**Contents:-**

1.Taxi Fare Prediction

1.1. Abstract

1.2. Introduction

1.3. Objective of Research

1.4. Problem Statement

2. Review of Literature

3. Data Collection

4. Methodology

4.1. Exploratory Data Analysis

4.1.1. Figures and table

4.2 Statistical techniques and visualization

4.3 Data Modelling and visualization

5. Findings and Suggestions

6. Conclusion

**TAXI FARE PREDICTION**

**1.About project**

**1.1 Abstract:-**

* Present days it is difficult to estimate the taxi fare using dynamic conditions such as Day, time, traffic conditions and it’s also vary from locationto location.
* To avoid this situationwe are deploying a model that will be used for Taxi fare estimation.
* Goal of this model is to predict the fare of a taxi trip given information about the pickup location and drop off location, the pickup date & time and also number of passengers travelling.

**1.2Introduction:-**

* There are a great many programs for predicting a taxi fare both on the Website and on the smart phone application market, and there has also been some research into the routing options of the electronic map. However, the results are not precise enough and are even limited to the user setting the fare rate and other information, not to mention safety protection.
* To facilitate the user’s choice of the best route and to predict the fare in order to save time and money by using electronic map real-time information, our design uses machine learning method.
* Electronic maps and GPS to create a mash up application which then combines both short message service functions to share information between users, so that our program can conduct actual simulation, send the current location to friends and family to ensure safety and concurrently store the coordinates of the driving route trace.
* The millions of rides taken each month can provide insight into traffic patterns, road blockage, or large-scale events that attract many New Yorkers.
* It is increasingly important for taxi companies to provide visibility to their estimated fare and ride duration, since the competing apps provide these metrics upfront.
* Predicting fare and duration of a ride can help passengers decide when is the optimal time to start their commute, or help drivers decide which of two potential rides will be more profitable.
* In order to predict duration and fare, only data which would be available at the beginning of a ride was used.
* This includes pickup and dropoff coordinates, trip distance,start time,number of passengers.
* Decision tree models were used to predict duration and fare amount.

**1.3. Objective of research**

The above problem made me to go for a research about how can we predict the Fare of Taxi making easier. Through many documentation and cases, it came out that machine learning and data science can make the work easier and faster. The objective of this project is to analyse dataset which consist of Taxi Fare criteria depending upon dataset consisting of previous data.

**1.4. Problem Statement**

Finding out the prediction of Taxi Fare by usingMachine learning algorithm using python as core.

**2. Review of literature**

The Main Objectives of Taxi Fare Prediction:

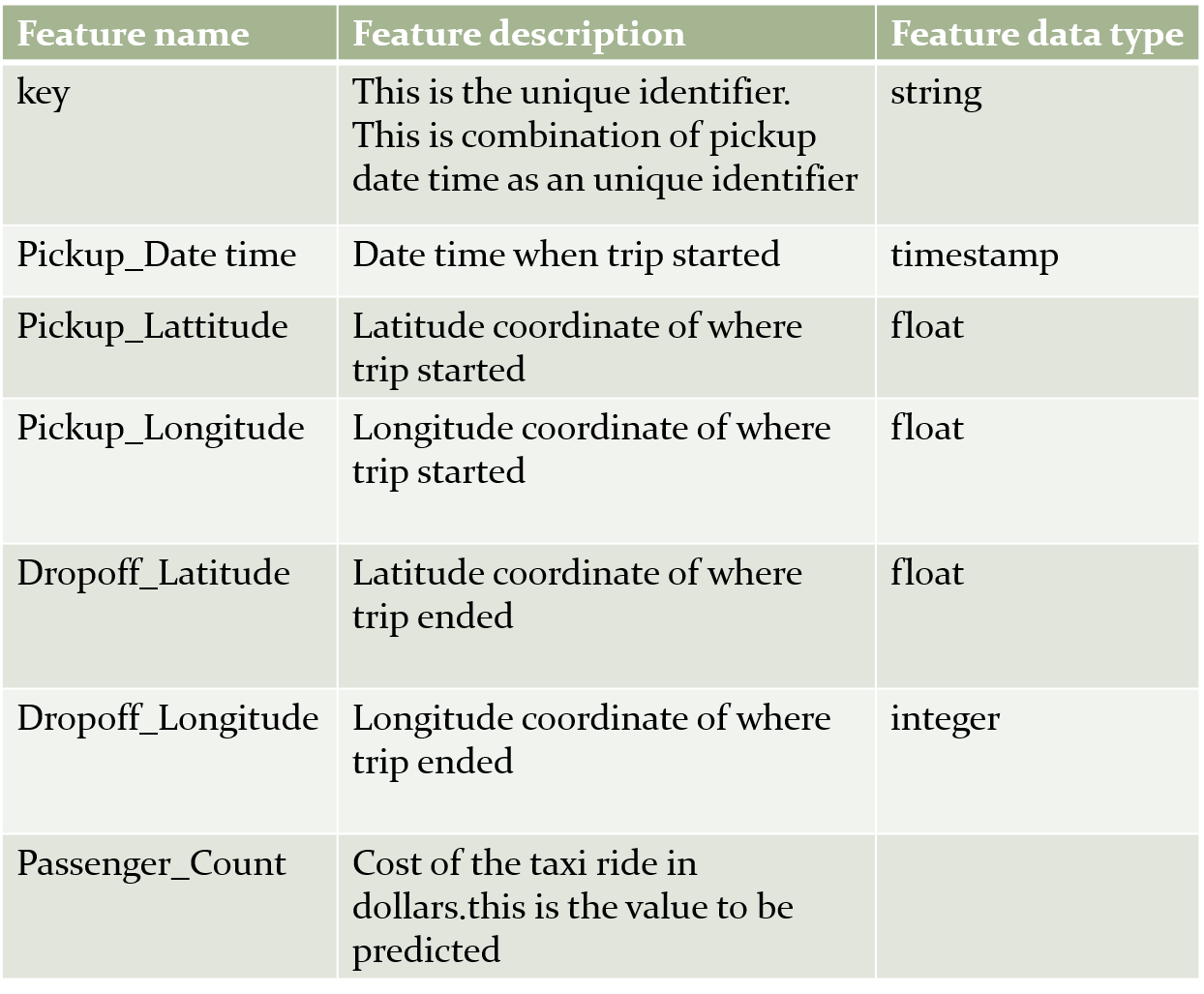
* To predict the Taxi fare of an Individual between two locations.
* To know the individual fare with the help of Pickup\_Lattitude,Pickup\_Longitude,Dropoff\_Lattitude, Dropoff\_Longitude attributes.
* It will helps the passenger to known the cost between two different locations before taking ride.

**3. Data Collection:**

Thedata for the project was accessed from the UCI MachineLearning Repository([https://www.kaggle.com/uciml/new-york-city-taxi-fare-prediction#train.csv](https://www.kaggle.com/uciml/new-york-city-taxi-fare-prediction%20#train.csv) ).This dataset contains of taxi rides between the locations of New York Citizens.The dataset used in this project is taken from Kaggle. The dataset obtained from Kaggle is maintained and updated by the Taxi Drivers. Fare Prediction dataset from Kaggle is used in CSV format.

**4. Methodology**

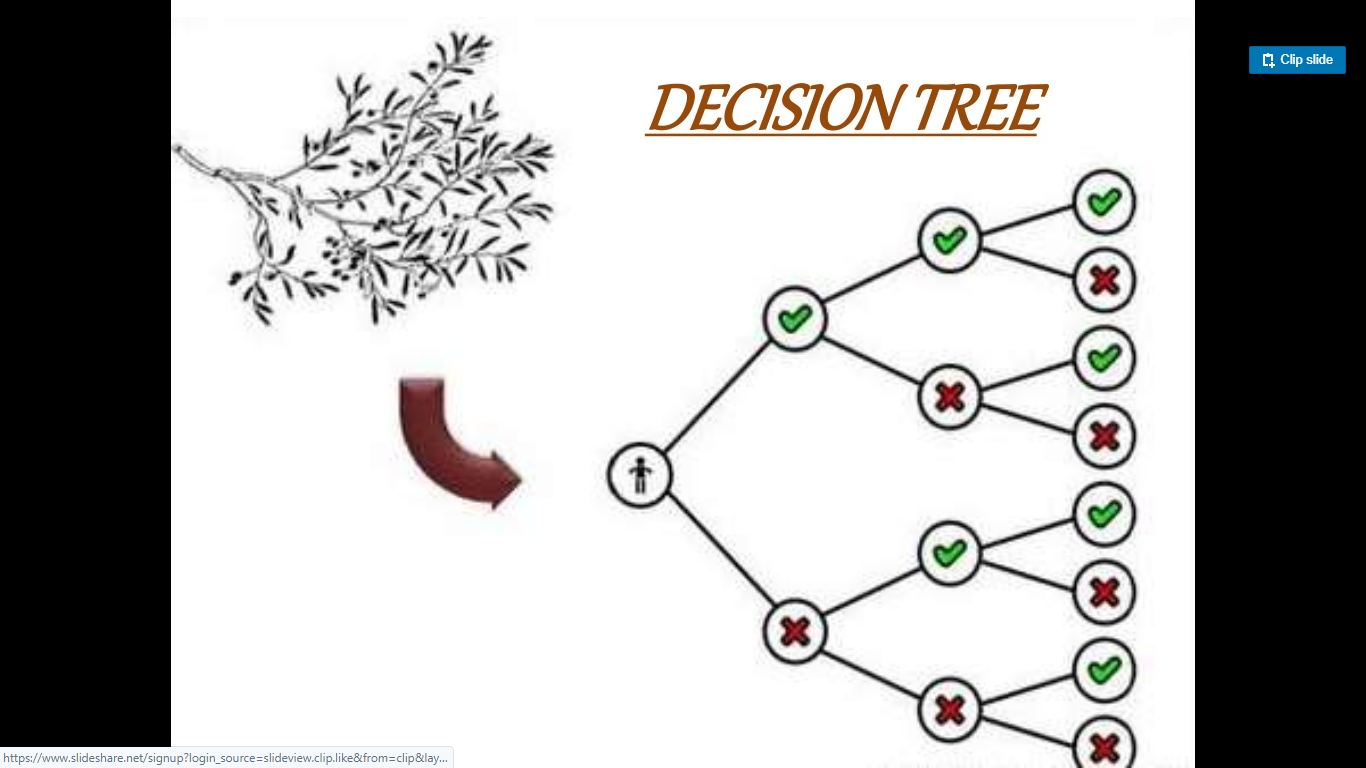
**4.1. Exploratory Data Analysis**



**4.1.1 Figures and Tables**

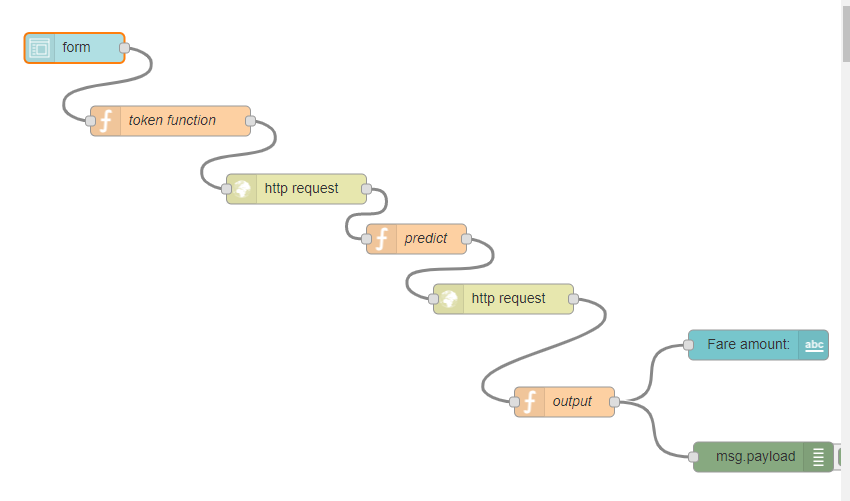
**Model used:-**

* [Decision Tree](https://www.geeksforgeeks.org/decision-tree/) is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.
* Here we used Decision Tree Regression than other models because the accuracy of dataset is better than other models.

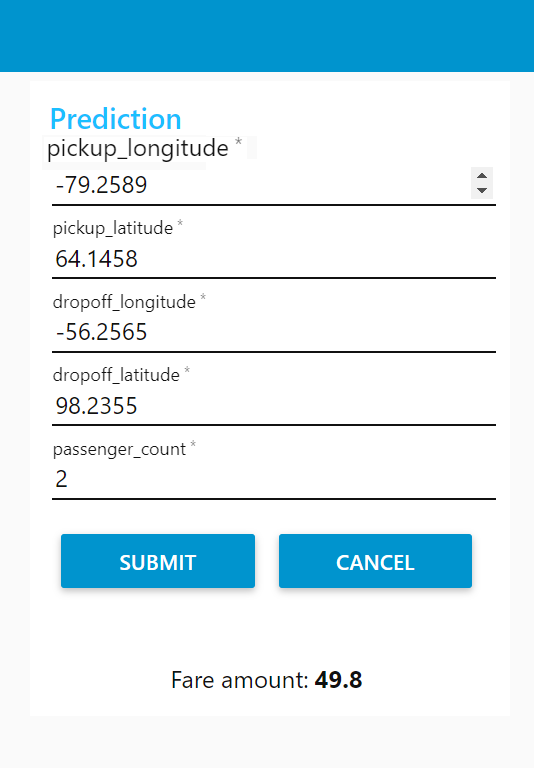
****

* Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
* The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node.

|  |  |  |
| --- | --- | --- |
| Decision trees can handle both categorical and numerical data.Datasets used:- WhatsApp Image 2019-07-16 at 12.30.05 PM.jpeg  **GRAPH:-**  **C:\Users\admin\Downloads\WhatsApp Image 2019-07-16 at 9.46.50 PM.jpeg** 4.2 Statistical techniques and visualization:  * + - * A supervised machine learning approach of Decision Treeis used for study. Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a [non-parametric](https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/) **supervised learning** method used for both **classification**and **regression**tasks.       * Tree models where the target variable can take a discrete set of values are called **classification trees**. Decision trees where the target variable can take continuous values (typically real numbers) are called **regression trees**. Classification And Regression Tree (CART) is general term for this.       * From "sklearn.model\_selection" we import "train\_test\_split " ,and we declare variables x\_train, x\_test, y\_train,y\_test and these are assigned to train\_test\_split of x,y with test\_size =0.2, and we declare here random\_state as zero.       * Random Sampling (Train and Test)   **Training Sample:**Model will be developed on this sample. 70% or 80% of the data goes here.   * + - * **Test Sample:** Model performances will be validated on this sample. 30% or 20% of the data goes here. |  |  |
| **4.3 Data Modelling and Visualization:**   * + - * Node-RED is a flow-based development tool for visual programming developed originally by IBM for wiring together hardware devices, APIs and online services as part of the Internet of Things. Node-RED provides a web browser-based flow editor, which can be used to create JavaScript functions. Elements of applications can be saved or shared for reuse. The runtime is built on Node.js. The flows created in Node-RED are stored using JSON. Since version 0.14 MQTT nodes can make properly configured TLS connections. In 2016, IBM contributed Node-RED as an open source JS Foundation project. |  |  |



**Output:**



**5. FINDINGS AND SUGGESSTIONS**

Decision tree classifier is applied to the data to answer the following questions:

* What is Taxi Fare of an individual with certain attributes?

Assessing whether an individual with certain attributes (key,Pickup\_Date time,Pickup\_Lattitude,Pickup\_Longitude,Dropoff\_Lattitude,Dropoff\_Longitude) will be checked. A function that predicts Taxi Fare based on the fitted model and individual attribute is provided in the ‘python’ notebook.

* What are the key features determining Taxi Fare?

Identifying what are the top 5 features explaining much of the difference between Pickup\_Lattitude,Pickup\_Longitude,Dropoff\_Lattitude,Dropoff\_Longitude. Determining the key features can help in policy formulation by identifying the few factors that can give most of the gains in Fare.

**6.Conclusion:**

The conditional inference tree was more effective at classifying the Over Fifty column because the prediction of the dataset is more effective when based on a permutation of variables. The conditional inference tree seems to avoid the bias that decision trees can have due to the information gain of individual variables. Instead the condition inference tree selects the variable to split on based on the statistical significance of a permutation of variables. The applications of this prediction model are mostly towards understanding which qualities lead individuals to Predicting Taxi Fare.

With the help of machine learning technology, it has become easy to find out relation and patterns among various data’s. The work in this project mainly revolves around predicting the Taxi Fare. Using the concept of machine learning we have built a model. We generated the Fare value which helps the Taxi Passengers.

The Taxi Fare dataset describes passengers for the past few years in New York city.

To explore and understand the dataset using a suite of line plots for the series data and histogramfor the data distributions.

How to use the new understanding of the problem to consider different framings of the prediction problem, ways the data may be prepared, and modeling methods that may be used.