injured casualties. By deleting this row, the research can concentrate on the distinct counts of fatalities and injuries, offering more detailed information and enabling more focused study. Similarly "All ages" value from the column age group is also removed as it is the aggregated data for all age groups.The Road User type column is another field that has aggregated data in it. The value " All road users" is an accumulation of data that includes counts of different road user categories, such as pedestrians, cyclists, motorcyclists, all car users, and other road users. "All road users" is also removed as a result. The value " All Car Users" is the sum of "Car drivers" and " Car Passengers". Hence that value is also omitted from the dataset.

- Data Transformation: Age group column in the dataset has the value "Age unknown". Hence it has been replaced with the mode value and it can be considered as a form of data imputation, that substitutes a statistically valid estimate based on the data already present. By replacing the age unknown values with the age group that makes up the majority of the dataset, the dataset is made more complete, and the analysis can be carried out with more accuracy.

### 3.1.4 Modeling

Modelling is a crucial phase in which machine learning and deep learning models are created, trained, and evaluated using the prepared data. This phase is intended to convert the preprocessed data into actionable knowledge, allowing for predictions to be made to answer the research questions.

In this project various machine learning and deep learning algorithms are used to predict Irish road accidents.

- Machine Learning Algorithms

Five machine algorithms are applied for the prediction of the road accidents in Ireland. A comprehensive review of relevant literature on accident prediction reveal that the machine learning models such as Random Forest, Decision Tree, XGBoost, Ridge Regression, and KNN are increasingly adopted by researchers. Across numerous datasets, these models have consistently demonstrated superior performance. The choice of these algorithms for research of Irish road accident data is based on the fact that they have been used successfully in other studies. The main goal is to figure out how well they work and how effective they are in the context of this unique dataset.

- Random Forest is known for being able to handle data that is noisy or inconsistent. In real-world datasets like road crashes, there may be times when information is missing or wrong. The ensemble method of Random Forest, which combines the predictions of several trees, can help lessen the effect of such noisy data points.
- Decision trees are particularly suitable for capturing non-linear relationships among variables. Despite the conversion of categorical data into numerical representations, it is possible that complex connections persist among these transformed features, which can have an impact on number of accidents. Decision trees have the ability to accurately detect and depict non-linear patterns, hence offering valuable insights that may not be immediately evident using basic linear modelling methods.
- The utilisation of K-Nearest Neighbours (KNN) algorithm has proven to be a viable approach for detecting anomalies in the domain of road accident prediction. The K-nearest neighbours (KNN) algorithm can be utilised to detect cases that exhibit substantial deviations from the average in terms of feature values. This capability enables the identification of uncommon or infrequent circumstances that may want specific attention or action. The capacity to identify and emphasise exceptional data points and irregularities holds significant potential in augmenting the precision of accident forecasting models and raising the overall state of road safety.
- Ridge Regression is capable of addressing potential multicollinearity concerns with the incorporation of a regularisation term into the cost function, taking into account factors such as year and age. This phenomenon improves the robustness of coefficient estimations, hence enhancing the accuracy and dependability of predictive outcomes.
- The XGBoost algorithm integrates regularisation approaches that are highly effective in mitigating overfitting, particularly in the context of datasets that are of moderate size. XGBoost promotes the balance between model complexity and generalisation by incorporating penalties for too complicated models, hence enhancing the stability and dependability of predictions.

A brief description of each algorithm is mentioned below.

### A) Random Forest with and without hyperparameter optimization

The Random Forest is one of the most efficient forecasting machine learning algorithms. How a random forest algorithm works is mentioned in detail in this study by Mbaabu (2020). A random forest method is made up of numerous decision trees. Bagging or bootstrap aggregation is used to train the 'forest' produced by the RF algorithm. Bagging is a meta-algorithm for ensembles that enhances the precision of machine learning models. This model determines the result according to the decision tree predictions. It predicts by averaging or calculating the mean of the results obtained from multiple trees. Increasing the total number of trees improves the accuracy of the results. It eliminates the constraints of a decision tree. It decreases dataset overfitting and improves precision.

Several hyperparameters govern the behaviour of Random Forest Regression, including the number of trees (n_estimators), the highest depth of each tree (max_depth), and the number of attributes evaluated at each split (max_features). The optimisation of hyperparameters involves determining suitable values for these parameters.

### B) XGBoost with and without hyperparameter optimization

XGBoost (Extreme Gradient Boosting) is a well-known and potent ML model that has been demonstrated to be successful in tasks such as classification and regression. It creates a robust model for prediction by integrating various weak learners, generally decision trees. Hyperparameter tuning can be used to improve the efficacy of XGBoost by searching for the most effective combination of hyperparameter values. XGBoost's hyperparameters include the number of trees (n_estimators), the learning rate (eta), and the maximal depth of trees (max_depth). The optimisation of hyperparameters requires the specification of a search space for those values. XGBoost's key characteristics are mentioned in detail in the blog by Hachcham (2023). Some of them are: XGBoost has various regularisation penalties to prevent the overfitting, the model can generalise effectively with the help of a penalty regularisations, and XGBoost can recognise and understand non-linear data patterns.

### C) Decision Tree with and without hyperparameter optimization

Decision Tree Regression is a method for supervised machine learning utilised in regression analysis. The paper by Prasad (2021) shows how this model works. Decision tree creates a structure resembling a tree by recursively splitting the feature space into sections according to the values of the features. Each leaf node in the tree indicates an estimated value, that is the mean of the desired variable in that subset. This model is very easy to understand and implement. Hyperparameters such as the maximum depth of the tree (max_depth) and the minimum number of samples needed to divide an internal node (min_samples_split) govern the complexity of this model. Optimisation of hyperparameters seeks to identify the optimal values for these parameters.

### D) KNN with and without hyperparameter optimization

K-Nearest Neighbours (KNN) Regression is another most efficient ML models utilised for regression tasks. KNN does not directly learn a model from training data, unlike conventional regression models. Rather, it uses the resemblance of the data points in the feature space to generate forecasts. This model can manage nonlinear connections among features and the target variable, which makes it appropriate for complex data patterns. In addition, it makes no assertions regarding the pattern of distribution of the underlying information, resulting in a flexible and versatile model. Hyperparameter optimisation aids in locating the optimal K value, resulting in enhanced predictive accuracy and reduced risk of underfitting or overfitting. The advantages of KNN model are mentioned in the study by Soni (2020).In terms of improvisation for random modelling on the available data, KNN modelling is incredibly time-efficient. This is due to the fact that algorithm doesn't need a training period, as data itself serves as a model for future predictions. More information can be added at any moment because there is no training period and model will not be affected.

### E) Ridge Regression with and without hyperparameter optimization

Ridge Regression, also known as L2 regularisation, is a method for modelling and forecasting continuous numerical values that is based on linear regression. The regularisation term is added to the cost function in order to improve on the conventional least squares method. The regularisation term is determined by the total of the squared magnitudes of the model's coefficients multiplied by a hyperparameter known as ”alpha” . The main goal of this model is to discover the optimum coefficients that minimise the sum of the squared variations among the forecasted and real values. The efficacy of Ridge Regression can be enhanced through hyperparameter optimisation in which the optimal value of alpha is identified.

- Deep Learning Algorithms

In this research paper, two deep learning algorithms are used to predict road accidents in Ireland. Following is a concise description of each algorithm.

### A) Long Short-term Memory(LSTM)

LSTM networks are successful in capturing dependencies and trends in sequential data due to their distinctive architecture, which allows them to store data for extended periods. LSTM is intended for handling consecutive data and has demonstrated exceptional success in a variety of applications, including NLP, time series forecasting, and so on. This model can acquire appropriate characteristics automatically from the provided data, lowering the requirement for manual feature engineering. In addition, each of the steps can be trained simultaneously which leads to shorter training periods compared to traditional RNNs. Pretrained models based on LSTM can be fine-tuned for particular uses, saving computational power and employing the knowledge gained from large datasets. LSTMs store and alter data over time using memory cells. These memory cells enable the framework to keep a memory of pertinent information from previous time steps, thereby facilitating a deeper comprehension of context. LSTMs are less susceptible to overfitting than other deep learning algorithms, particularly when handling large quantities of sequential data.

### B) Feedforward Neural Network(FNN)

FNN, which stands for Feedforward Neural Network, is one of the most basic and straightforward deep learning algorithms. It is the building element for more complex deep learning architectures and is also called as a multi-layer perceptron (MLP). The paper by Kamali (2023) shows how FNN model works. With no loops or recurrent connections, data moves in a single direction from input to output in a Feedforward Neural Network. A number of layers make up the network: an input layer, one or more hidden layers, and an output layer. Each layer is composed of neurons (also referred to as nodes or units) that conduct computations on the input data. There are various advantages of using FNN model. FNNs are very easy to implement and comprehend, which makes them an excellent starting point for learning deep learning principles. It can be scaled up to tackle challenging tasks and huge datasets by incorporating more hidden layers and neurons.

### 3.1.5 Evaluation

CRISP-DM's evaluation phase assesses the model's effectiveness in achieving the the objectives of the project. It involves verifying the model's forecasts, analysing the results, and assessing the model's accuracy with suitable evaluation metrics. There are four evaluation metrics used in this study to assess the performance of the models and a brief description of each of them is given below.

- Mean Squared Error (MSE): It is the average squared deviation among forecasted and actual values. By aggregating the squared errors, the MSE determines how well a regression model fits the data. The smaller the MSE, the closer the model's predictions are to the actual values.
- Mean Absolute Error(MAE): The average absolute variance between predicted and observed values is the MAE. It indicates the extent of the model's predictive errors.
- R-squared score: R-squared measures how much of the variation in the objective variable can be explained by the model. If R-squared is near to 1, it indicates that model is a good fit for the data.
- Root Mean Squared Error(RMSE):The root-mean-square error (RMSE) assesses the average magnitude of differences between predicted and observed values. It assesses how closely the predicted and actual values correspond. A lower RMSE value indicates that model's predictions are closer to actual values, indicating that model is more accurate.

### 3.1.6 Deployment

In the Deployment part of the CRISP-DM method, the trained and tested model is ready to be put to work and can be used to predict crashes in real time.

# 4 Design Specification

This project is carried out on a computer equipped with 64-bit processor, running on Windows operating system and have 16 GB of RAM. The models are developed using Python language. Jupyter notebook IDE is used to execute Python. The Central Statistics Office (CSO) website was utilised to collect the dataset to predict Irish traffic accidents for this study. The dataset is in csv format. Five machine learning models and two deep learning machines have been constructed for predicting road accidents in Ireland. The performance of the models are evaluated using the metrics: RMSE, MSE, MAE and R-squared score.

# 5 Implementation

The machine learning models applied in this project are: Random Forest, XGBoost, Ridge Regression, KNN and Decision tree. The 2 deep learning models employed are LSTM and FNN. The hyperparameter optimisation method employed in this study is the grid search since it is straightforward, inclusive and very simple to apply. Grid Search is simple and systematic method that attempts every possible combination of hyperparameters within the given ranges in order to explore entire hyperparameter space. This thorough search guarantees that all the possible hyperparameter values are examined, offering thorough assessment of model performance. Grid Search produces findings that are easily replicated and free of randomness because it assesses all the hyperparameter combinations. Grid Search is also user-friendly and easy for beginners.

## 5.1 Machine Learning Algorithms

### 5.1.1 Random Forest with and without hyperparameter optimization

Using random forest algorithm 2 models have been developed. In the first model, default hyperparameters are used to construct the Random Forest regression model. The number of trees in the forest ('n_estimators') is set to 100, maximum depth of the trees ('max_depth') is set to None (unlimited), and minimal number of samples needed to split the internal node ('min_samples_split') is set to 2. On the basis of evaluation metrics such as MSE, RMSE, MAE and R-squared score, the performance of the model is measured.

Similar to the first model, the second model begins with a data preparation, one-hot encoding of the categorical variables, and data partitioning. However, it also incorporates optimisation of hyperparameters. Grid Search conducts an exhaustive search across a grid of the predefined hyperparameter combinations. It iterates through all the possible combinations of the hyperparameters provided in the grid, constituting a systematic and a deterministic method. In contrast, Random Search selects hyperparameter combinations from a given distribution or range for a specified number of iterations. It focuses mainly on randomly exploring various regions of the hyperparameter space as opposed to covering every possible combination. Hence Grid search is selected for this project. n_estimators, max_depth, and min_samples_split are the hyperparameters that needs to be optimised. The Grid Search is carried out using GridSearchCV class from the'sklearn.model_selection' package. After determining the optimal hyperparameter combination, grid_search.best_estimator_ is used to derive the optimal Random Forest model. This optimal model is then applied to the test set to generate predictions.

### 5.1.2 XGBoost with and without hyperparameter optimization

The XGBRegressor class of the 'xgboost' library is used to generate the XGBoost regression model with its default hyperparameters. The predictive model is then trained with the fit method on training data. The trained XGBoost model is applied to the test set to predict the target variable. Several evaluation metrics were then employed to assess the model's performance.

The second method entails optimising the hyperparameters. A dictionary named param_grid containing various hyperparameter values to be used during the grid search is defined. It features different learning rates, maximum tree depths, and number of estimators (trees). A regressor is constructed with the default hyperparameters for the XGBoost. The GridSearchCV from scikit-learn is utilised to carry out grid search with 5-fold cross-validation. It iterates through all combinations of hyperparameters specified in the param_grid and determines which combination provides the greatest performance on the training data. After grid search is complete, optimal hyperparameters for the best model (XGBoost regressor) are obtained. On the test set, the optimal model is employed to predict the dependent variable.

### 5.1.3 Decision Tree with and without hyperparameter optimization

Using DecisionTreeRegressor from the scikit-learn and a random_state=42 for reproducibility, a decision tree regressor object (decision_tree) is created in the first model. The fit method is used to train the decision tree regressor on training data. On the test set, predictions are made using the trained decision tree model.

Using default hyperparameters, a decision tree regressor object (dt_regressor) is constructed in the second model. A dictionary named param_grid containing various hyperparameter values to try over grid search is defined. It contains basically two hyperparameters:'max_depth', which defines maximum depth of tree, and 'min_samples_split', which describes minimal number of samples needed for splitting an internal node. Following the completion of grid search, an ideal model optimal hyperparameters is determined. The optimal model is saved in the best_dt_model variable, which contains the decision tree regressor with best-performing hyperparameters during grid search. The optimal model is then employed to forecast the dependent variable.

### 5.1.4 KNN with and without hyperparameter optimization

KNeighborsRegressor from the scikit-learn is used to generate K-nearest neighbours (KNN) regression model. The n_neighbors parameter has been set to 5, indicating that the algorithm should take into account 5 neighbours when making predictions. n_neighbors is a hyperparameter which specifies the number of adjacent neighbours to take into account when predicting a new data point. It is an essential parameter of the KNN algorithm, as it directly affects the behaviour and efficacy of the model. The KNN model is trained on the training data, and the trained model is then utilised to make predictions on test set.

For the second model, KNeighborsRegressor class was constructed with no hyperparameters being specified. By default, it takes number of neighbors (k) as 5 and weight function as 'uniform'. Next is defining hyperparameter grid for the grid search. Two hyperparameters: n_neighbors and weights are used to define param_grid . n_neighbors indicates the number of neighbouring data points to take into account when making forecasts for a data point. Here n_neighbors hyperparameter is set to a list [3, 5, 7], indicating that the grid search will attempt these three different k values during model training and the assessment. The algorithm will select the k-nearest neighbours for each value of k in order to make predictions. The weights parameter defines the weight function used for predicting the objective value for data point. The optimal set of n_neighbors and weights that results in the most precise KNN regression model is discovered by combining these hyperparameters with cross-validation using grid search.

### 5.1.5 Ridge Regression with and without hyperparameter optimization

StandardScaler function is used to scale the features for ridge regression in order to normalise them and give them zero mean and unit variance. By doing so, we can standardise scale of all the attributes, thus guaranteeing their magnitudes are comparable. PolynomialFeatures generates polynomial features of degree 2 to capture potential nonlinear connections among features. Using the Ridge function, Ridge Regression with a preset alpha value of 0.5 is performed. Using negative mean squared error (neg_mean_squared_error) as the scoring metric, cross-validation with 5 folds are applied to the training data to assess the effectiveness of the model. By taking the negative of the MSE, scikit-learn consider higher values of the neg_mean_squared_error as the better scores. So scikit-learn will select the model with the highest neg_mean_squared_error during model evaluation or cross-validation. The algorithm is trained on training data containing polynomial features.

The second model is ridge Regression with optimisation of hyperparameters. This model attempts various combinations of hyperparameters, such as different degrees of polynomial features and different alpha values, to find optimal combination that yields smallest mean squared error.

## 5.2 Deep Learning Algorithms

### 5.2.1 Long Short-term Memory(LSTM)

Using MinMaxScaler from scikit-learn (scaler), the input features are scaled to a range of 0 to 1 following the separation of data into training and test sets. Scaling parameters to a specific range can improve the neural network's training performance. LSTM models require (samples, timesteps, features) as 3D input data. The data is reshaped so that each sample has one timestep and a number of features. Keras and TensorFlow are used to develop the LSTM model. It includes an LSTM layer with 64 units and a Dense layer with one unit for regression. In order to replicate the 3D shape of the input data, the input_shape parameter of the LSTM layer is set to (1, X_train_scaled.shape[1]). The model is compiled with model.compile using mean_squared_error as the loss function and the Adam optimizer. The LSTM model is trained with 50 epochs and 32 batches of training data. On the test set, the trained LSTM model is used to make predictions.

### 5.2.2 Feedforward Neural Network(FNN)

The FNN is constructed using Keras with the Sequential API. The network architecture is comprised of three dense layers (layers that are interconnected). The model.add(Dense(64, activation='relu', input_shape=(X_train_scaled.shape[1],) function inserts the first dense layer with 64 units and the ReLU activation function. In the first layer, the input_shape parameter is used to inform the model of the input shape. model.add(Dense(32),'relu') adds the second dense layer with 32 units and ReLU activation function. model.add(Dense(1)) applies the final dense output layer for regression tasks. Due to the fact that this is a regression task, no activation function has been defined in this layer. The algorithm is then compiled with the model.compile using mean_squared_error as the loss function and the Adam optimizer. Mean squared error is a common loss function employed in regression assignments to determine the variation between predicted and actual target values. The FNN model is trained for 50 epochs and 32 batches on the training data . On the test set, the trained FNN algorithm is employed to make the predictions.

# 6 Evaluation

For evaluating the models developed for the prediction of road accidents in Ireland, 4 evaluation metrices are used and those are: RMSE, MSE, MAE and R-squared score. Several experiments are carried out and the findings are described in detail in this section.

## 6.1 Machine Learning Algorithms

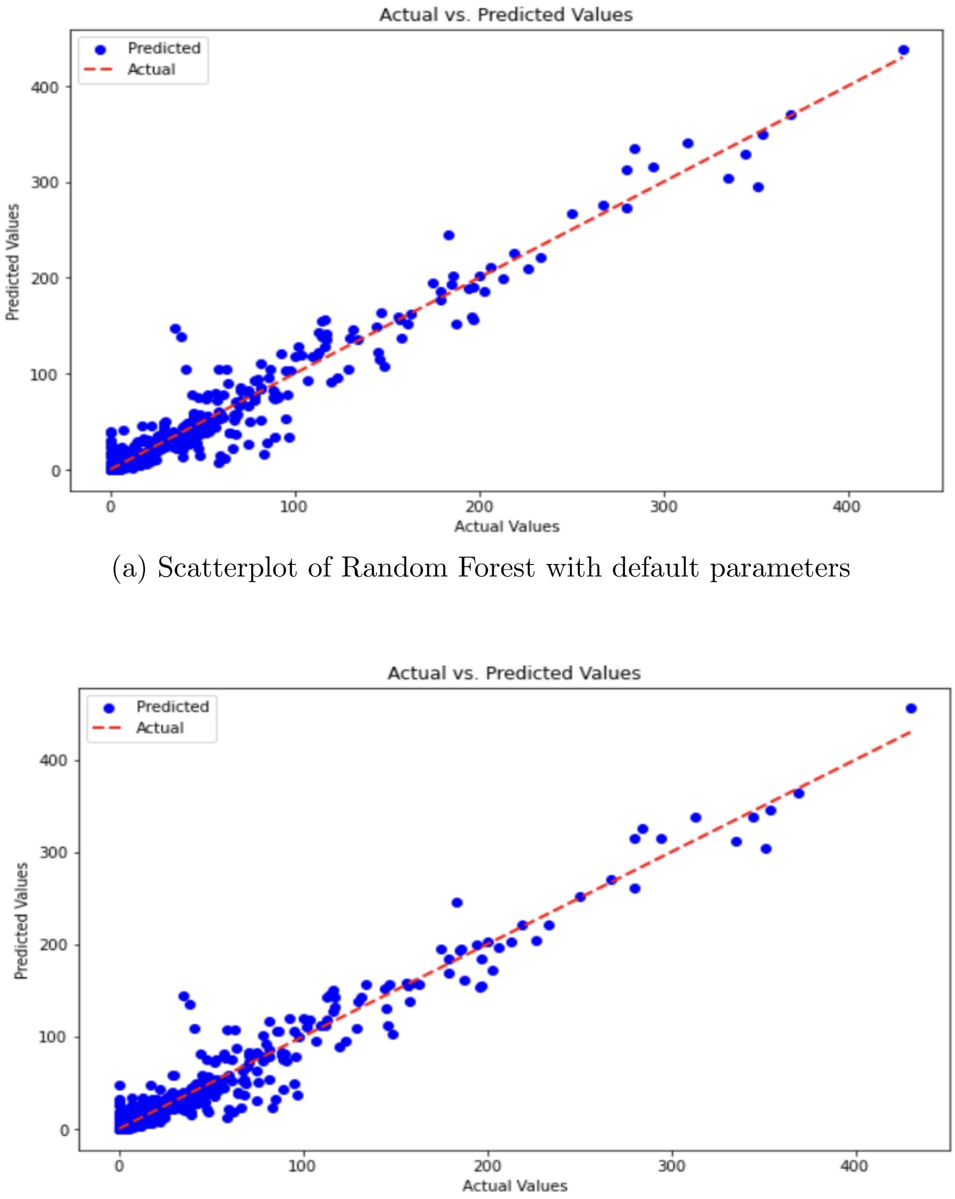
### 6.1.1 Experiment 1 - Random Forest with and without hyperparameter optimization

Let's compare the effectiveness of the Random Forest model with and without hyperparameter optimisation for forecasting the crashes on roads in Ireland. The table 1 shows the comparison of both the models. The MSE of RF model without hyperparameter optimisation is 141.35, but it decreased to 136.8 after hyperparameter optimisation. The decrease in MSE shows that hyperparameter tuning enhanced the predictive accuracy of algorithm, as the values that were anticipated are now closer to actual values. Likewise, value of MAE dropped from 5.23 to 5.18 after the optimisation of hyperparameters. The decline in MAE provides additional evidence that hyperparameter optimisation is effective for lowering the average magnitude of errors. The RMSE values are 11.88 without hyperparameter optimisation and 11.69 with optimisation. The reduction in RMSE indicates that hyperparameter tuning contributed to better-fitting predictions and decreased the magnitude of total error. After optimising the hyperparameters, R-squared score increased from 0.94 to 0.95. The rise in R-squared shows that the hyperparameter tuning improved the model's ability to account for the variance in the target variable, resulting in a better fit.

Table 1: Comparison of Random Forest and Random Forest with hyperparameter optimization

| Models | Random Forest | Random Forest with hyperparameter optimization |
|---|---|---|
| MSE | 141.35 | 136.82 |
| MAE | 5.23 | 5.18 |
| RMSE | 11.88 | 11.69 |
| R-Squared score | 0.94 | 0.95 |

The figure 7 represents the scatterplot of Random Forest with and without hyperparameter optimization. The x-axis corresponds to the actual values ('y_test'), while the y-axis corresponds to the expected values ('y_pred'). The scatter points are illustrated in the colour blue in order to indicate the predicted values.The line of perfect prediction is represented by a red dashed line . The scatter plot exhibits a concentration of blue dots in close proximity to the red dashed line, suggesting a favourable indication of a proficient predictive model.



(a) Scatterplot of Random Forest with default parameters

(b) Scatterplot of Random Forest with hyperparameter optimization

Figure 7: Scatterplot of Random Forest with and without hyperparameter optimization

#### 6.1.2 Experiment 2 - XGBoost with and without hyperparameter optimization

The second experiment compares the XGBoost model with default setting of parameters to the XGBoost model with hyperparameter optimisation. By employing hyperparameter optimisation to XGBoost model, a consistent enhancement in the forecasting accuracy across multiple evaluation metrics is noticed. The tuned model outperformed the default XGBoost model, indicating that fine-tuning procedure improved the model's precision. From the table 2 it is noticeable that, MSEwas significantly reduced by hyperparameter optimisation from 101.79 to 99.96. However the MAE value rose marginally from 5.35 to 5.79 after the optimisation of hyperparameters. Despite a slight increase in MAE, it is important to note that MAE emphasises the absolute variation between predicted and actual values. Despite the increase, the MAE values remain comparatively low, demonstrating the model's predictive accuracy. Optimisation of hyperparameters improved the RMSE from 10.08 to 9.99. The R-squared score for both default and tuned XGBoost algorithms stayed steady at 0.96.

Table 2: Comparison of XGBoost and XGBoost with hyperparameter optimization

| Models | XGBoost | XGBoost with hyperparameter optimization |
|---|---|---|
| MSE | 101.79 | 99.96 |
| MAE | 5.35 | 5.79 |
| RMSE | 10.08 | 9.99 |
| R-Squared score | 0.96 | 0.96 |

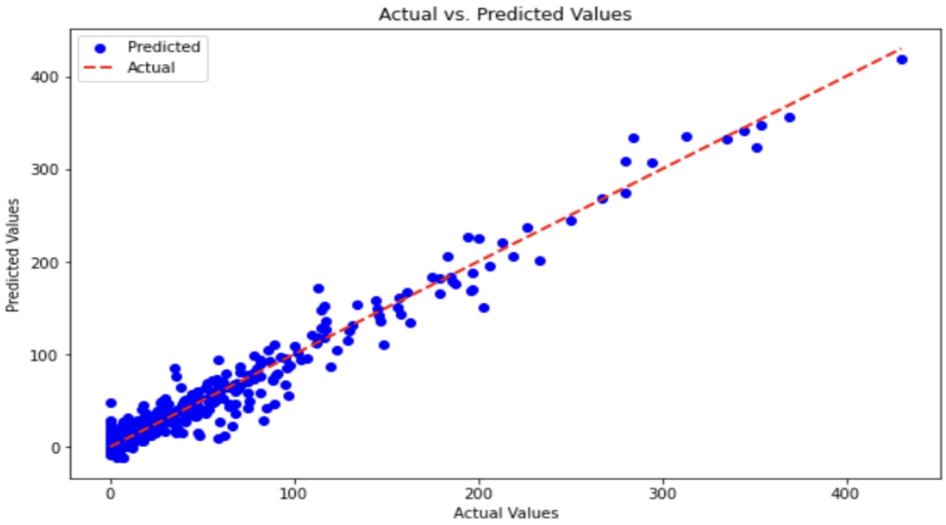


Figure 8: Scatterplot of XGBoost model with hyperparameter optimization

The figure 8 represents the scatterplot of XGBoost model with hyperparameter optimization. Most of the data points are closer to the straight line. The graph suggests that model's predictions exhibit a high degree of reliability and have effective generalisation capabilities when applied to novel, previously unknown data.

### 6.1.3 Experiment 3 - Decision Tree with and without hyperparameter optimization

The implementation of hyperparameter optimisation to the Decision Tree model resulted in notable enhancements to the model's performance. As shown in table 3, there is a reduction in MSE, which implies that tuned Decision Tree model offers more accurate predictions. The MAE dropped from 6.34 to 5.85. Optimisation of hyperparameters reduced the RMSE from 15.30 to 14.36. R-squared indicates that both models account for a comparable amount of variance in the objective variable.

Table 3: Comparison of Decision Tree and Decision Tree with hyperparameter optimization

| Models | Decision Tree | Decision Tree with hyperparameter optimization |
|---|---|---|
| MSE | 234.21 | 206.45 |
| MAE | 6.34 | 5.85 |
| RMSE | 15.30 | 14.36 |
| R-Squared score | 0.91 | 0.92 |

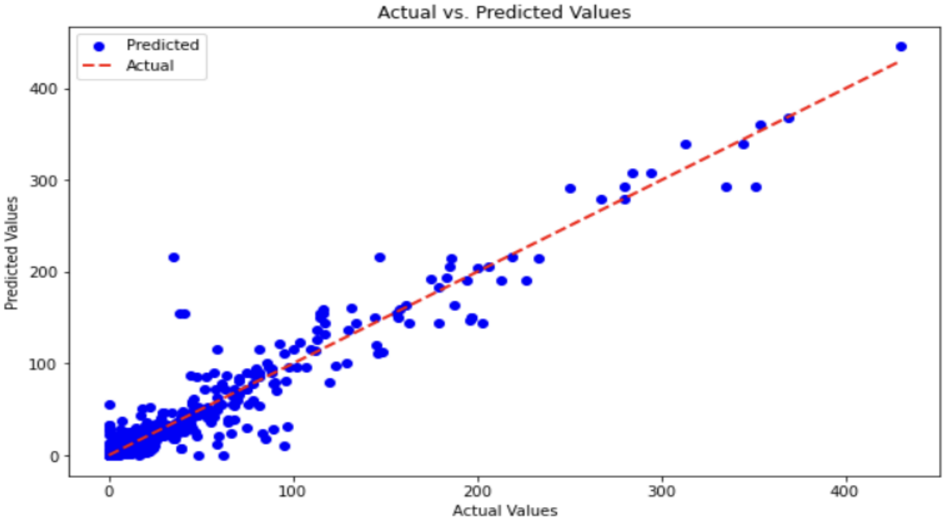


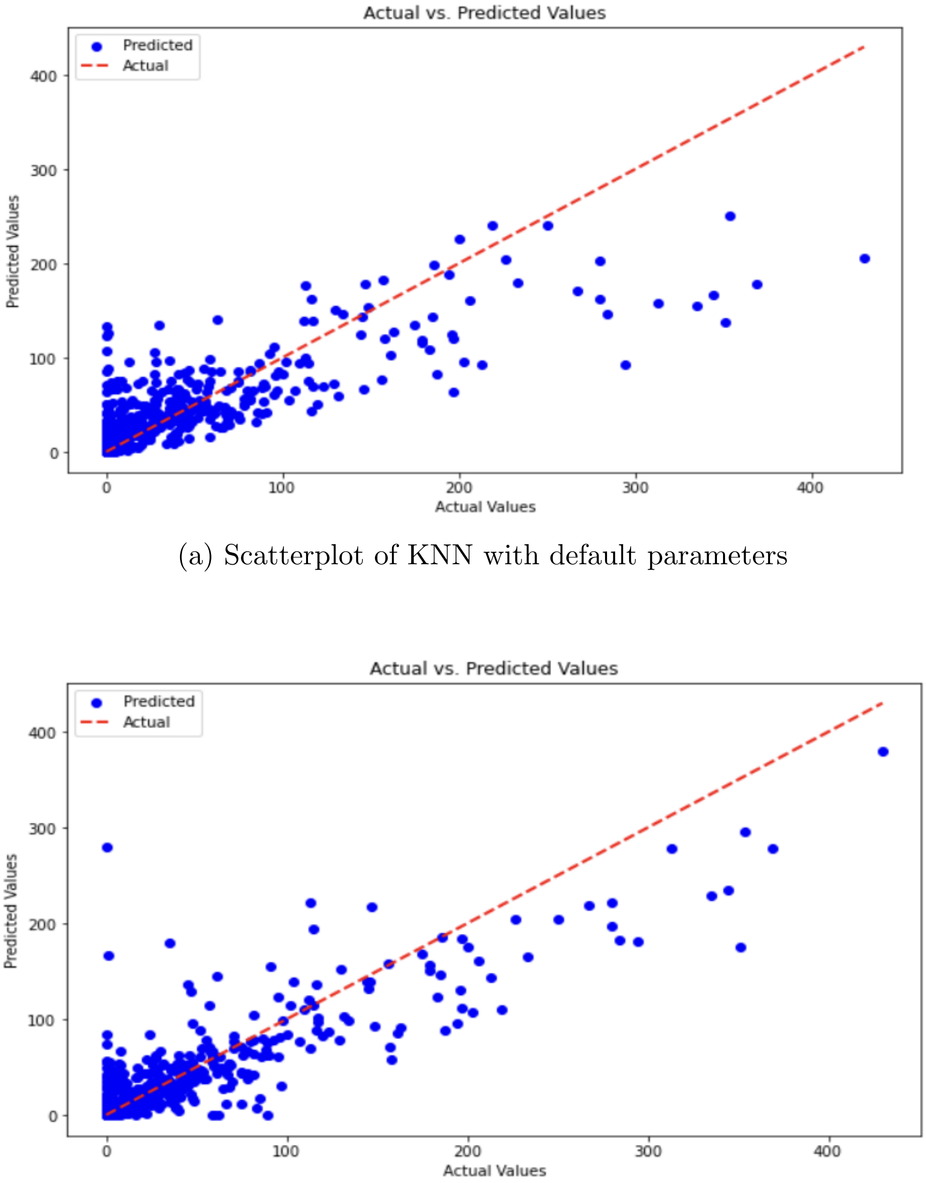
Figure 9: Scatterplot of Decision Tree model with hyperparameter optimization

### 6.1.4 Experiment 4 - KNN with and without hyperparameter optimization

MSE was substantially reduced by hyperparameter optimisation, from 844.13 to 627.24. The reduction in MSE shows that tuned KNN algorithm generates better predictions, which leads to fewer squared variances between the forecasted and actual values. After hyperparameter optimisation, MAE declined from 13.45 to 10.98. Optimisation of hyperparameters reduced RMSE value from 29.05 to 25.04. After hyperparameter optimisation, the R-squared value increased from 0.69 to 0.77. Therefore, it is evident that the KNN with optimised hyperparameters performed better and the result is shown in table 4.

Table 4: Comparison of KNN and KNN with hyperparameter optimization

| Models | KNN | KNN with hyperparameter optimization |
|---|---|---|
| MSE | 844.13 | 627.24 |
| MAE | 13.45 | 10.98 |
| RMSE | 29.05 | 25.04 |
| R-Squared score | 0.69 | 0.77 |



(a) Scatterplot of KNN with default parameters

(b) Scatterplot of KNN with hyperparameter optimization

Figure 10: Scatterplot of KNN with and without hyperparameter optimization

The figure 10 (a) represents the scatterplot of KNN with default parameters. It is evident that most of the data points are far away from the straight line, which means the model is not performing well. It is also evident from the figure 10(b) that after hyperparameter optimization, the data points are much more closer to the straight line.

### 6.1.5 Experiment 5 - Ridge Regression with and without hyperparameter optimization

With implementation of hyperparameter optimisation, the Ridge Regression model also demonstrated significant improvements. MSE decreased significantly from 646.39 to 152.91. After adjusting the hyperparameters, MAE dropped from 18.03 to 6.36. Optimisation of hyperparameters led to a reduction in the model's Root Mean Squared Error (RMSE) from 25.42 to 12.36. The R squared score also shown a better performance after the tuning. This result is shown in table 5.

Table 5: Comparison of Ridge Regression and Ridge Regression with hyperparameter optimization

| Models | Ridge Regression | Ridge Regression with hyperparameter optimization |
|---|---|---|
| MSE | 646.39 | 152.91 |
| MAE | 18.03 | 6.36 |
| RMSE | 25.42 | 12.36 |
| R-Squared score | 0.76 | 0.94 |

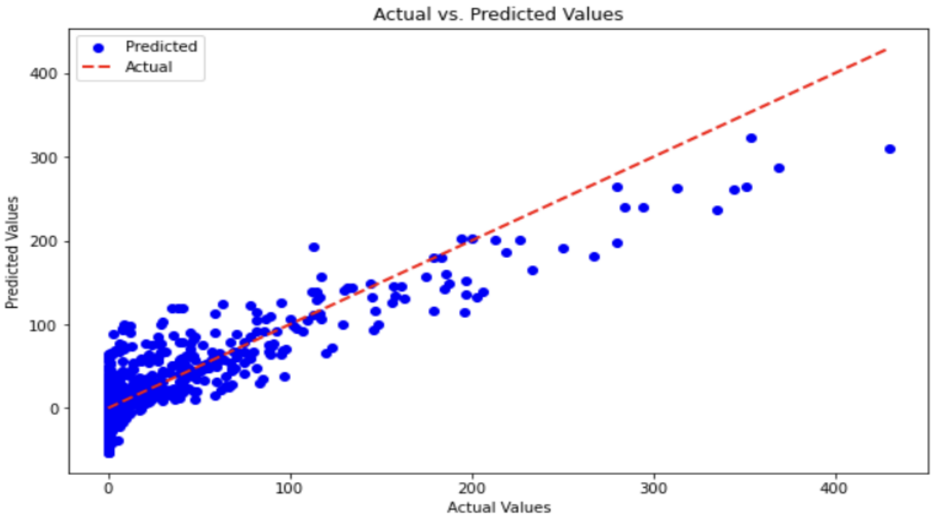


Figure 11: Scatterplot of Ridge Regression model with default parameters

Figure 11 depicts the scatterplot resulting from the implementation of Ridge regression using the default parameters. The scatterplot visually demonstrates significant dispersion of the data points deviating from the theoretical linear relationship. Observed discrepancy indicates that model's performance is not optimal given the original settings. The scatterplot exhibits a noticeable absence of correspondence between the expected and actual values, suggesting that model has a restricted ability to accurately predict outcomes.

On the other hand, Figure 12 displays scatterplot resulting from the implementation of hyperparameter optimisation techniques on the Ridge regression model. There is a noticeable change in the observed pattern, as the data points exhibit a greater degree of convergence towards the expected linear trendline in comparison to the model utilising default parameters.

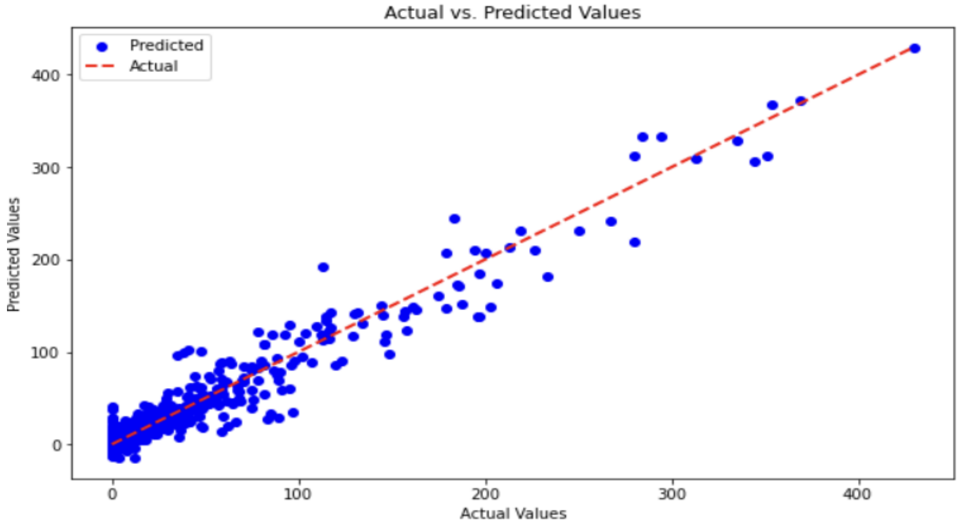


Figure 12: Scatterplot of Ridge Regression model with hyperparameter optimization

## 6.2 Deep Learning Algorithms

### 6.2.1 Experiment 6 - Long Short-term Memory(LSTM)

The performance of deep learning model is shown in table 6. MSE is the average squared deviation between the predicted and actual values of the target variable. MSE value of 542.83 for the LSTM algorithm shows that the squared deviation between the LSTM model's predictions and the actual number of road accidents is, on average, 542.83. A greater MSE value indicates that the model's predictions are less accurate than expected. MAE value of 9.38 shows that LSTM model's predictions differ from actual number of road incidents by an average of 9.38. RMSE value of this model is 23.29 and R-squared score is 0.80 .

Table 6: Performance of LSTM model

| Model | Long Short-term Memory(LSTM) |
|---|---|
| MSE | 542.83 |
| MAE | 9.38 |
| RMSE | 23.29 |
| R-Squared score | 0.80 |

For Long Short-Term Memory (LSTM) model, the scatterplot depicted in the Figure 13 exhibits a discernible pattern characterised by a notable separation between the data points and reference linear line. The observed spatial discrepancy suggests a significant divergence between the projected values and the actual results. The scatterplot reveals a discernible attribute indicating a restricted correspondence between the model's predictions and the actual data.

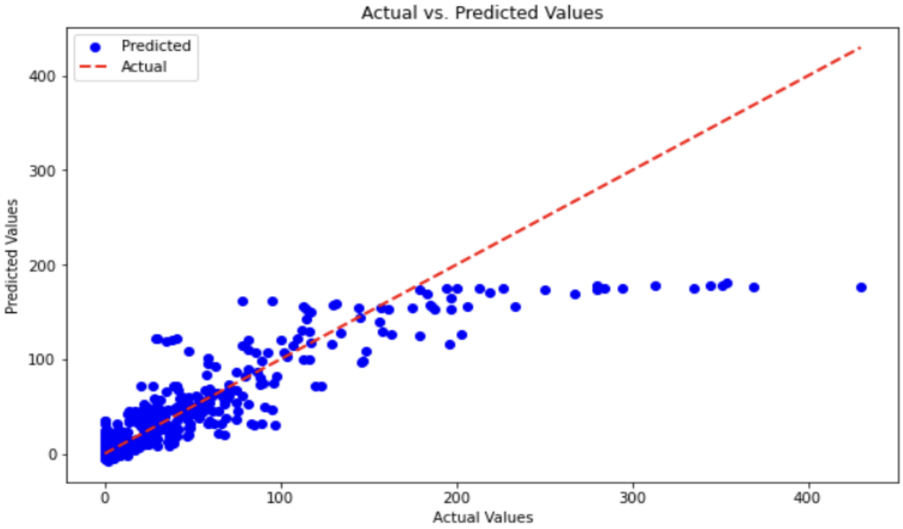


Figure 13: Scatterplot of LSTM model

### 6.2.2 Experiment 7 - Feedforward Neural Network(FNN)

When compared to LSTM model, the FNN model performs much better for this dataset. The result of FNN is shown in table 7. MAE is the average absolute deviation between predicted and the actual values. Having an MAE of 7.32, the FNN model's predictions deviate from actual quantity of crashes on the roads by an average of 7.32. The model's MSE value is 201.37. RMSE value of 14.19 implies that this algorithm's forecasts vary from actual number of road accidents by an average of 14.19. A high R-squared score of 0.92 suggests that the FNN model predicts approximately 92% of the variance in the number of road incidents. This indicates that algorithm closely matches the observed data.

Table 7: Performance of FNN model

| Model | Feedforward Neural Network (FNN) |
|---|---|
| MSE | 201.37 |
| MAE | 7.32 |
| RMSE | 14.19 |
| R-Squared score | 0.92 |

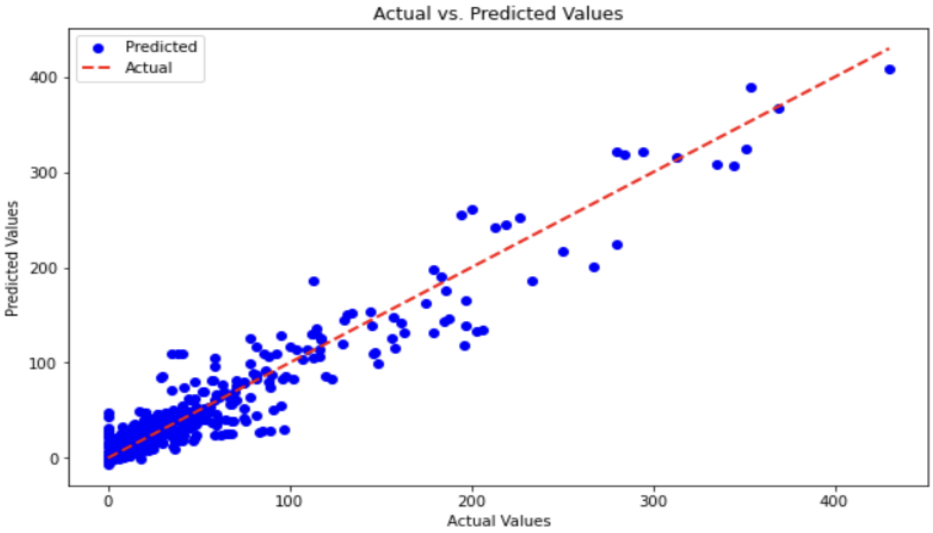


Figure 14: Scatterplot of FNN model

Figure 14 shows the sctterplot of FNN model. FNN model predicts better when compared to LSTM model.

### 6.3 Discussion

In this section, the major findings are evaluated critically. The comparative analysis of all the machine learning models with and without hyperparameter optimization are represented in table 8 and 9 respectively. As demonstrated by a comparison of machine learning models with default parameters and after the hyperparameter optimisation, the performance metrics such as MSE, MAE, RMSE, and R-squared score improved substantially after hyperparameter tuning. This means that hyperparameter optimisation has significantly contributed to improved model performance, generalisation, and prediction because the models are better able to recognise the data's underlying connections and patterns. In general, all models performed better following hyperparameter optimisation. Prediction errors are relatively low when using the Random Forest model. Across all the evaluation metrics, XGBoost algorithm performs better than Random Forest model. With respectably low MSE, MAE, and RMSE values, the Ridge Regression model performs well. Even though KNN has made progress, it still lags behind the other three models. It has larger prediction errors and it didn't perform well for this dataset. Having lower MSE, MAE, and RMSE values, Decision tree model performed decent. Although Decision Tree performs decent, it might not be as reliable as other algorithms. XGBoost outperforms all other machine learning models, followed by Random Forest and Ridge Regression.

The comparative analysis of deep learning models are shown in table 10. According to the performance measures, the MSE, MAE, RMSE, and R-squared score of FNN algorithm are better than those of LSTM model. Comparing the FNN model to the LSTM method , FNN indicates that it is more accurate and better fits the data. It is evident that LSTM model didn't perform well in predicting the road accidents.

Table 8: Comparative analysis of the Machine Learning models with default parameters

| Model | Random Forest | XGBoost | Decision Tree | KNN | Ridge Regression |
|---|---|---|---|---|---|
| MSE | 141.35 | **101.79** | 234.21 | 844.13 | 646.39 |
| MAE | 5.23 | **5.35** | 6.34 | 13.45 | 18.03 |
| RMSE | 11.88 | **10.08** | 15.30 | 29.05 | 25.42 |
| R-Squared score | 0.94 | **0.96** | 0.91 | 0.69 | 0.76 |

Table 9: Comparative analysis of the Machine Learning models after performing hyperparameter optimization

| Model | Random Forest | XGBoost | Decision Tree | KNN | Ridge Regression |
|---|---|---|---|---|---|
| MSE | 136.82 | **99.96** | 206.45 | 627.24 | 152.91 |
| MAE | 5.18 | **5.79** | 5.85 | 10.98 | 6.36 |
| RMSE | 11.69 | **9.99** | 14.36 | 25.04 | 12.36 |
| R-Squared score | 0.95 | **0.96** | 0.92 | 0.77 | 0.94 |

Table 10: Comparative analysis of the Deep Learning models

| Models | LSTM | FNN |
|---|---|---|
| MSE | 542.83 | **201.37** |
| MAE | 9.38 | **7.32** |
| RMSE | 23.29 | **14.19** |
| R-Squared score | 0.80 | **0.92** |

# 7 Conclusion and Future Work

This study used a variety of machine learning and deep learning methods to forecast the traffic incidents in Ireland. The dataset used in the project is obtained form the official website of Central Statistics Office (CSO) , which included information on the type of road user, age group, sex, statistic label, and year. Hyperparameter optimization was employed on machine learning algorithms to see if it will improve the performance of the model and Grid search was the chosen technique for that. Seven accident prediction models have been developed and evaluated in order to contrast their performances. The comparative analysis revealed that the optimisation of hyperparameters greatly improved the performance of all machine learning models. XGBoost outperformed the other models after tuning their hyperparameters, followed by Random Forest and Ridge Regression. The KNN algorithm performed unfavourably in comparison to other models. The Feed-forward Neural Network (FNN) performed better than the Long Short-Term Memory (LSTM) model for deep learning models. Overall, the study demonstrated that the machine learning and deep learning techniques are capable of predicting traffic accidents in Ireland. The accuracy of the models varied, and hyperparameter optimisation played an essential role in improving their predictive capabilities.

Future research in this field could involve additional experimentation with various hyperparameter optimisation techniques such as Random Search, Bayesian Optimisation and Genetic Algorithms in order to identify optimal set of hyperparameters for each of the model. Various other models can also implement on this dataset. Incorporating additional features or investigating the potential of feature engineering may also enhance the model performance. In addition, analysing the impact of external factors such as the weather, traffic volume, and road maintenance could improve predictive potential of the models. Incorporating real-time data from the roadway sensors or additional sources could also make forecasts more flexible and responsive. In addition, deploying best-performing model in the real-world applications, like creating accident prediction tools or giving real-time accident notifications could improve the road safety and reduce collision rates in Ireland.

# References

Alagarsamy, S., Malathi, M., Manonmani, M., Sanathani, T. and Kumar, A. S. (2021). Prediction of road accidents using machine learning technique, *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1695–1701.

Aldhari, I., Almoshaogeh, M., Jamal, A., Alharbi, F., Alinizzi, M. and Haider, H. (2023). Severity prediction of highway crashes in saudi arabia using machine learning techniques, *Applied Sciences* **13**(1).
**URL:** *https://www.mdpi.com/2076-3417/13/1/233*

Bao, J., Liu, P. and Ukkusuri, S. V. (2019). A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data, *Accident Analysis Prevention* **122**: 239–254.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0001457518303877*

Biswas, A. A., Mia, M. J. and Majumder, A. (2019). Forecasting the number of road accidents and casualties using random forest regression in the context of bangladesh, *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–5.

David, D. (2020). Hyperparameter optimization techniques to improve your machine learning model's performance, *freeCodeCamp* .
**URL:** *https://www.freecodecamp.org/news/hyperparameter-optimization-techniques-machine-learning/*

Dia, Y., Faty, L., Sarr, M. D., Sall, O., Bousso, M. and Landu, T. T. (2022). Study of supervised learning algorithms for the prediction of road accident severity in senegal, *2022 7th International Conference on Computational Intelligence and Applications (ICCIA)*, pp. 123–127.

Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M. and Yuan, J. (2020). Predicting real-time traffic conflicts using deep learning, *Accident Analysis Prevention* **136**: 105429.
**URL:** *https://www.sciencedirect.com/science/article/pii/S000145751930973X*

*Global Road Safety* (2023).
**URL:** *https://www.cdc.gov/injury/features/global-road-safety/index.html*

Hachcham, A. (2023). Xgboost: Everything you need to know, *Neptune* .
**URL:** *https://neptune.ai/blog/xgboost-everything-you-need-to-know*

Humera Khanum, Anshul Garg, M. I. F. (2023). Accident severity prediction modeling for road safety using random forest algorithm: an analysis of indian highways, *F1000Research* .
**URL:** *https://f1000research.com/articles/12-494*

Jiang, F., Yuen, K. K. R. and Lee, E. W. M. (2020). A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions, *Accident Analysis Prevention* **141**: 105520.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0001457519317713*

Kamali, K. (2023). Deep learning (part 1) - feedforward neural networks (fnn) (galaxy training materials).

Lee, J., Yoon, T., Kwon, S. and Lee, J. (2020). Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms: Seoul city study, *Applied Sciences* **10**(1).
**URL:** *https://www.mdpi.com/2076-3417/10/1/129*

Li, P., Abdel-Aty, M. and Yuan, J. (2020). Real-time crash risk prediction on arterials based on lstm-cnn, *Accident Analysis Prevention* **135**: 105371.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0001457519311108*

Malik, S., El Sayed, H., Khan, M. A. and Khan, M. J. (2021). Road accident severity prediction — a comparative analysis of machine learning algorithms, *2021 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, pp. 69–74.

Mbaabu, O. (2020). Introduction to random forest in machine learning, *Section* .
**URL:** *https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/*

Prasad, A. (2021). Regression trees — decision tree for regression — machine learning, *Medium* .
**URL:** *https://medium.com/analytics-vidhya/regression-trees-decision-tree-for-regression-machine-learning-e4d7525d8047*

Purbasari, A., Rinawan, F. R., Zulianto, A., Susanti, A. I. and Komara, H. (2021). Crisp-dm for data quality improvement to support machine learning of stunting prediction in infants and toddlers, *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pp. 1–6.

Rahim, M. A. and Hassan, H. M. (2021). A deep learning based traffic crash severity prediction framework, *Accident Analysis  Prevention* **154**: 106090.
**URL:** *https://www.sciencedirect.com/science/article/pii/S0001457521001214*

RSA (2023). 13%rise in road deaths recorded in 2022.
**URL:** *https://www.rsa.ie/news-events/news/details/2023/01/01/13-rise-in-road-deaths-recorded-in-2022*

Soni, A. (2020). Advantages and disadvantages of knn, *Medium* .
**URL:** *https://medium.com/@anuuz.soni/advantages-and-disadvantages-of-knn-ee06599b9336*

Yamparala, R., Challa, R., Valeti, P. and Chaitanya, P. S. (2022). Prediction of cyclist road accidents in india using machine learning and visualization techniques, *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pp. 476–481.