1

Vehicle Accident Prediction Model

Based On Weather Factors

Center Coordinator: Mr. Prashant Karhale

External Guide: Mr. Akshay Tilekar

Presented by:

Snehal Padekar

Pooja Maske

Introduction

1. Traffic accidents are extremely common. If you live in a sprawling metropolis like I do, chances are that you've heard about, witnessed, or even involved in one. Because of their frequency, traffic accidents are a major cause of death globally, cutting short millions of lives per year.
2. Therefore, a system that can predict the occurrence of traffic accidents or accident-prone areas can potentially save lives.
3. Although difficult, traffic accident prediction is not impossible. Accidents don't arise in a purely stochastic manner; their occurrence is influenced by a multitude of factors such as driver's physical conditions, vehicle types, driving speed, traffic condition, road structure and weather.
4. Studying historical accident records would help us understand the (potentially causative) relationships between these factors and road accidents, which would in turn allow us to build an accident predictor.

How to prevent accidents?

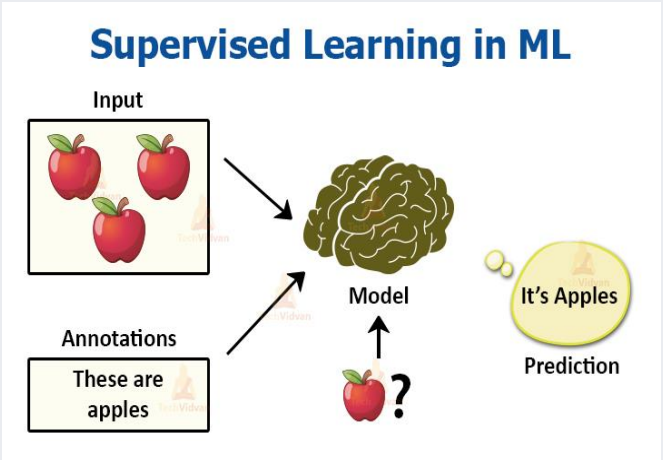
To prevent this situation we need more data information of about the various factors for prediction.

To predict the accident will happen or not..? We are going to used various factors like humidity, wind speed, intensity, wind direction, condition of road with respect to weather, light condition etc.

Here the classification & prediction comes into picture to predict the best outcomes using the various supervised machine learning algorithms

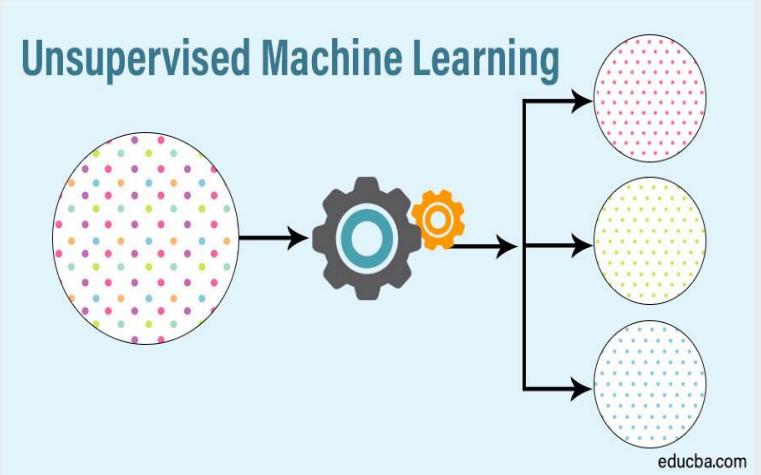
What is supervised learning?

- The concept of supervised learning depends on the labelled data.
- Supervised learning means having a full set of labeled data while training an algorithm.
- The collected data are labelled so that we know what input needs to be mapped to what output.
- The labelled dataset has both input and output parameter.
- It helps us to correct the algorithm if we make any mistake in the answer or output.
- Supervised learning can make new predictions for new unseen data.
- Mathematically, supervised learning can be shown as a linear function, i.e. *y=f(x),* where x is the input and y is the output.



What is unsupervised learning?

- In unsupervised learning, a deep learning model is handed a dataset without explicit instructions on what to do with it.
- The training dataset is a collection of examples without a specific desired outcome or correct answer.
- Depending on the problem at hand, the unsupervised learning model can organize the data in different ways:
  - Clustering
  - Anomaly Detection.
  - Association
  - Autoencoders
- Because there is no “ground truth” element to the data, it’s difficult to measure the accuracy of an algorithm trained with unsupervised learning.

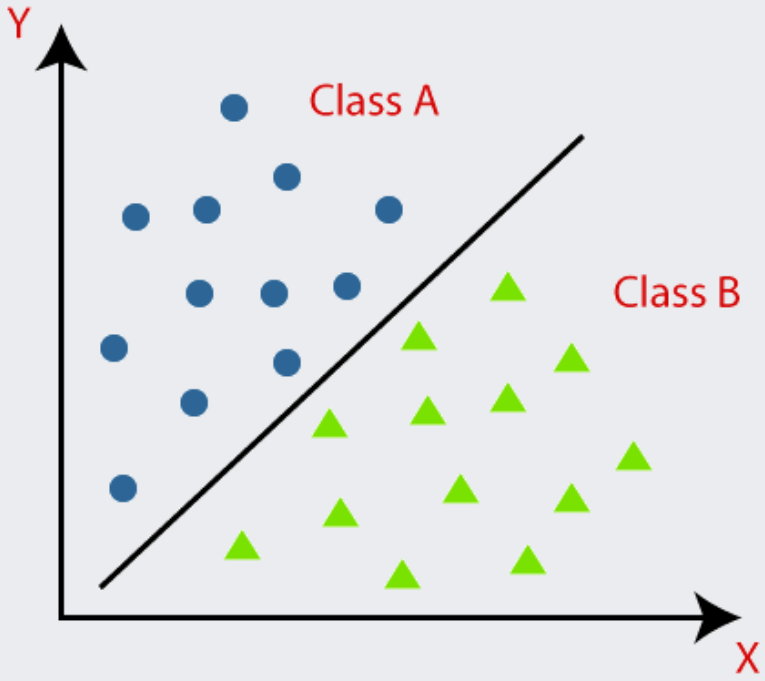


Difference between learning techniques.

	Supervised Learning	Unsupervised Learning
Input Data	Uses Known and Labeled Data as input	Uses Unknown Data as input
Computational Complexity	Very Complex	Less Computational Complexity
Real Time	Uses off-line analysis	Uses Real Time Analysis of Data
Number of Classes	Number of Classes are known	Number of Classes are not known
Accuracy of Results	Accurate and Reliable Results	Moderate Accurate and Reliable Results

What is classification learning?

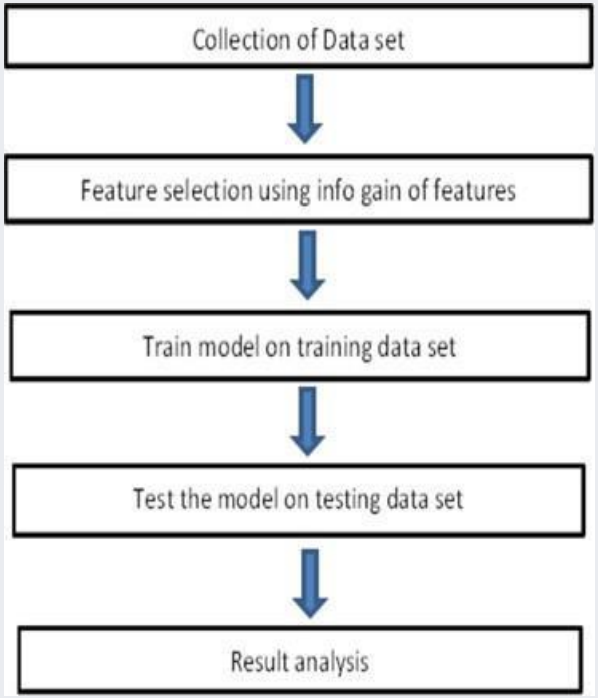
- In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.
- Classification predicts the category of the data it belongs to. Example: Spam detection, Gender, Sentiment analysis.
- From a modeling perspective, classification requires a training dataset with many examples of inputs and outputs from which to learn.
- A model will use the training dataset and will calculate how to best map examples of input data to specific class labels. As such, the training dataset must be sufficiently representative of the problem and have many examples of each class label.
- There are perhaps four main types of classification tasks that you may encounter; they are:
    - Binary Classification
    - Multi-Class Classification
    - Multi-Label Classification
    - Imbalanced Classification



Classification models used in this project.

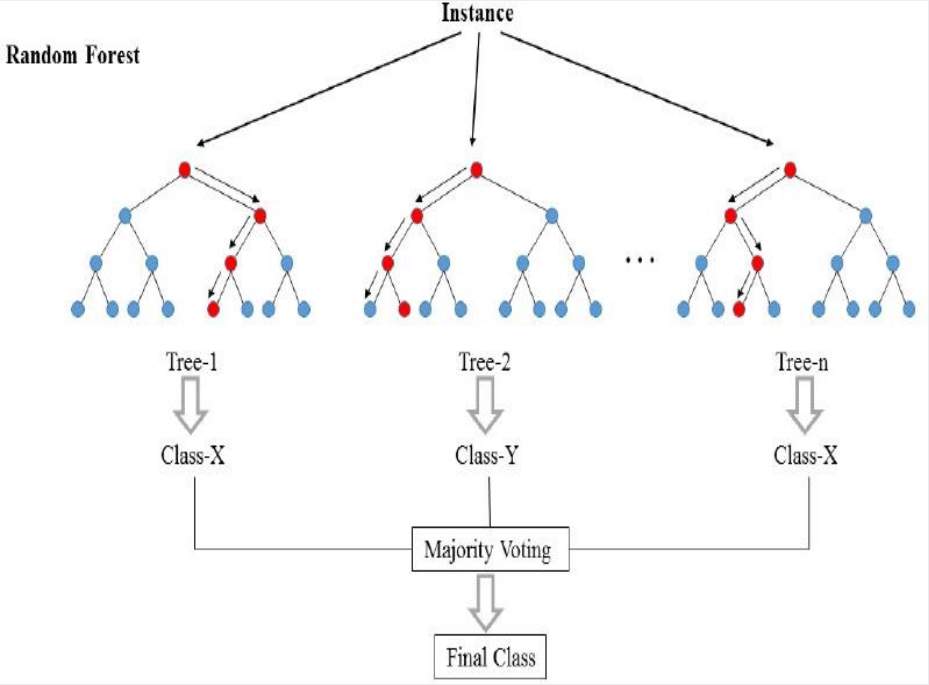
- **Logistic Regression:**
  Logistic Regression is a classification algorithm. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.
- **Random forest:**
  Random forest is a combination of bagging idea and random selection of features.
- **Decision Tree:**
  We can use tree-based algorithms for both regression and classification problems.
- **XGBoost:**
  XGBoost is an implementation of gradient boosted decision trees designed for speed and performance
- **Ridge Classifier:**
  The Ridge Classifier, based on Ridge regression method, converts the label data into [-1, 1] and solves the problem with regression method.
- **Adaboost:**
  This is a type of ensemble technique, where a number of weak learners are combined together to form a strong learner. Here, usually, each weak learner is developed as **decision stumps.**

Flowchart of the System:



Why was Random forest suitable algorithm?

- The random forest algorithm is based on supervised learning. It can be used for both regression and classification problems. As the name suggests Random Forest can be viewed as a collection of multiple decision trees algorithm with random sampling. This algorithm is made to eradicate the shortcomings of the Decision tree algorithm.
- Random forest is a combination of bagging idea and random selection of features. The idea is to make the prediction precise by taking average or mode of the output of multiple decision trees. The greater the number of decision trees is considered the more precise output will be.



Accuracy: Accuracy (ACC) measures the fraction of correct predictions.

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Accuracy of model using Random Forest-

```
0.7032380296245263
```

Confusion Matrix: It is an error matrix is a specific table layout that allows visualization of the performance of an algorithm.

Confusion Matrix of model using Random Forest-

```
[[14227  6198]
 [ 5863 14354]]
```

Precision: Precision calculates the ability of a classifier to not label a true negative observation as positive.

```
Precision=TP/(TP+FP)
```

Precision of model using Random Forest- 0.6984235110938108

Recall (Sensitivity): Recall calculates the ability of a classifier to find positive observations in the dataset.

$$Recall = TP/(TP+FN)$$

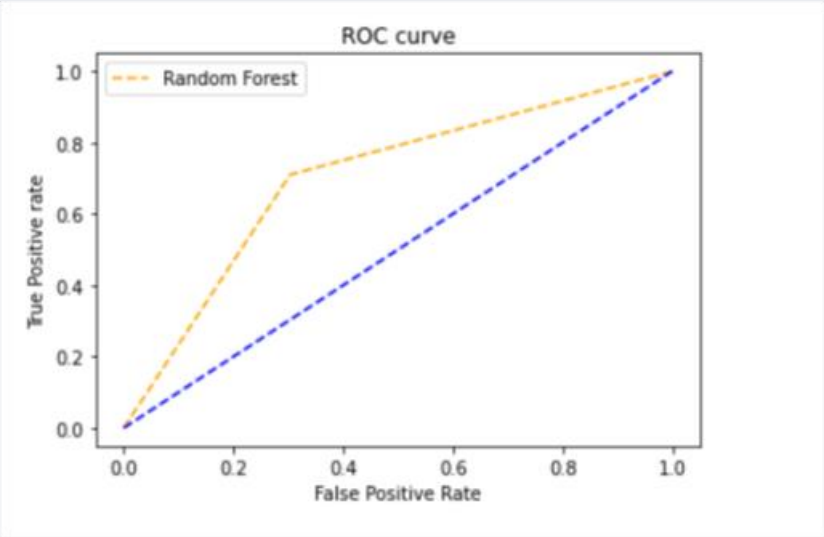
Recall of model using Random Forest- 0.7099965375673938

F1-Score: Harmonic mean of precision and recall.

$$\frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

F1-Score of model using Random Forest- 0.7041624763913757

ROC Curve of Random Forest-



Conclusion

The study analysed the vehicle accident risk for different weather conditions. According to the results the vehicle accident risks were higher if we consider condition of road with respect to weather i.e. in snow covered. For the precipitation the relative accident risk was higher for precipitation type i.e. during clear compared to the other precipitation type. If we consider light condition then during night chances of getting accident is higher than other light conditions. Applying various model we got best accuracy by using random Forest.

Future Scope.

The scope of Machine Learning is not limited to only predictions. We pose the vehicle accident risk prediction as a classification problem with two labels (accident and no accident).

There are several parameters we use to predict the model such as type of road i.e. one way, two ways, highway or bypass etc.

Among many other parameters we also take into consideration weather conditions like the chances of road being wet due to rain, dew or are the road dry.

Thanks!

Any questions?