

**PUNE INSTITUTE OF COMPUTER TECHNOLOGY,  
DHANKAWADI PUNE-43.**

***Report On***  
***Data Analytics Mini Project***

**Analysis of road accident using machine learning algorithms**

**SUBMITTED BY**

<b>Name</b>	<b>Roll No.</b>
Rutuja Adkine	4102
Sammed Admuthe	4103
Ahmed Khwaja	4104
Shameem Ajare	4105

**CLASS: BE-1**

**GUIDED BY**

**Prof. A. A. Chandorkar**



**COMPUTER ENGINEERING DEPARTMENT**  
**Academic Year: 2018-19**

PUNE INSTITUTE OF COMPUTER TECHNOLOGY,  
DHANKAWADI PUNE-43.

## ***CERTIFICATE***



This is to certify that

Ms. ***Rutuja Adkine***, Roll No. **4102**,

Ms. ***Sammed Admuthe***, Roll No. **4103**,

Mr. ***Ahmed Khwaja***, Roll No. **4104**,

Ms. ***Shameem Ajare***, Roll No. **4105**

students of B.E. (Computer Engineering Department) Batch 2018-2019 , has satisfactorily completed a report on “**Analysis of Road Accidents using Machine Learning algorithms**” under the guidance of ***Prof.A.A.Chandorkar*** towards the partial fulfillment of the fourth year Computer Engineering Semester I of Pune University.

Prof. A.A.Chandorkar  
**Internal Guide**

Dr. R.B.Ingle  
**Head of Department,  
Computer Engineering**

**Date:**

**Place:**

**Abstract:**

*The traffic has been transformed into the difficult structure in points of designing and managing by the reason of increasing number of vehicle. This situation has discovered road accidents problem, influenced public health and country economy and done the studies on solution of the problem. This work analyzes the road accidents using previous data sets and classification algorithms namely decision tree, KNN . Performance measures used are accuracy and standard deviation. Based on the performance accuracy we can determine the best algorithm amongst the used ones.*

**Keywords:**

*Machine learning,classification algorithms:Decision Trees ,KNN , Road accident dataset.*

# Analysis of Road Accidents using Machine Learning algorithms

## Contents

<b>1 INTRODUCTION</b>	<b>5</b>
1.2 History . . . . .	5
<b>2 LITERATURE SURVEY</b>	<b>6</b>
2.1 Analysis of road accidents in India using data mining classification algorithms	7
2.2 Analysis for status of the road accident occurrence and determination of the risk of accident by machine learning in istanbul . . . . .	7
<b>3 METHODOLOGY</b>	<b>7</b>
3.1 ALGORITHMS USED . . . . .	7
3.1.1 Decision trees . . . . .	7
3.1.2 KNN . . . . .	8
3.2 CHALLENGES . . . . .	8
3.2.1 Data Preprocessing . . . . .	8
3.2.2 Re-Sampling of data . . . . .	8
3.2.3 Need large dataset . . . . .	8
3.3 DESIGN AND ANALYSIS OF SYSTEM . . . . .	10
<b>4 DISCUSSION ON IMPLEMENTATION RESULTS</b>	<b>11</b>
<b>5 CONCLUSION AND FUTURE ENHANCEMENT</b>	<b>12</b>
5.1 Conclusion . . . . .	12
5.2 Future Enhancements . . . . .	13

## List of Figures

1 KNN . . . . .	8
2 Flow Chart of implementation . . . . .	10
3 PCA analysis . . . . .	11
4 KNN score( $X_{label}$ : no.of epochs, $Y_{label}$ : score) . . . . .	12

## List of Tables

1	Literature survey . . . . .	6
---	-----------------------------	---

# 1 INTRODUCTION

One of the most complicated and difficult daily needs is overland transportation. Studies on traffic have executed that road accidents and death-laceration ratio will increase. Design and control of traffic by advanced systems come in view as the important need. Assumptions on the risks in traffic and the regulations and interventions in the end of these assumptions will reduce the road accidents. An assumption system which will be prepared with available data and new risks will be advantageous. Machine learning which is sub-branch of artificial intelligence supplies learning of computer taking advantage of data warehouses. Assumption abilities of computer systems have advanced in the event of machine learning. Utilization of machine learning is a widespread and functional method for taking authentic decisions by using experience. Machine learning is able to attain extract information from data and use statistical methods. Classification algorithms in Machine learning analyzes the data and classifies the data into a predefined set of classes. Here we classified data into three categories: Serious, Fatal, Slight according to its effect on the casualty. Thus it is a multi level classification. There are one or more machine learning algorithms are available out of which decision tree, KNN are used here. The main aim of this work is to analyze the features that are more prone for occurrence of an accident. The performance measures are accuracy rate and standard deviation. This research paper shows that : Road Surface\_Wet or Damp, Road Surface\_Dry, Age of Casualty are the three most important features.

## 1.1 Motivation

The number of people undergoing injuries also deaths is increasing day by day. It has really become difficult to manage the traffic due to increasing number of vehicles, increasing population and people always rushing to work. A person can face death even due to a minor mistake.

Machine learning is a very overwhelming field which can help us to achieve our goal of reducing the number of road accidents. One of the way is to finding out the major factors which are responsible for causing an accident. And taking extra care about them.

The motivation behind our research is the understanding of specific conditions that affect the severity of an automotive accident. The purpose of this project is to highlight impactful variables while operating a vehicle in order to improve accident prevention.

## 1.2 History

History of this work includes working with the old versions of classification algorithms like old versions of decision trees and KNN. They have low accuracy as compared to today's more

developed algorithms.

Again if we face any complex or huge data set it's difficult to achieve required results with accuracy we achieve today.

Other old way of dealing it would be manually finding out the fatal features causing accident with help of human brain. But again this task is very complex and if we have large data set then it is nearly impossible.

Thus using Machine learning algorithms gives the more appropriate result in more efficient way.

## 2 LITERATURE SURVEY

sr.no	Paper Title	Author	Machine learning algorithms used
1.	Analysis of road accidents in India using data mining classification algorithms	Ms. E. Suganya Dr. S. Vijayarani	Linear regression, logistic regression, SVM, Naive bayes, KNN, decision trees, random forest, gradient boosting
2	Analysis for status of the road accident occurrence and determination of the risk of accident by machine learning in istanbul	Halil İbrahim BÜLBÜL Tari k KAYA	Decision tree, Statistical classification algorithm, Naive bayes

Table 1: Literature survey

## **2.1 Analysis of road accidents in India using data mining classification algorithms**

This paper analyzes the road accidents in Indian data set using classification algorithms namely linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm. And achieves precision of about 90.5% using KNN as the highest precision.

## **2.2 Analysis for status of the road accident occurrence and determination of the risk of accident by machine learning in istanbul**

In this paper, assignment of the most compatible machine learning classification techniques for road accidents estimation by data mining has been intended.

# **3 METHODOLOGY**

## **3.1 ALGORITHMS USED**

### **3.1.1 Decision trees**

A decision tree is a decision support tool that uses a tree like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision trees often mimic the human level thinking so its so simple to understand the data and make some good interpretations. Decision trees actually make you see the logic for the data to



interpret

### 3.1.2 KNN

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure like distance measures for example euclidean measure, minkowski measure, manhattan measure. In the following diagram for a model, if we consider nearest neighbours  $k=3$  then the data set belongs to class B. But if  $k=6$ , then data set belongs to class A.

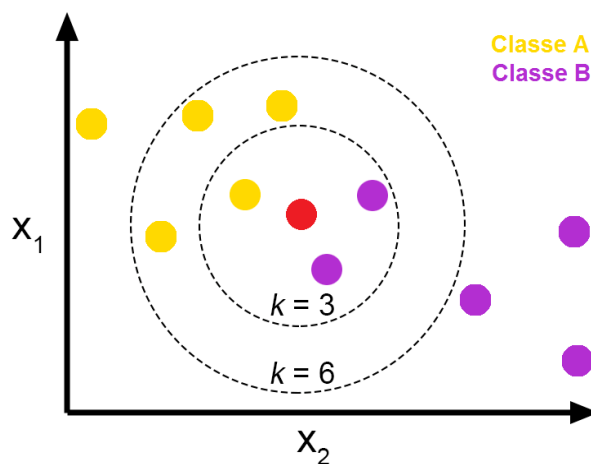


Figure 1: KNN

## 3.2 CHALLENGES

### 3.2.1 Data Preprocessing

Data preprocessing was one of the major challenges. Dropping of undesirable columns, creating dummy variables, converting the features was needed to be carried out.

### 3.2.2 Re-Sampling of data

Re-sampling the data set was also quite tedious as oversampling and undersampling of data set were to be performed.

### 3.2.3 Need large dataset

As every Machine learning problem this problem also requires large amount of dataset. It was

a daunting task to find dataset. Fortunately data set for this problem statement could be found from online resources

### 3.3 DESIGN AND ANALYSIS OF SYSTEM

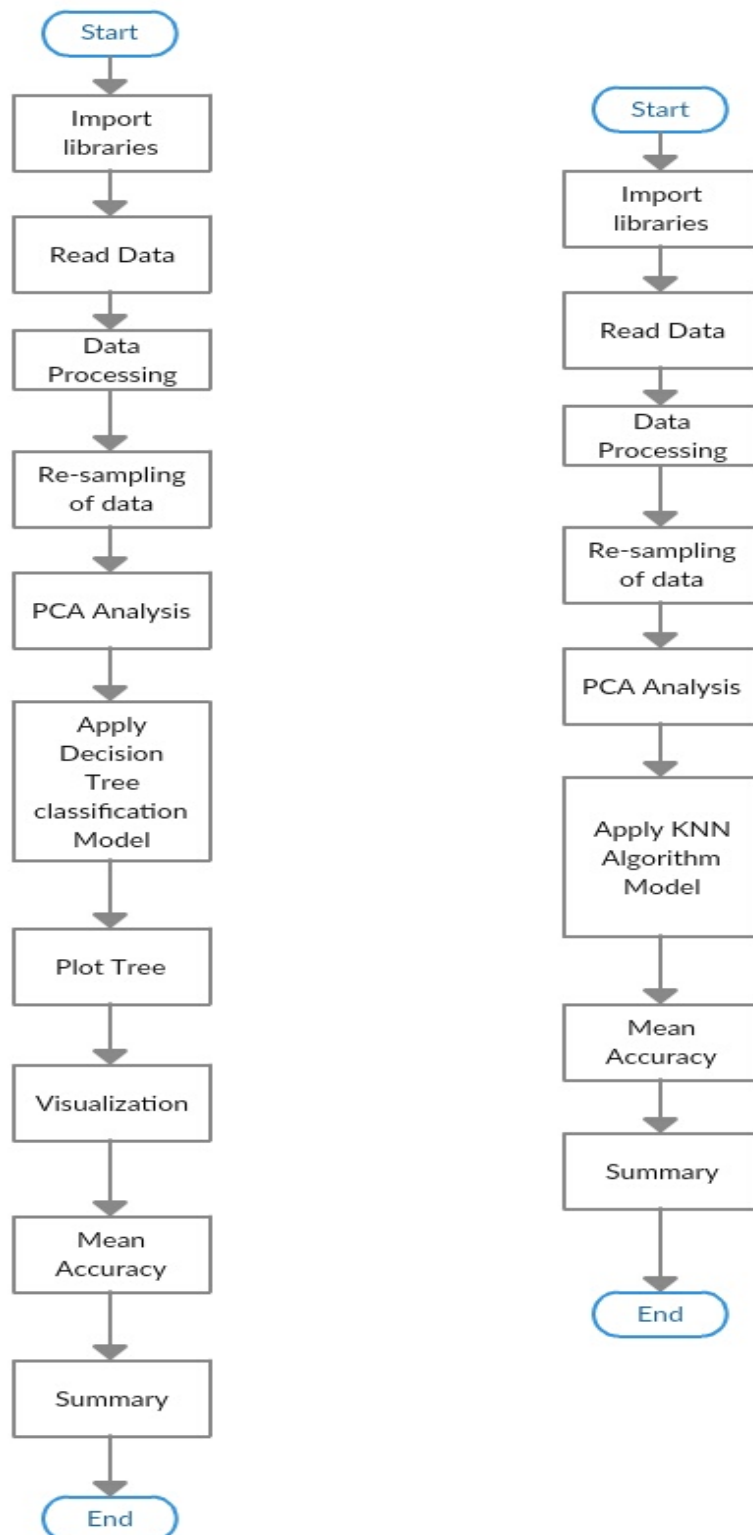


Figure 2: Flow Chart of implementation

## 4 DISCUSSION ON IMPLEMENTATION RESULTS

Decision trees and KNN models were implemented in python using scikit-learn library.

For data visualization, matplotlib library was used.

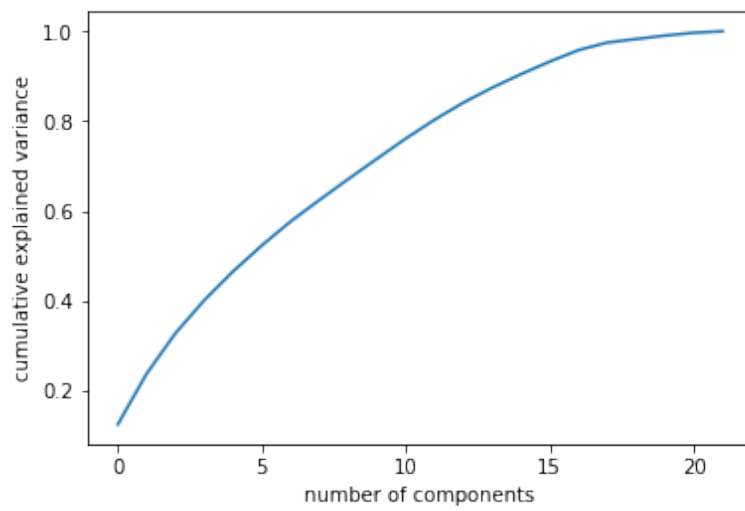


Figure 3: PCA analysis

The score for decision tree model for 40 epoch is shown below, its mean accuracy is 70.29% .

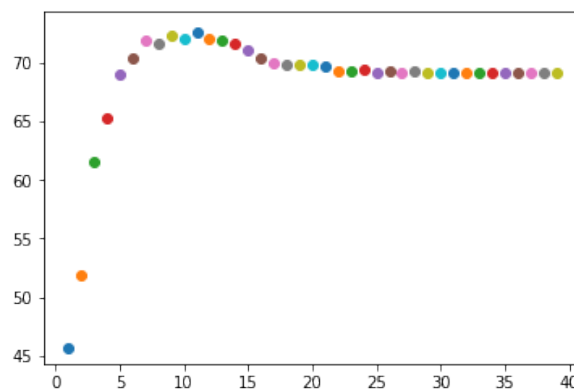


Figure 4: Decision tree score

The score for KNN model for 40 epochs is shown below, its mean accuracy is 72.73% .

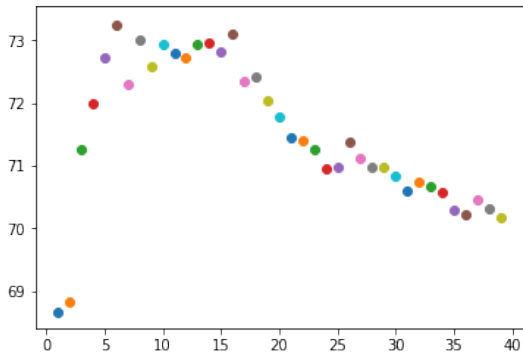


Figure 4: KNN score( $X_{label}$  : *no.of epochs*,  $Y_{label}$  : *score*)

## 5 CONCLUSION AND FUTURE ENHANCEMENT

### 5.1 Conclusion

Hence using machine learning algorithms we can predict the most important factors which cause accidents. To do so, we followed common steps to solve any task with machine learning:

- Load and preprocess data
- Analyse patterns in data
- Train different models
- Interprete the trained model

## 5.2 Future Enhancements

Although results are satisfactory there a big scope for improvement in this problem.

- Training on large dataset can improve accuracy and model can be generalized.
- Data loss can be minimized using more efficient architecture.

## References

- [1] Ms. E. Suganya ,Dr. S. Vijayarani, “ Analysis of road accidents in India using data mining classification algorithms” , Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017), IEEE Xplore ,2017.
- [2] Halil İbrahim BÜLBÜL,Tarık KAYA,Yusuf TULGAR,” Analysis for status of the road accident occurrence and determination of the risk of accident by machine learning in istanbul” ,2016 15th IEEE International Conference on Machine Learning and Applications.
- [3] <https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1>
- [4] <https://medium.com/@chiragsehra42/k-nearest-neighbors-explained-easily-c26706aa5c7f>