iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44256

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Accident Data Analysis Using Machine Learning

Dr. Venkata. Koti Reddy. G[1], Bandaru Venkataramana[2], P. Bhasakarareddy[3], P. Ram Nivesh Reddy[4]
[1]*Ph.D Dept of CSE, Holy Mary Institute of Engineering and Technology, Keesara, TS, India*
[2] *PhD Scholar, Dept of CSE, Holy Mary Institute of Engineering and Technology, Keesara, TS, India*
[3]*Professor, Dept of ECE, Holy Mary Institute of Engineering and Technology, Keesara, TS, India*
[4]*Student, Dept of CSE, Holy Mary Institute of Technology and Science,Keesara, TS, India*

Abstract: ***The main objective of this project is toanalyze the road side accidents by scrutinizing accident-prone or hotspot areas and their root causes. Accidents through roadways have been a greatthreat to developed as well asunderdeveloped countries. Roadaccidents and its safety have been a major concern for the world, and everyone is trying to handle this since years. Road traffic and reckless drivingoccur in every part of the world. Because of this, many pedestrians are affected too. With no fault, they become victims. Many road accidents occur because of numerous factors likeatmospheric changes, sharp curves, andhuman faults. Injuries caused by road accidents are major but sometimes imperceptible, which later on affect health too. This study aims to analyze road accidents in one of the popular metropolitan cities, i.e., Bengaluru,through Linear Regression, Polynomial Regression, Decision Tree Regressor, Support Vector Regressor, Random Forest Regressor algorithms and machine learning by scrutinizing accident-prone or hotspot areas and their root causes.***

Keywords: Machine Learning Algorithms such as Random Forest, Super vector Machine, Linear Regression, Polynomial Regression, Decision Tree.

I. INTRODUCTION

The main objective of this project is to analyze the road side accidents byscrutinizing accident-prone or hotspot areas and their root causes. Accidents through roadways have been a great threat to developed as well as underdeveloped countries. Road accidents and its safety have been a major concern for the world, and everyone is trying to handle this since years. Road traffic and recklessdriving occur in every part of theworld. Because of this, manypedestrians are affected too. With nofault, they become victims. Many road accidents occur because ofnumerous factors like atmospheric changes, sharp curves, and humanfaults. Injuries caused by road accidents are major but sometimesimperceptible, which later on affect health too[1]. This study aims to analyze road accidents in one of the popular metropolitan cities, i.e., Bengaluru, through Linear Regression, Polynomial Regression, Decision Tree Regressor, Support Vector Regressor, Random Forest Regressor algorithms and machine learning by scrutinizing accident- prone or hotspot areas and their root causes[2].

II. SYSTEM STUDY

A. Related Work

In the area of road safety traditional statisticalmodel-based techniques were used to predictaccident fatal and severity. Mixed logit modeling approach, ordered Probit model, logit model are few of adopted conventional statistical-based studies. some studies believed the conventional statistical model better identify dependent and independentaccident factors. But conventional statistical-based approach lacks the capability to deal with multidimensional datasets. In order to combat traditional statistical models limitations; Nowadays many studies used ML approach due to its predictive supremacy, time consuming and informative dimension.In these decade ML approach employed occupational educational classification insurance[4].construction industry, accident,argiculutre , classification ,sentiment and banking and Neural Network (ANN), Convolution Neural Network (CNN) and Logistic Regression (LR) are in front to build accident severity model. Kwon et al. Adopted Nave Bayes (NB) and Decision Tree (DT) on California dataset collected from 2004 to 2010[4]. Authors used binary regression to compare the performance of the developed model but Nave Bayes were more sensitive to risk factors than the Decision Tree model[3].

B. Proposed System

The K-means clustering model produced a low accuracy. Using K-means there werequite a few wrong predictions which wrongly got detected as accident spots. Therefore, K-means would not be the preferred model as itdoesn't correctly analyze accidents and it also produced a lot of false positives.

C. Existing System

Data Analysization will be done by DataAnalyst or Data Scientist. Most of the data Scientists are using K-Means clutering model algorithm. For predicting the major occuring accident hotspot area's. And mostlyit is preferred by data scientist too.

The K-means clustering model produced a low accuracy. Using K-means there werequite a few wrong predictions, which wrongly got detected as Accident spots. Therefore, K-means would not be the preferred model, as it doesn't correctly Analyze Accidents and it also produced a lotof false positives Hence in this project we are using some of the machine learning algorithms they are Linear Regression, Polynomial Regression, Decision Tree Regression, Super Vector Regression, Random Forest Regression algorithms. Here we will choose the best algorithm to produce a good estimate of the generalizationerror and to be resistant to over fitting out of5 algorithms[3]. This algorithms has been found to produce a good accuracy and precision. By analysing or predicting from thesealgorithms we will be finding the algorithm which predicts and correct predictions then K-means algorithm. By this best shown regression value will be assumed as best algorithm to analysize the data. Hence that algorithm will correctly analyze accidents data where it was occurred.

III. METHODOLOGY

A. Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable[3]. The variableyou want to predict is called the dependent variable. The variable you are using to predictthe other variable's value is called theindependent variable[3].

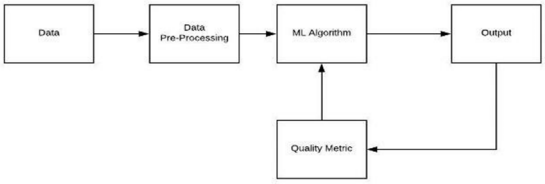


Fig 3.1 Linear Regression use-case diagram

B. Polynomial Regression

Polynomial Regression is a form of Linear regression known as a special case of Multiple linear regression which estimates the relationship as an nth degreepolynomial [3]. Polynomial Regression issensitive to outliers so the presence of one ortwo outliers can also badly affect the performance [3].

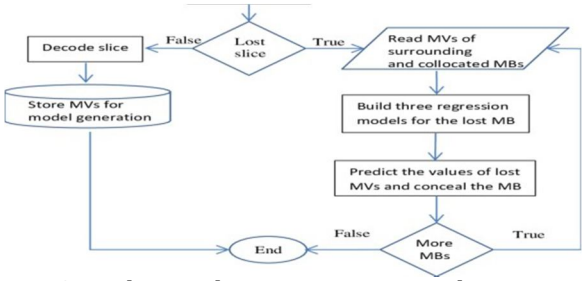


Fig 3.2 Polynomial Regression use-case diagram

C. Decision Tree

Decision tree builds regression orclassification models in the form of a tree structure[4]. It breaks down a dataset into smaller and smaller subsets while at the sametime an associated decision tree is incrementally developed. The final result is atree with decision nodes and leaf nodes[4].

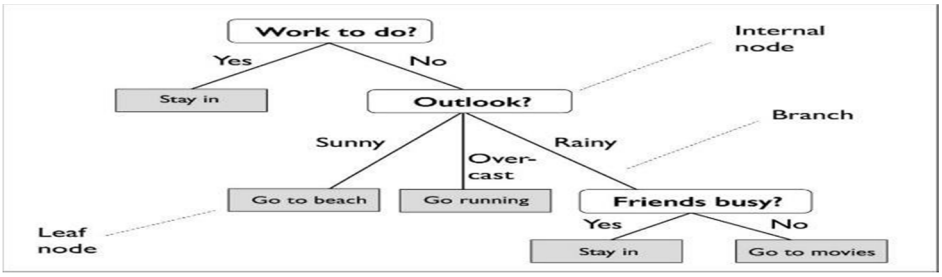


Fig 3.3 Decision Tree use-case diagram

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

D. Support Vector Machine

Supervised Machine Learning Models with associated learning algorithms that analyze data for classification and regression analysisare known as Support Vector Regression. SVR is built based on the concept of Support Vector Machine or SVM. It is one among the popular Machine Learning models that can be used in classification problems or assigning classes when the data is not linearlyseparable.

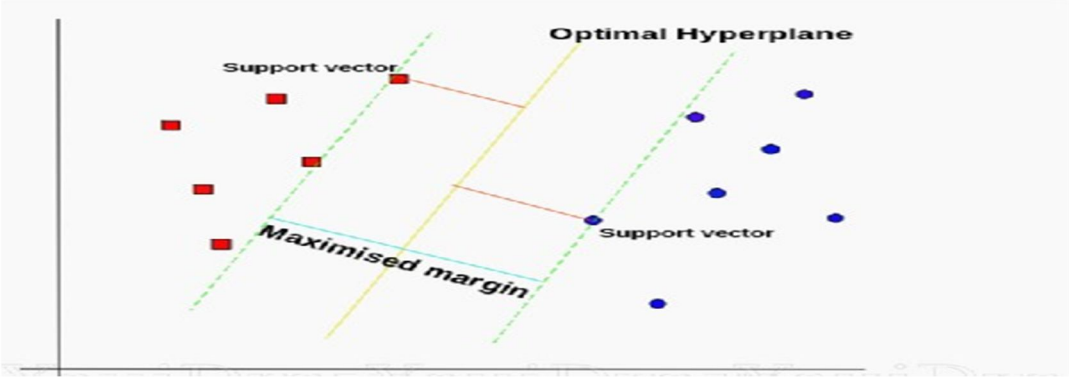


Fig 3.4 Support Vector Machine

E. Random Forest

Random Forest Regression is a supervisedlearning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combinespredictions from multiple machine learning algorithms to make a more accurate prediction than a single model[4].

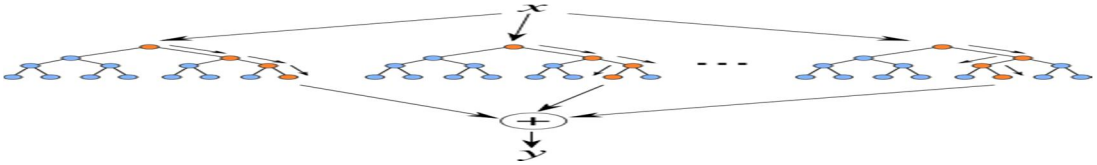


Fig 3.5 Random Forest Classification

IV. IMPLEMENTATION

A. Data Collection and Analysis

The data was collected from external websites like kaggale.com for this project. And during the project data analysis we have installed some python modules that are required they are

1) *PANDAS* - Pandas is a Python library for data analysis and it is most widely used in machine learning tasks.
2) *NUMPY* – It stands for 'Numerical Python' module, it can be utilised to perform numberof mathematical operations on arrays such astrigonometric, statistical and algebraic routines.
3) *MATPLOTLIB[6]* – It is a python library which is used for data visualization[6].
4) *SEABORN[6]* - It is a python library which is used for making statistical graphs[6].

B. Data Cleaning

Datasets are collected from different manually recorded materials. Some of the data values are incomplete, noisy andinconsistent. Real world data set is not feasible for analysis or to make efficient decisions. Real world datasets are most of the time contains unintentionally 'dirty data.' In case of machine learning datasets shouldbe in a proper format to get a classification model. Further to get quality output from the classification model the input must be quality. Prior to analysis dataset should be preprocessed intensively to get valuable trends from the historical data.

V. CONCLUSION

The principal aim of our project is to choice the best machine learning regression algorithms out of 5 algorithms which helps the data scientist to analyze the the data efficiently and easily. By the end of the project we can conclude that we can use 1 out 5 of the best algorithm to analyze the data.

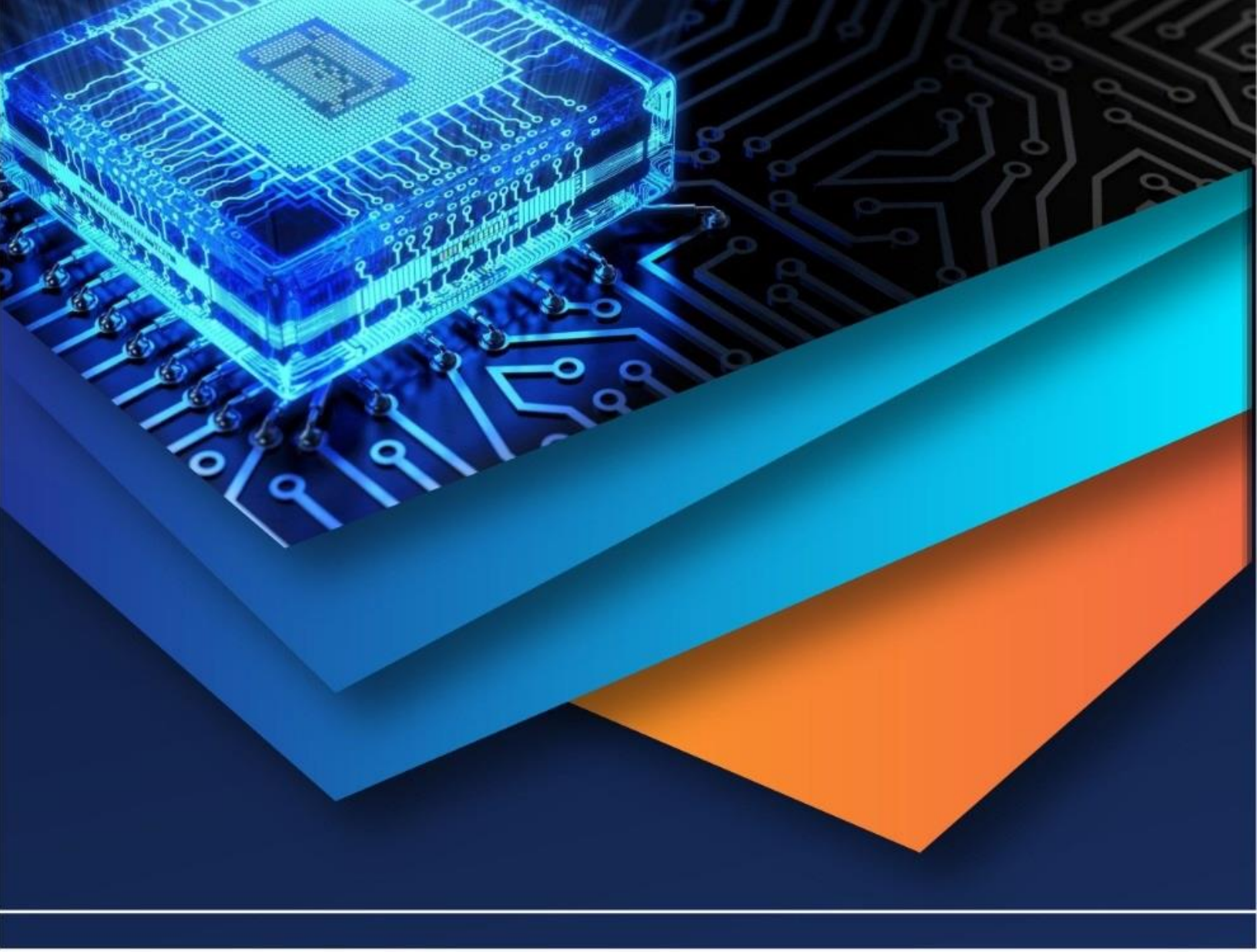
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

VI.FUTURE SCOPE

The main goal of this project is to decrease the number of accidents rate.
For this goal it is necessary to analyse the accidents rate first.
This helps the data scientist to use their timeefficiently as data scientists invest most of their time in data analysing.

REFERENCES

[1] Python Crash Course, Eric Matthes
[2] Head-First Python, Paul Barry
[3] Classification and Regression Treesby Leo Breiman, Jerome Friedman
[4] Decision trees, discriminant analysis,logistic regression, svm, ensamble methods and knn with matlab.
[5] The Hanford Plaintiffs, Trisha T. Pritikin, Richard C. Eymann
[6] Python Data Science Essentials - Third Edition by Alberto Boschetti, Luca Massaron



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY