**Research Proposal/Synopsis for MS Thesis**

**Department of Computer Science and Information Technology**

**1. Name of the Student:** \_\_\_\_\_\_Muhammad Shujaat Abid\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2. VU ID:** \_\_\_\_MS210400061\_\_\_\_\_\_\_ **3. Session:** \_\_\_\_\_\_2021-23\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**4. Semester:** \_\_\_\_\_\_\_3rd\_\_\_\_\_\_\_\_\_\_\_\_\_\_ **5. Field of Specialization:** \_\_\_\_\_Database\_\_\_\_\_\_

**6. Title of Research Proposal:** \_\_\_\_\_ROAD ACCIDENT ANALYSIS AND PREDICTION

OF INJURY SEVERITY IN PEDESTRIANS USING MACHINE LEARNING PARADIGMS\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**7. Date of Enrolment in Research:** \_\_\_\_\_\_\_\_\_\_02/12/2022\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**8. Duration of Proposed Research:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2 Semesters\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**9. Total Funds Requested (if any) Rs.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_(**Rupees **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_)**

**===============================================================================**

**Supervisor, Supervisory Committee (SC) Information**

**1. Name of Supervisor:** \_\_\_\_\_\_\_\_\_\_\_\_\_\_ Dr. Mushtaq Hussain \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Designation:** \_\_\_\_\_ Assistant Professor \_\_\_\_ **Email ID:** \_\_mushtaq.hussain@vu.edu.pk\_

**Affiliation:** \_\_\_\_\_\_\_\_\_\_Virtual University of Pakistan\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**2. Name of Supervisory Committee (SC) Member 1:** \_\_\_\_\_\_\_\_ Dr Said Nabi \_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Designation:** \_\_\_ Lecturer \_\_\_\_\_ **Email ID:** \_\_\_\_ said.nabi@vu.edu.pk \_\_ **Affiliation:** \_\_\_\_\_\_\_Virtual University of Pakistan\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**3. Name of Supervisory Committee (SC) Member 2:** \_\_\_\_\_\_\_\_\_\_Saad Ahmed\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Designation:** \_\_\_\_Lecturer\_\_\_\_\_\_\_\_\_\_\_\_ **Email ID:** \_\_\_\_\_\_\_\_said.nabi@vu.edu.pk\_\_\_\_\_\_ **Affiliation:** \_\_\_\_\_Virtual University of Pakistan\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Topic**

PREDICTION OF INJURY SEVERITY IN PEDESTRIANS AND EXPLORING IT’S FACTORS USING MACHINE LEARNING PARADIGMS

**Abstract/Summary**

Pedestrian safety is a critical issue in urban environments. The World Health Organization (WHO) estimates that globally, about 270,000 pedestrians lose their lives in road crashes every year. Besides these mortalities, many more also suffer major and non-fatal injuries. These figures highlight the urgent need for comprehensive road collision analysis. It also highlights the importance of developing proper safety measures for the protection of vulnerable road users, specifically pedestrians.

The study uses the Killed or Seriously Injured (KSI) dataset that is freely available on Public Safety Data Portal and managed by Toronto Police Services. This dataset contains all the traffic collisions data from 2007 to 2022 of a Canadian city Toronto. The dataset includes information about pedestrian demographics, location of the collision, time of day, weather conditions, vehicles involved in the collision and other factors that may contribute to pedestrian major or fatal injuries.

The purpose of this study is to use the Traffic Collisions data to identify the trends in the dataset. Furthermore, it will do a deeper analysis of the data and try to figure out the characteristics of collisions. The study will also try to find out the underlying factors that result in fatal or major pedestrian injuries by applying a data mining technique known as association rule mining. Association rules are used to detect the relationship among the factors and the hidden patterns in the data. It will be helpful in discovering the factors that lead to the likelihood of fatal or major pedestrian injuries.

In addition to this, the study also aims to apply several other machine learning techniques such as clustering, and other classification algorithms. The researcher will try to train a model that can predict the fatal or major injuries in pedestrians. The performance of these developed models will be evaluated by using different measures. Most commonly used measures include precision, F1-score, recall and accuracy. The researcher will choose the best performing model and compare its results with the already developed models from literature and try to tune up the model to achieve better performance as compared to the already developed models. Lastly, the study will deploy the best performing model according to the given criteria.

**Introduction (Need/Motivation of the Present Research)**

In today's fast-paced world, roads, transportation systems and vehicles play a crucial role in connecting people, transferring goods, and services across cities, countries, and even continents.

The development and maintenance of road infrastructure have significant economic, social, and environmental impacts, affecting industries, governments, and societies at large. With the increase in the global population and urbanization, there is also a tremendous increase in the number of road vehicles over the years. So, the demand for efficient, secure and sustainable transportation systems has become more pressing than ever.

However, along with the benefits of road infrastructure come risks and challenges. Despite efforts to improve infrastructure, implement education and awareness campaigns, and enforce traffic laws, pedestrians remain vulnerable to collisions with vehicles. Road accidents, particularly those involving pedestrians, are a major public health concern worldwide. Pedestrians are the most vulnerable road users, and they account for a significant proportion of fatalities and injuries in road collisions as mentioned earlier in the WHO report statistics.

The Killed or Seriously Injured dataset presented by the Toronto Police Services provides a valuable resource for studying pedestrian collisions and fatalities. This dataset includes detailed information on the characteristics of collisions that result in fatal pedestrian injuries, including demographic data, collision location and time, and other factors that may contribute to the risk of injury. The data of road accidents consist of many dimensions and is quite heterogeneous and traditional statistical analyses may not be able to portray the relationship among the features and may not be able to detect the hidden patterns among these features that are important for predicting the likelihood of fatal pedestrian injuries.

Machine learning techniques, such as association rules, can provide a solution to this problem. Association rule mining refers to a class of machine learning methods which is based on rules and it is used to identify remarkable relationships between attributes within large datasets (Piatetsky-Shapiro et al 1991). The primary objective of this technique is to generate those rules that can provide help to predict the likelihood of occurrence of an item based on the incidences of some other items (Antonie et al 2008).

Association rules allow researchers to identify relationships and dependencies among variables, which can help to predict outcomes and identify risk factors. By applying association rules to the KSI dataset, this thesis aims to identify the factors that contribute to fatal or major pedestrian injuries in Toronto and to develop a predictive model that can be used to inform interventions and improve pedestrian safety. Once the factors are known to us, then we can take proper measures to reduce the chances of fatal or major pedestrian injuries.

Most of the work done in this field is of general nature and they aim to develop a model that predicts only the severity of collisions or the factors that lead to collisions. However, this study, with the aid of machine learning algorithms, specifically focuses on the pedestrian’s injury severity and will unveil the factors that lead to fatal or major pedestrian injuries. Moreover, the Toronto Police has recently updated the Killed or Seriously Injured (KSI) dataset that includes records of accidents from 2006 to 2022, this study will use this latest dataset for performing analysis. Lastly, the study will suggest some recommendations based on the analysis of data and the findings of the study that will help to reduce the fatal or major injuries in pedestrians.

This study has important implications for urban planning and transportation policy, as well as for the advancement of targeted interpositions to minimize the incidences of fatal or major pedestrian injuries by using data-driven approaches and machine learning techniques to address public safety issues. The study also demonstrates the value of applying innovative methods to complex problems.

**Research Question(s)**

Q1: What will be the estimated injury severity level of pedestrians in road accidents?

Q2: What are the underlying hidden factors that contribute to fatal or major pedestrian’s injuries?

**Research Objectives**

The objectives to fulfill this study are as follows

* To find out the factors that are more involved in pedestrian’s collisions using association rule mining
* To build a prediction model that will predict the injury severity in pedestrians by using the different machine learning algorithms.
* Evaluating the performance of these prediction models.

**Socio-Economic Importance/Benefits (if applicable)**

Human lives are the most critical capital for a country’s growth and development. The accidents could cost a lot to countries so it is required to minimize this loss in this age where the traffic is increasing day by day. The in-depth analysis of this study will present the factors that lead to pedestrians fatal or major injuries. These results could be used to minimize these factors so that the traffic accidents and pedestrian’s fatal injury rate could be reduced. Moreover, the study will try to predict the injury severity level in pedestrians that could be used in different ways to ensure the pedestrian’s safety.

**Review of Literature**

*Emu M et al in 2022* conducted research to develop a model that will be able to predict the consequence of a collision is either fatal or non-fatal. The focus of the research was the prediction of collision severity with the help of machine learning algorithms. For developing the required model, the researchers have used four machine learning methodologies that include Support Vector Machine, Random Forests, K-Nearest Neighbour, and Convolutional Neural Network. This paper has employed methods that are related to deep learning on a large dataset that is collected over 20 years and contains 5.8 million entries. The overall dataset contains 104 attributes that are reduced only to 67 without losing any performance of prediction models by using information content analysis technique. An in-depth analysis reveals that road traits, vehicle type, weather conditions, usage of safety devices, time of collisions, involvement class, and the status of traffic control are the aspects that participate most in the prediction models. The dataset has the problem of class imbalance and to tackle this issue in the study the method of under sampling has been used. The study has used accuracy, recall, precision and F1-score as performance measures for each method. The accuracy of all of these methods ranges from 69.38% to 75.56%. The KNN has the lowest accuracy of 69.38% and the CNN performed significantly higher than all of these methods with an accuracy of 75.56%. CNN also has better precision (75.74%) and recall (75.45%) than the other methods. So, these results show that Convolutional Neural Network (CNN) achieves the best performance in the prediction of accident seriousness by using both soft and majority voting. It has been concluded from the study that CNN is the most suitable option that can be deployed for the purpose of developing an ITS (Intelligent Transportation System) to perform the task of real-time fatality hazard predictor on various navigation apps.

*Shweta et al (2021)* conducted research to identify the factors that lead to road accidents. They had tried to investigate the attributes that have an influence on road accidents by using the machine learning techniques and then based on these causes they have strained to give some valuable suggestions to avoid and reduce the road accidents. For this purpose, they have used the Killed or Seriously injured (KSI) dataset available on Public Safety Data Portal managed by Toronto Police Services. This dataset contains all the information of traffic accidents that have been reported from 2007 to 2017. The data has been cleaned before using in the analysis to avoid any misleading information. The input features are selected using feature selection technique to reduce the dimensionality and the attribute of ‘FATAL’ was selected as an output. The researcher has applied several supervised machine learning approaches that includes K-Nearest Neighbor (KNN), Naïve Bayes, Decision Trees, and AdaBoost. The researcher has also applied several unsupervised machine learning techniques including K-mean clustering algorithm, CART and ROC value. The researcher has achieved 81.5% performance with the CART algorithm and AdaBoost also achieves the best performance among these. By analyzing the data, it has been revealed that the number of accidents has been slightly reduced over the years, and the most number of accidents happens from June to October. It has been concluded that most of the accidents that lead to fatal injuries happen in the east Toronto region and the victims are mostly pedestrians. Aggressive driving (62.9%) is the major cause of accidents, then over speeding (21.4%), signal violation at red light (10.4%) and then alcohol (5.3%). It has been concluded that the results of this study are not very general, but they have classified the secure and risky zones very well.

*Shanshal D et al in 2020* conducted research aiming to build a model that can predicts and identifies the hidden patterns in deadly and severe collisions happening in Toronto and predict the mortalities and major injuries of individuals that are involved in the collisions such as drivers, pedestrians and bicyclists using machine learning techniques and by doing analysis of the data. The researchers have used KSI dataset in their study that is available on Public Safety Data Portal and managed by the Toronto Police Services. The finalized dataset after cleansing includes 8922 observations and 26 variables. The initial analysis revealed that in the year 2007 453 collisions happened that reduces down to 331 in 2017. It was the lowest number of collisions since 2007. It has been analyzed that pedestrians, cyclists and motorcyclists are the most affected ones that get major or fatal injuries in the collisions. The researchers have used the data mining approach of association rules, classification algorithms including Lasso Regression and Random Forest. The researcher has found that Random Forest achieves best performance as a predictor of injury severity with an accuracy of 80% for both drivers and pedestrians and 89% accuracy for cyclists. The researcher has used the Apriori Algorithm to mine the dataset and for finding association rules in order to establish relationships between the variables and to uncover the hidden patterns in the dataset. Many association rules for each category have been created by the researchers. The researchers have used the two popular predictive model’s algorithms i.e. Lasso Regression (LR) and Random Forest (RF) and compared their performances and found out that Random Forest is more accurate. The accuracy, sensitivity and specificity are taken as performance measures for these two models. It has been concluded that the Random Forest model is more accurate in capturing the overall picture than Lasso Regression model. It is concluded that the factors of failing to yield right of way, driver’s distraction, aggressive driving, improper turns lead to severe collisions. Several environmental conditions such as clear and dry weather in summer and spring seasons lead to more accidents. The cyclist in the 6th and 7th calendar month gets fatal injuries. The cases of rainy weather, wet surfaces, and dark light causes a pedestrian's probability of deadly or serious injury. The cyclist gets fatal injuries when both the driver and cyclist are moving in the same direction and driver sideswipes the cyclist, a motorcyclist making a left turn through the cyclist’s path and the cyclist is hit by the opened door of vehicles. It has been recommended by the researchers to send off more patrolmen on to the roads to ensure road safety, traffic safety campaigns and public awareness messages should be spread to overcome aggressiveness and inattentiveness during driving.

*Gan J et al in 2020* got inspired by the Deep Forest algorithm that is grounded on the decision tree ensembled and conducted a study with the intent of developing a predictor model that will be able to predict the severity of traffic collisions built upon this DF algorithm. According to the researchers, it is the leading study that uses the Deep Forest algorithm to make the predictions related to the seriousness of traffic crashes. The researchers have used the UK road safety dataset in 2016 obtained from the Kaggle website. The dataset includes 18 items in total including latitude and longitude, light conditions, time characteristics, vehicle type, weather, gender, age, road surface, vehicle age, speed, and some other data features. First the researchers did the reliability verification of the dataset and then the pre-processing methods were applied. Then the researchers used a mixture of Grid and Randomized Search methods for the optimization of parameters that resulted into eight features, and these eight features are incorporated as the primary data attributes. For the validation of their proposed model and to compare the performance of Deep Forest algorithm, the researchers have incorporated other famous machine learning algorithms also that includes LightGBM, DNN, K-Nearest Neighbor, Random Forest, XGboost, and decision trees (DT). Then after the comparison it is concluded that Deep Forest has achieved the best performance among all of these. The Deep Forests achieves best performance i.e. recall is 0.92, an accuracy equal to 90.69%, highest F1 score and ROC of 0.91 and 0.93 respectively. Then the researchers explained several advantages of Deep Forests algorithm as compared to traditional machine learning methods. The high prediction accuracy of Deep Forests algorithms reveals that it can be used more effectively for the prediction of accident severity, and it will be more conductive to the transplantation of models as it requires fewer hyper-parameters as compared to other models. This model can be adopted certainly in order to tackle most of the problems that are related to traffic accidents.

The summarized related work is shown in the following table:

| **Sr No** | **Title** | **Citation** | **Objective** | **Strengths** | **Weaknesses** | **ML Technique** | **Performance** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Fatality Prediction for Motor Vehicle Collisions: Mining Big Data Using Deep Learning and Ensemble Methods | Emu, M., Kamal, F. B., Choudhury, S., & Rahman, Q. A. (2022). Fatality prediction for motor vehicle collisions: mining Big Data using deep learning and ensemble methods. *IEEE Open Journal of Intelligent Transportation Systems*, *3*, 199-209. | Developing a model that predicts whether the outcome of a collision will be fatal or non-fatal. | Sufficient dataset used for training and necessary pre-processing has been done | Missing some important features (related to driver behavior and car condition)  More models can be trained based on other classifiers  Performance can be further improved by tuning-up the models | Random Forests, K Nearest Neighbor, SVM, Convolutional Neural Networks | Accuracy of all these ranges from 69.38% to 75.56%.   CNN is best with Accuracy= 75.56% Precision = 75.74% Recall = 75.45% |
| 2 | A Framework for Analyzing Road Accidents Using Machine Learning Paradigms | Yadav, J., Batra, K., & Goel, A. K. (2021, August). A Framework for Analyzing Road Accidents Using Machine Learning Paradigms. In *Journal of Physics: Conference Series* (Vol. 1950, No. 1, p. 012072). IOP Publishing. | Identifying factors that have an impact on road accidents. | An in depth analysis of the dataset has been shown in the paper | Its results are not very significant as it is unable to reveal major facts  Performance measures are not discussed explicitly for all the algorithms | Supervised: AdaBoost  Naïve Bayes K-Nearest Neighbour (KNN) Decision tree  Unsupervised:  K-means clustering algorithm  CART, ROC value | 81.5% accuracy with CART algo  AdaBoost achieves best performance. |
| 3 | PREDICTION OF FATAL AND MAJOR INJURIES OF DRIVERS, CYCLISTS, AND PEDESTRIANS IN COLLISIONS | Shanshal, D., Babaoglu, C., & Başar, A. (2020). Prediction of Fatal and Major Injury of Drivers, Cyclists, and Pedestrians in Collisions. *Promet-Traffic&Transportation*, *32*(1), 39-53. | Building a prediction model that predicts the fatal accident and serious injuries in traffic crashes. | An in depth analysis of the dataset is provided using both association rules and predictive models  Performance of predictor mode in satisfactory | Blank or missing values are not handled properly  Does not provide a direct performance comparison of the developed model | Data mining through Association rules and Classification algo:  Random Forest (RF), Lasso regression (LR) | 80% accuracy of random forest for drivers and ped and 89% accuracy for cyclist  Accuracy, sensitivity, specificity |
| 4 | An Alternative Method for Traffic Accident Severity Prediction: Using Deep Forests Algorithm | Gan, J., Li, L., Zhang, D., Yi, Z., & Xiang, Q. (2020). An alternative method for traffic accident severity prediction: using deep forests algorithm. *Journal of advanced transportation*, *2020*, 1-13. | Developing a model that will predict the seriousness of traffic collisions based on Deep Forest algorithm. | Strong analysis and data pre processing techniques have been used to ensure the quality of the data  The developed model has significant accuracy and reliable | Limited number of features are employed in the study | Deep Forest algorithm | Accuracy = 90.69% Recall = 0.92 False alarm rate = 0.09 F1 score = 0.91 ROC = 0.93 |

After conducting a thorough review of the literature, we may conclude that a good deal of effort has been put down in the field of accident data mining, particularly in the domain of prediction models. However, most of the studies focus only on one domain i.e. building the prediction model or finding the factors that involve traffic accidents, only a few studies combining both of these aspects. The majority of these studies utilizes machine learning methods on several different datasets containing attributes related to accidents such as year, month, time, road surface, weather, vehicle type and many more and demographic information of the involvements with the injury severity level.

This study aims to find out the injury severity level of one involvement type i.e. pedestrians and the factors that contribute to fatal or major injury in the pedestrians which is a distinguishing factor of this study. The dataset will be analyzed statistically and predictively using several different machine learning procedures in order to find out the hidden patterns and obtain the rules and factors that contribute to pedestrians' fatal or major injury. The study will also provide suggestions and recommendations to ensure the pedestrian’s safety.

**Methodology/Research Design**

This study aims to specify the previously done general work in the field of accident mining and will use the Apriori algorithm for mining the dataset and uncovering rules and hidden patterns between the variables. Agrawal et al provides a proper description of association rules given below:

Let A is a set of 'n' attributes A = {a1, a2, …, an} and T is a set of 'm' transactions represented as T = {t1, t2, …, tm} that make up a database. Each transaction has a few of the items from A in it. A rule is the same as saying X leads to Y (X => Y), where both X and Y belong to A, the former is called as antecedent and later as consequent. Basically, a rule is a way to anticipate a transaction in the database.

The significance of rules is evaluated using the following three measures:

**Confidence:** It measures the frequency of the rule occurrence within the database. It is calculated as given below:



**Support:**  It indicates the frequency of appurtenance of an item in the transaction. Basically, the support of a data item X in relation to a set of transactions T is calculated as follows:



**Lift:** It Figures out the likelihood of a rule happening based on the chances of the antecedent and consequent being unrelated. It is calculated as follows:



The lift value >1 indicates that the two data items are dependent on each other and form a true rule.

A bunch of algorithms have been created in the last few decades for discovering association rules and we have chosen the Apriori algorithm because of several benefits that it provides. The Apriori algorithm is well-known for its ability to process huge datasets with ease. It is an iterative approach that eliminates the need to scan the entire database over and over again, instead employing a candidate generation and pruning strategy that reduces computational complexity. It is also relatively simple to understand and implement, as it is based on the concept of frequent itemsets and uses a straightforward, breadth-first search strategy, making it a good starting point for those new to the area of association rule mining. (Agarwal et al, 1994).

Besides uncovering the frequent itemsets, the Apriori algorithm also generates associated rules that can be used in a variety of contexts, such as decision-making, market basket analysis, and recommendation systems. Furthermore, the Apriori algorithm provides the ability to set thresholds for both support and confidence, allowing users to adjust the algorithm to their own needs and ensure that meaningful and statistically significant rules are generated (Han et al, 2011). However, this algorithm may work slowly because it needs to look through multiple sets of data to come up with the rules. Its speed depends upon size of the dataset, how many items it has, and the minimum measures chosen.

This study uses Python for the implementation of *Apriori* algorithm. Python provides an extensive range of libraries and packages such as ‘mlxtend’, ‘pyfim’, and ‘Orange’. These are dedicated to data mining and provide support for data preprocessing, rule generation, and evaluation. These are also able to handle large amounts of datasets in an efficient manner.

**Research Design:**

The proposed methodology for this study has been divided into six steps from data collection, data pre-processing, training and evaluating model, tuning model to the deployment of predictor model and rules extraction. The flow chart of the research design is shown below:



The proposed framework is as follows:

**(i)** **Data collection:**

The first step is data collection. We will obtain a Killed or Seriously Injured (KSI) dataset that is freely available on public safety data portal and managed by the Toronto Police Services (https://data.torontopolice.on.ca/). The KSI contains over 18,000 records, where each row represents a unique involvement in the accident. This dataset contains all the accidents that are reported by Toronto Police from 2006 to 2022.

**(ii)** **Data pre-processing:**

After collecting data, the next step is the data pre-processing that is the most crucial one as the working and performance of machine learning algorithms highly depends on the quality of input dataset. In this step, we will take the subset of the KSI dataset that includes only those records in which pedestrians are involved. All the data pre-processing techniques such as data cleaning, removing the null values and standardizing the values of certain attributes including date and time are applied in this step.

**(iii)** **Feature selection:**

In this step, we first understand our dataset and then we take an insight analysis of the dataset using techniques such as Pearson correlation and chi-square tests that help us in identifying and selecting the best features that are most appropriate to our machine learning model for predicting the rules. We also spend some time in creating certain new features from our input dataset to achieve better performance and to find out the hidden patterns in the dataset. We also transform all the descriptive data values into nominal form so that our machine learning algorithm i.e. apriori can work with this dataset.

**(iv)** **Extracting rules:**

At this stage, we will use Python for finding hidden patterns and creating association rules. We will select different input features at a time to get the related rules. In this stage, while extracting rules will select several different parameters such as the minimum values of confidence, support and other measures that can be helpful in finding best rules within the given range of parameters.

**(v)** **Training and evaluation:**

After getting rules, we will train our model. We'll split up our data set into two pieces using k-fold cross validation technique, such that we will use 90% of it to train our model and the other 10% for testing it. We will apply several machine learning algorithms and then evaluate their performances in order to get the best model.

**(vi)** **Model tuning and deployment:**

In the last stage, after doing analysis and evaluating performance of different models, we will try to further tune our developed model to achieve the maximum performance. Then we will be in a position in which we recommend and deploy the best predictor.

Lastly, based on rules and after doing analysis of different factors that lead to fatal or major pedestrian’s injuries, we will present some suggestions to minimize the pedestrian’s hit rate and will recommend several safety measures to ensure the pedestrian’s safety.

**References/Bibliography**

[1] Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowledge discovery in database*, 229-248.

[2] Antonie, L. (2008). INTRODUCTION TO DATA MINING ASSOCIATION RULES. *University of Guelph.*

[3] Emu, M., Kamal, F. B., Choudhury, S., & Rahman, Q. A. (2022). Fatality prediction for motor vehicle collisions: mining Big Data using deep learning and ensemble methods. *IEEE Open Journal of Intelligent Transportation Systems*, 3, 199-209.

[4] Yadav, J., Batra, K., & Goel, A. K. (2021, August). A Framework for Analyzing Road Accidents Using Machine Learning Paradigms. In *Journal of Physics: Conference Series* (Vol. 1950, No. 1, p. 012072). IOP Publishing.

[5] Shanshal, D., Babaoglu, C., & Başar, A. (2020). Prediction of Fatal and Major Injury of Drivers, Cyclists, and Pedestrians in Collisions. *Promet-Traffic&Transportation*, 32(1), 39-53.

[6] Gan, J., Li, L., Zhang, D., Yi, Z., & Xiang, Q. (2020). An alternative method for traffic accident severity prediction: using deep forests algorithm. *Journal of advanced transportation, 2020*, 1-13.

[7] Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207-216).

[8] Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

[9] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques,‖ Morgan Kaufmann Publishers Inc. *San Francisco, CA, USA.*

**Gantt Chart (to be used as guideline)**

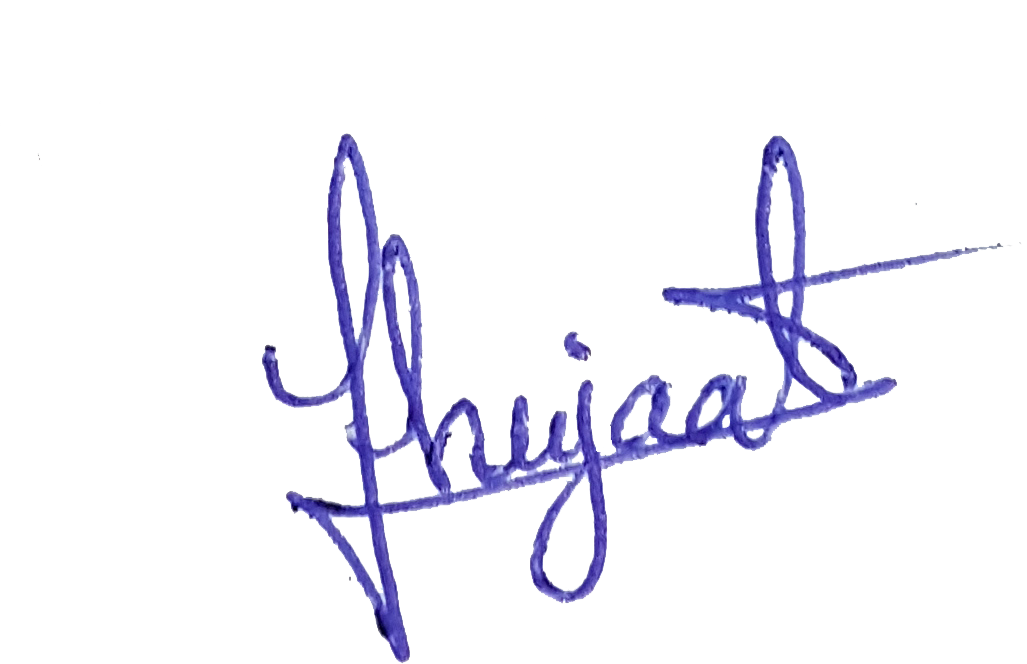
Activity Plan is as follows:

| **Specific Objectives** | **Activities** | **Semester-I** | | | | **Semester-II** | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Month-I** | **Month-II** | **Month-III** | **Month-IV** | **Month-I** | **Month-II** | **Month-III** | **Month-IV** |
|  | Literature’s Survey | Dec 2022 |  |  |  |  |  |  |  |
| I | Brainstorm ideas |  |  |  |  |  |  |  |  |
| Select candidate research ideas |  |  |  |  |  |  |  |  |
| Feasibility of research ideas |  |  |  |  |  |  |  |  |
| Finalize research idea |  |  |  |  |  |  |  |  |
| II | Literature Review |  |  |  |  |  |  |  |  |
| Refine research idea |  |  |  |  |  |  |  |  |
| Synopsis writing and submission |  |  |  |  |  |  |  |  |
| II | Prepare dataset for model training |  |  |  |  |  |  |  |  |
| Develop machine learning based classification and regression models |  |  |  |  |  |  |  |  |
| Evaluate machine learning models and select best performing model |  |  |  |  |  |  |  |  |
| III | Deployment of proposed model |  |  |  |  |  |  |  |  |
|  | Thesis writing & Submission |  |  |  |  |  |  |  | Sep 2023 |

**Details of Funds/Expenditure (if applicable)**

Provide detail of funds and expected expenditure on this research such as Consumable/Chemicals/field Survey/Transport etc.

| **S. No.** | **Details** | **No(s)** | **Unit Rate** | **Total** |
| --- | --- | --- | --- | --- |
| 1 | Consumables & Chemicals |  |  |  |
| 2 | Surveys or Transport |  |  |  |
| 3 | Contingencies |  |  |  |
| 4 |  |  |  |  |
| 5 |  |  |  |  |
| **Grand Total** | | | |  |

** Student Signature**

**Date:** \_ 01/04/2023\_

**DECLARATION**

We hereby agree to supervise the research work as per above proposal/synopsis.



**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Signature of Supervisor**

**** 

**Signature of SC Member 1 Signature of SC Member 2** 

**Date: \_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_**

*Note: Hard and soft copy of synopsis/research proposal must be submitted to secretary ASRB for final approval.*

**Profile of Supervisor**

**Name of Supervisor:** \_\_Dr Mushtaq Hussain\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Designation:** \_\_\_\_Assistant Professor

* Total No. of Impact Factor Research Publications during last 5 years: \_\_\_\_7
* Total No. of Publications without Impact Factor during last 5 years: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_4



| **Ongoing**  **Research students** | |
| --- | --- |
| Number of MS/M.Phil. students | Number of PhD students |
| 4 |  |

**Signature of Supervisor**

| Endst. No. \_\_\_\_\_\_\_\_\_\_\_ | Dated:\_4-3-2023\_\_ |
| --- | --- |

The Proposal entitled “ ”duly recommended by the Graduate Research Committee (GRC) in its meeting held on \_\_\_\_\_\_\_\_\_\_ is forwarded to ASRB through the Dean of the Faculty for approval and allocation of funds (if requested).

|  | **Signature / Seal**  **Chairperson of the Department**  **Date: \_\_\_\_\_\_\_\_\_\_\_** |
| --- | --- |
| **Signature / Seal**  **Dean of the Faculty**  **Date: \_\_\_\_\_\_\_\_\_\_\_** | **Signature / Seal**  **Secretary ASRB**  **Date: \_\_\_\_\_\_\_\_\_\_\_** |