Gyalpozhing College of Information Technology
Royal University of Bhutan

ITS307 Data Analytics

Bachelor of Science in Information Technology

Taxi Fare Prediction System

Researchers:
Nima,12200067
Purna Kumar Limbu,12200075
Sonam Wangchuk,12200084
Tshering Dekar, 12200091

Module Tutor: Mrs. Nima Dema

Table of Contents

1. Proposed Methods **2**

1.1 System Overview **2**

1.2 Algorithm **3**

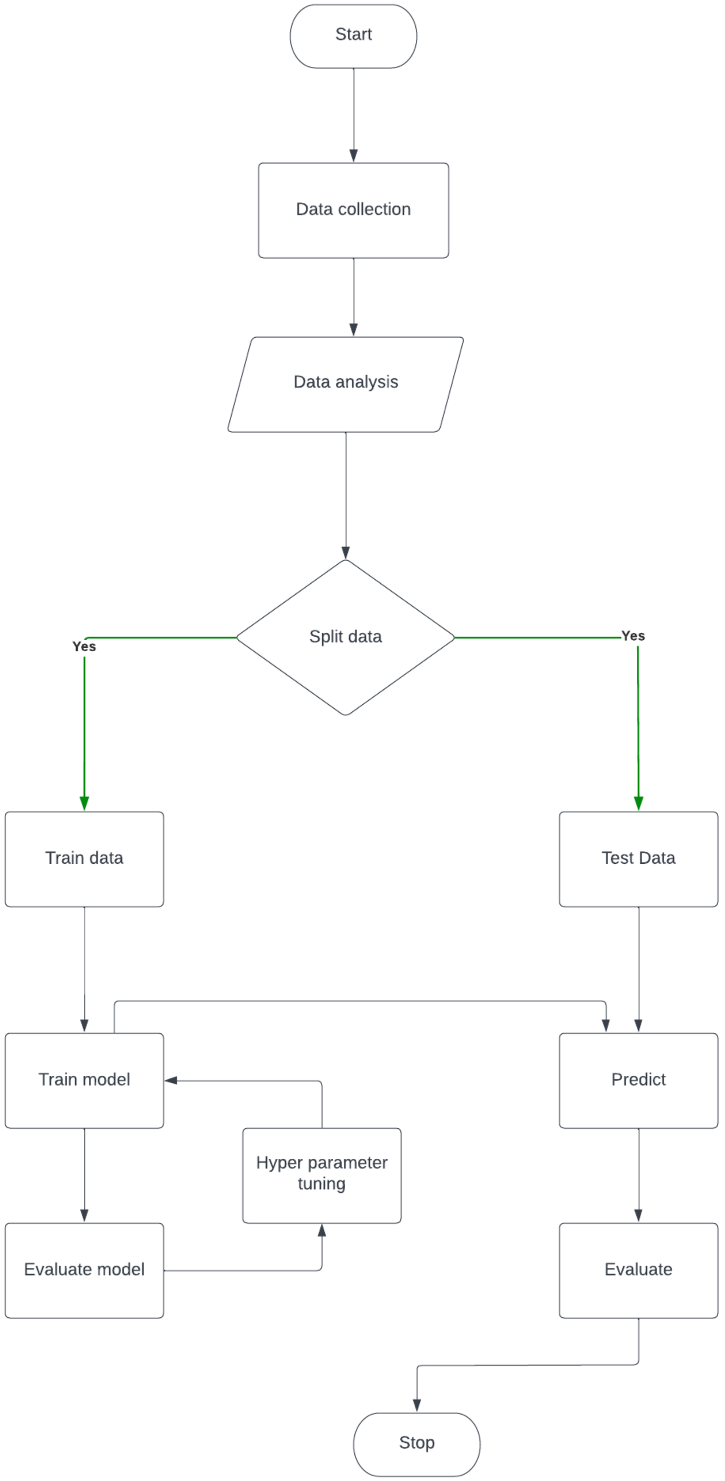
1.3 Dataset **4**

1.4 Evaluation Metrics **4**

1.5 Experimental Setup **5**

1. Proposed Methods

1.1 System Overview



The web application will be based on three-tier architecture; presentation layer, application layer and data layer. The data provided by the user in the presentation layer (user interface) will be processed in the application layer and the data will be stored in the data layer where prediction takes place.

1.2 Algorithm

To train the model, we used the KNN algorithm. Among other algorithms, the KNN algorithm proved to be the best algorithm. Both training data and test data's accuracy were comparatively higher and close to each other.

As for the parameter, we chose n_neighbors as 3 and obtained 91% accuracy score in train data and 93% accuracy score in test data. We tried changing the n_neighbors up to 10 and observed no considerable changes in the accuracy of the train data and test data, therefore, n_neighbor of 3 was kept unchanged.

The KNN algorithm works on the similarity between new and stored data points (training points) and classifies the new test point into the most similar class among the available classes. The KNN algorithm is non-parametric, and it is called the lazy learning algorithm, meaning that it does not learn from the training dataset, but rather stores the training dataset. When classifying the new dataset (test data), it classifies the new data based on the value of k, where it uses the Euclidean distance to measure the distance between the new point and the stored training points. The new point is classified into a class with the maximum number of neighbors. The Euclidean distance function (Di) was applied to find the nearest neighbor in the feature vector.

1.3 Dataset

The project uses the existing datasets acquired from Road Safety Transport and Authority (RSTA) to build the model. The existing datasets of taxi fare provided by RSTA will be used or analyzed for this project. This dataset has a mixture of both numerical and categorical features. Features include pickup point, destination point, distance(km), Fuel price, Fuel consumption and number of seats.

1.4 Evaluation Metrics

We used Mean Squared Error and R2 Score to check the performance of the model.

Mean Squared Error

$$\boxed{MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

R2 Score

$$R2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y_i})^2}{\sum_{i=1}^{N}(y_i - \bar{y})}, \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

1.5 Experimental Setup

Programming language: Python
Platform for training model: Jupyter Notebook
Python libraries for data processing and data visualization
Pandas, Numpy, Sklearn, Matplotlib, Seaborn,