International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 758-767 www.ijaem.net ISSN: 2395-5252

Predicting Road Crash Using Ensemble Learning

Mrs. J.Sanchana[1] Mrs.R.Mohanabharathi[2]
[1]PG scholar [2]Assistant Professor[1,2]Department of Computer Science Engineering[1,2]SelvamCollegeofTechnology,Namakkal, India

Date of Submission: 12-01-2023 Date of Acceptance: 24-01-2023

ABSTRACT— The number of vehicles increasing on the road inthe recent years which leads to increase in the number ofaccidents. Accident prediction and prevention is the majorchallenge faced by the government / transport department.TheObjectiveofthissystemistodevelopa machinelearningmodelforreal-timeaccidentforecastingbycomparing supervised algorithms with mean value of votingclassifier results. Recent technologies like automated trafficcontrol signals and IOT based GPS Technology helps inpreventingaccidentsontheroad.TheMachinelearningalgorithms has been implemented to predict the occurrenceofaccidentsontheroad.Ensemblelearning methodisoneof the best method for accident forecasting and finding thebestroadselection.Thissystem isproposedtocompareensemblelearningalgorithmwithotheralgorithmslikesupervisedmachinelearningalgorithmsuchaslogisticregression, decision tree, random forest, and support vectorclassifier, K nearest neighbor and Naive Bayes. EnsembleLearning produces better predict performance compared to asinglemodel.InEnsemblelearningtechnique,various models will be combined and the best prediction result willbe found. The comparative analysis helps to prove that theensemblelearningalgorithmprovideshighaccuracyofresults than other model. The voting classifier method inensemblelearninghelpstodocomparativeanalysisandthere by forecast accident and to find the best road. Datasetof previous accident reports available in government websitehas been used as input data to find the best road and topredictthe accidents.

Keywords: MachineLearning,EnsembleLearning,PredictionofAccuracy

I. INTRODUCTION

Machine learning is to predict the future from past data.Machine learning (ML) is a typeof artificial intelligence(AI)thatprovidescomputerswiththeabilitytolearnwithoutbeingexplicitlyprogrammed.Machinelearningfocuses on the development of Computer Programs that canchange when exposed to new data and the basics of MachineLearning,implementationofasimplemachinelearningalgorithm using python. Process of training and predictioninvolves use of specialized algorithms. It feed the trainingdatatoanalgorithm,andthealgorithmusesthistrainingdata to give predictions on a new test data. Machine learningcan be roughly separated in to three categories. There aresupervisedlearning,unsupervisedlearningandreinforcement learning. Supervised learning program is bothgiven the input data and the corresponding labeling to learndatamustbelabeledbyahumanbeingbeforehand. Unsupervisedlearningisnolabels.Itprovidedtothelearningalgorithm.Thisalgorithmmustfigureouttheclusteringoftheinputdata.Finally,Reinforcementlearning dynamicallyinteractswithitsenvironmentanditreceivespositiveornegativefeedbackto improveitsperformance.

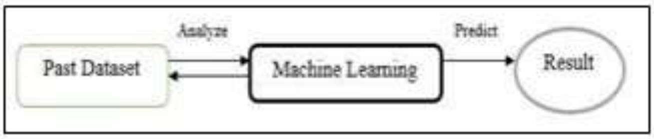


Fig.1:ProcessofMachinelearningmethod

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 758-767 www.ijaem.net ISSN: 2395-5252

Roadandtrafficaccidentsareunsureandundeterminable incidents and their valuation calls for theexpertiseofthefactorsaffectingthem.Roadandtrafficaccidents are defined by a set of variables which can beusuallyofdiscretenature.Theessentialdifficultyintheanalysis of coincidence records is its heterogeneous nature.Classificationismachinelearningtechniquethatcanbeusedasaninitialtasktoobtainvariousgoalsandtheclassificationcategorizetheaccidentdataintodifferentcategories.

SequentialEnsemblelearning(Boosting):Boosting is a machine learning ensemble meta-algorithm forprincipallyreducingbias,andfurthermorevarianceinsupervisedlearning,andagroupofmachinelearningalgorithms that convert weak learner to string ones. Boostingisameta-algorithmwhichcanbeviewedasamodelaveragingmethod.Itisthemostwidelyusedensemblemethod and one of the most powerful learning ideas. Thismethod was originally designed for classification but it canalsobeprofitablyextendedtoregression.Theoriginalboostingalgorithmcombinedthreeweaklearnerstogenerateastronglearnerandsequentialensemblemethodswherethebaselearnersaregeneratedsequentially.

ParallelEnsembleLearning(Bagging):Baggingisamachinelearningensemblemeta-algorithmintendedtoimprovethestrengthandaccuracyofmachinelearningalgorithms used in classification and regression purpose. Itadditionally diminishesfluctuation of data(variance)andhelp to from over-fitting. Bagging or Bootstrap Aggregationis a powerful, effective and simple ensemble method. Themethod uses multiple versions of a training set by using thebootstrap,i.e. samplingwithreplacementand tit can beused with any type of model for classification or regression.Bagging is only effective when using unstable (i.e. a smallchange in the training set can cause a significant change inthe model) non-linear models and parallel ensemble methodswherethebaselearnersaregeneratedinparallel.

StackingandBlending:Stacking isawayofcombiningmultiplemodelsthatintroducestheconceptofa Metalearner. Itislesswidelyusedthanbaggingandboosting.Unlikebaggingandboosting,stackingmaybeusedto combine models of different types.Stacking is concernedwithcombiningmultipleclassifiersgeneratedbyusingdifferentlearningalgorithmsonasingledatasetwhichconsistsofpairsoffeaturevectorsandtheirclassifications.

This technique consists of basically two phases, in the firstphase, a set of base-level classifiers is generated and in thesecondphase,ameta-levelclassifierislearnedwhichcombinestheoutputsofthebase-levelclassifiers.Blending is technique where we can do weighted averagingoffinalresult.

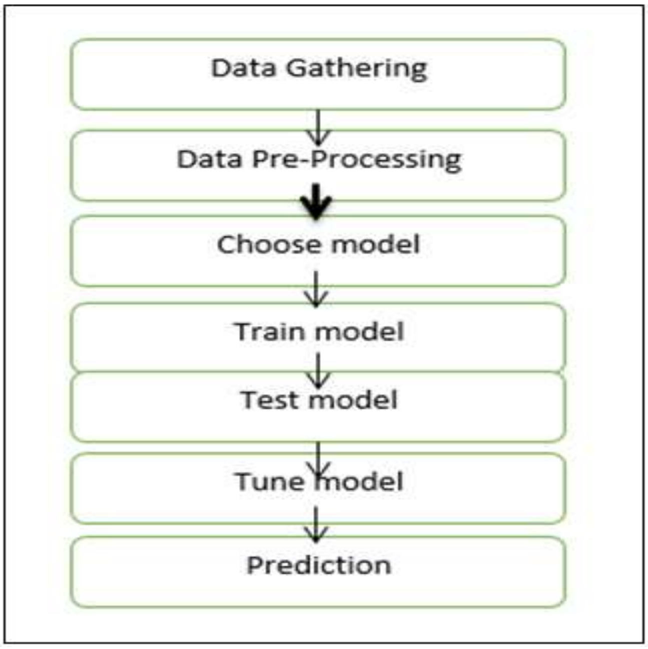


Fig.2:ProcessofDataflowdiagram

Machine learning needs data gathering have lot ofpast data's. Datagathering have sufficient historical dataand raw data. Before data pre-processing, raw data can't beuseddirectly.It's usedtopreprocess then,whatkind ofalgorithmwithmodel.Trainingandtestingthismodel workingandpredictingcorrectlywithminimumerrors. Tuned model involved by tuned time to time with improvingtheaccuracy.

II. SYSTEMMODEL

Ensemble learning helps improve machine learning resultsbycombiningseveralmodels.Thisapproachallowstheproduction of better predictive performance compared to asingle model and it is the art of combining diverse set oflearners together to improvise on the stability and predictivepower of the model. In the world of Statistics and MachineLearning, Ensemble learning techniques attempt to make theperformance of the predictive models better by improvingtheir accuracy. Ensemble Learning is a process using whichmultiplemachinelearningmodelsarestrategicallyconstructed tosolve a problem.

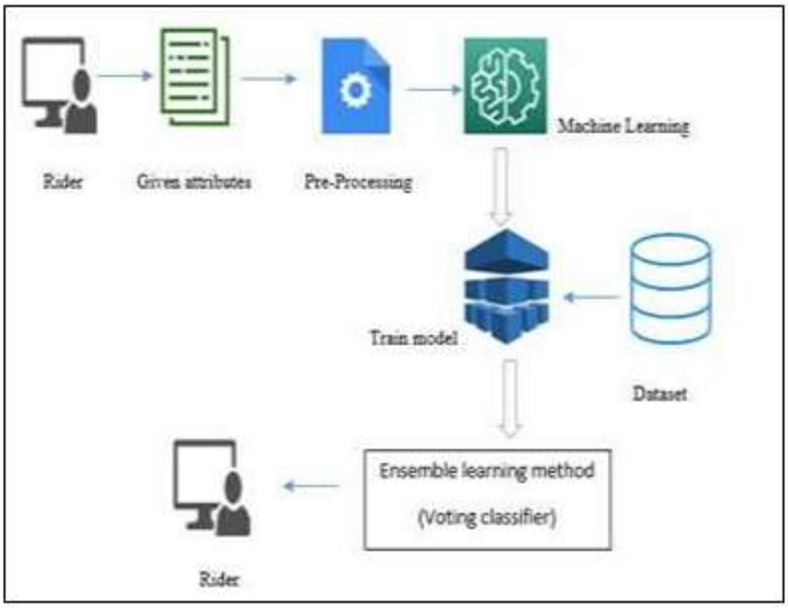


Fig.3:Systemarchitecturediagram

Max Voting: The max voting method is generallyused for classification problems. In this technique, multiplemodels are used to make predictions for each data point. Thepredictions by each model are considered as a 'vote' and thepredictions which we get from most of the models are usedasthefinalprediction.
Averaging: Similar to the max voting technique,multiplepredictionsaremadeforeachdatapointinaveraging. In this method, we take an average of predictionsfrom all the models and use it to make the final prediction.Averaging can be used for making predictions in regressionproblems or while calculating probabilities for classificationproblems.
WeightedAverage:Thisisanextensionoftheaveraging method. All models are assigned different weightsdefiningtheimportanceofeach modelforprediction.

A.PreparingtheDataset
The dataset is now supplied to machine learning model onthe basis of this data set themodel is trained. Every newdatadetails filledat the timeof application form acts asatestdata set.

Variable	Description
Accidentoccurs_Date	Dateofaccidentoccurs
Light_Cond	Roadlightcondition
Weathconds	Placeofweatherconditions
Mod	Vehicledriver orpedestrian
Age	Ageofdrivers

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 758-767 www.ijaem.net ISSN: 2395-5252

Vehicle_type	Typeofvehicle
Route1,2,3	Multitrafficroutes
Locations	Placeofaccident
cctv_footag	videosisonoroffcondition

Table1: Detailsofgivendataset

III. METHODOLOGY

A. DataValidationandPreprocessing

Validation techniquesin machinelearning are usedto getthe error rate of the Machine Learning (ML) model, whichcanbeconsideredasclosetothetrueerrorrateofthedataset.Ifthedatavolumeislargeenoughtoberepresentativeofthepopulation,youmaynotneedthevalidationtechniques.However,inreal-worldscenarios,to workwithsamplesofdatathatmaynotbeatruerepresentative of the population of given dataset. To findingthe missing value, duplicate value and description of datatypewhether itisfloat variableor integer.

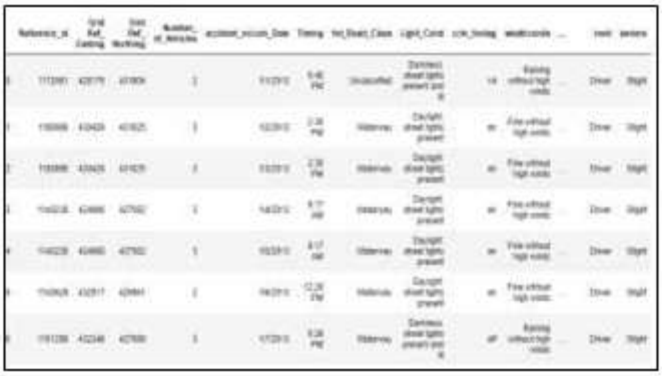


Fig.4:Givendataframe

Importing the library packages with loading givendataset.Toanalyzingthevariableidentificationbydatashape, data type and evaluating the missing values, duplicatevalues.A validation datasetisasampleof data heldbackfrom training your model that is used to give an estimate ofmodel skill while tuning models and procedures that you canuse to make the best use of validation and test datasets whenevaluatingyourmodels.Datacleaning/preparing byrenamethegivendatasetanddropthecolumnetc.toanalyze the uni-variate, bi-variate and multi-variate process.The steps and techniques for data cleaning will vary fromdataset to dataset. The primary goal of data cleaning is todetect and remove errors and anomalies to increase the valueofdata inanalyticsanddecisionmaking.

Pre-processing refers to the transformations appliedtoourdatabeforefeedingittothealgorithm.DataPreprocessing is a technique that is used to convert the rawdata into a clean data set. In other words, whenever the dataisgatheredfromdifferentsourcesitiscollectedinrawformat which is not feasible for the analysis. To achievingbetter results from the applied model in Machine Learningmethod ofthedata hastobeinapropermanner.

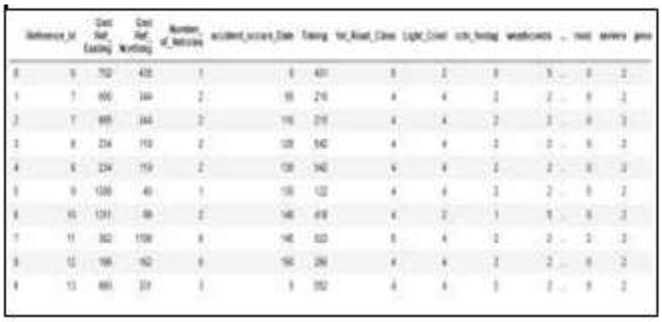


Fig.5:Pre-processeddataframe

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 758-767 www.ijaem.net ISSN: 2395-5252

B. ToTrainAModel ofVisualizationbyGiveAttributes

Data visualization is an important skill in applied statisticsandmachinelearning.Statisticsdoesindeedfocusonquantitativedescriptionsandestimationsofdata. Datavisualization provides an important suite of tools for gainingaqualitativeunderstanding.Thiscanbehelpful whenexploring and getting to know a dataset and can help withidentifying patterns, corrupt data, outliers, and much more.With a little domain knowledge, data visualizations can beusedtoexpressanddemonstratekeyrelationshipsinplots andchartsthataremorevisceralandstakeholdersthanmeasures of association or significance. Data visualizationand exploratory data analysis are whole fields themselvesand it will recommend a deeper dive into some the booksmentioned atthe end.

DataVisualizationAftertheclassificationandregressionprocessthepredictedresultsarevisualizedingraphical or tabular format for better understanding of theusers.Wecanalsogetthesummaryoftheresultsinnumerical format. Sometimes data does not make sense untilit can look at in a visual form, such as with charts and plots.Being able to quickly visualize of data samples and others isan important skill both in applied statistics and in appliedmachine learning. It will discover the many types of plotsthatyouwill needtoknowwhenvisualizingdatainPython.

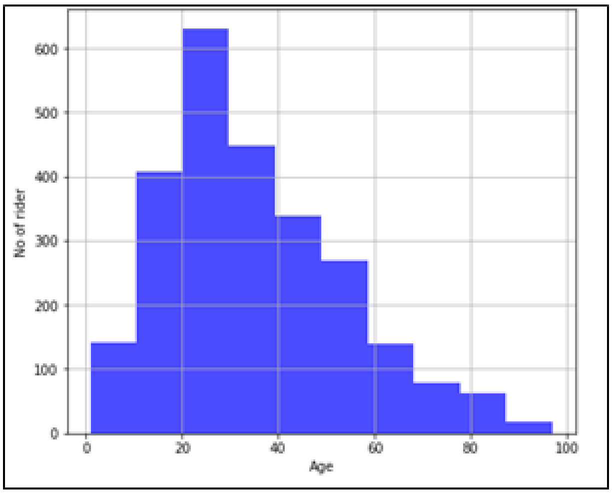


Fig.6:Ageofdistributionofeachriders

Evenbeforepredictivemodelsarepreparedontrainingdata,outlierscanresultinmisleadingrepresentationsandinturnmisleadinginterpretationsofcollecteddata.Outlierscanskewthesummarydistributionof attributevalues in descriptivestatistics likemean andstandarddeviationandinplotssuchashistogramsandscatterplots,compressingthebodyofthedata.Finally, outliers can represent examples of data instances that arerelevant to the problem such as anomalies in the case offraud detection and computersecurity.

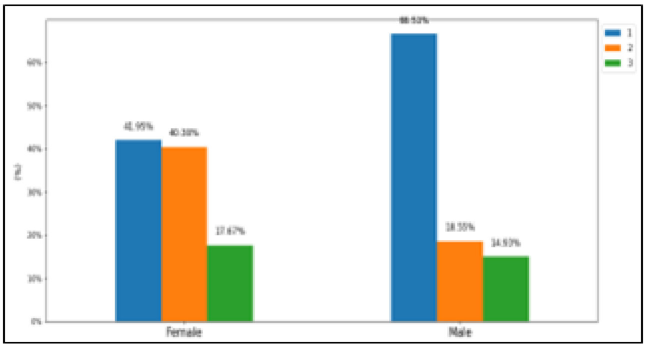


Fig.7: Accidentpredictionbyroutes

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 758-767 www.ijaem.net ISSN: 2395-5252

It couldn't fit the model on the training data andcan't say that the model will work accurately for the realdata. For this, we must assure that our model got the correctpatterns from the data, and it is not getting up too muchnoise.Cross-validationisatechniqueinwhichwetrainour model using the subset of the data-set and then evaluateusingthe complementarysubset ofthedata-set.

C. ComparisonofMachineLearning AccuracyResults

Itisimportanttocomparetheperformanceof multipledifferentmachinelearningalgorithmsconsistentlyanditwill discover to create a test harness to compare multipledifferent machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your ownmachinelearningproblemsandaddmoreanddifferentalgorithmstocompare.Eachmodelwillhavedifferentperformance characteristics. Using resampling methods likecross validation, you can get an estimate for how accurateeach model may be on unseen data. It needs to be able to usethese estimates to choose one or two best models from thesuite of models that you have created. When have a newdataset, it is a good idea to visualize the data using differenttechniquesinordertolookatthedatafromdifferentperspectives. The same idea applies to model selection. Youshould use a number of different ways of looking at theestimated accuracy of your machine learning algorithms inorderto choose the oneor twotofinalize.Away to dothis is to use different visualization methods to show theaverageaccuracy,varianceandotherpropertiesofthedistributionofmodel accuracies.

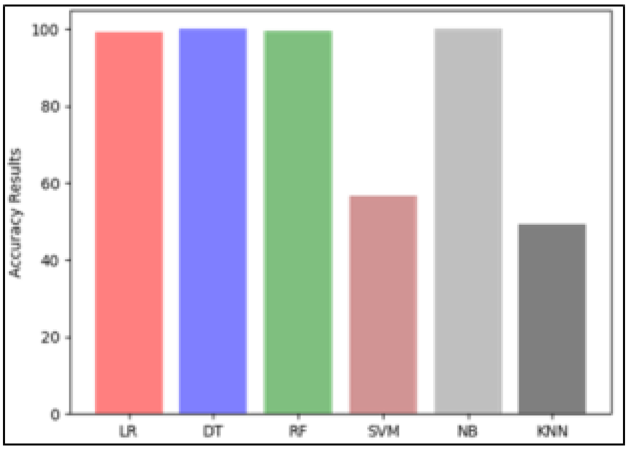


Fig.8:ComparisonofMLaccuracies

D. ImplementationofVotingClassifierAlgorithm

Voting is one of the most straightforward Ensemble learningtechniques in which predictions from multiple models arecombined.Themethodstartswithcreatingtwoormoreseparate models with the same dataset. Then a Voting basedEnsemble model can be used to wrap the previous modelsand aggregatethe predictionsofthosemodels.

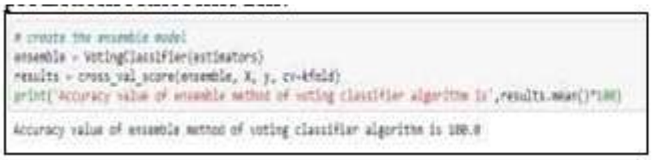


Fig.9: Accuracyresultofvotingclassifier

AftertheVotingbasedEnsemblemodelisconstructed, it can be used to make a prediction on new data.The predictions made by the sub-models can be assignedweights. Stacked aggregation is a technique which can beused to learn how to weigh these predictions in the bestpossibleway.Inthefieldofmachinelearningandspecificallytheproblemofstatisticalclassification,a confusionmatrix,alsoknownasanerrormatrix.Aconfusion matrix is a table that is often used to describe

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 758-767 www.ijaem.net ISSN: 2395-5252

theperformance of a classification of ensemble voting classifiermodel on a set of test data for which the true values areknown. It allows easy identification of confusion betweenclassesofaccidentoccurred and notoccurred accident.

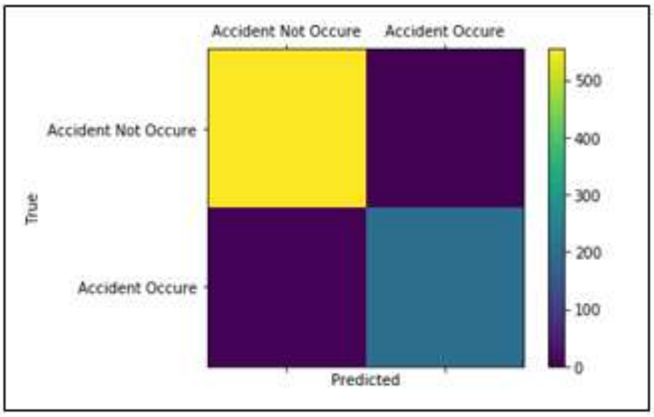


Fig.10:Confusion matrixof votingclassifier

Aconfusionmatrixisasummaryofprediction resultsonaclassificationproblemandthenumberofaccident not occurred and accident occurred predictions aresummarized with count values and broken down by eachclass. The confusion matrix shows the ways in which yourclassification model is confusedwhen itmakes predictions.It gives us insight not only into the errors being made by aclassifier but more importantly the types of errors that arebeing made.

DefinitionoftheTerms:

- Positive(P):Observationispositive(Accidentnot occurred).
- Negative(N):Observationisnotpositive(Accidentoccurred).
- TruePositive(TP):Observationispositive,andispredicted tobe positive.
- FalseNegative(FN):Observationispositive,butis predicted negative.
- TrueNegative(TN):Observationisnegative,andispredicted tobe negative.
- FalsePositive(FP):Observationisnegative,butis predicted positive.

To predicting the probability of a binary outcome isthe Receiver Operating Characteristic curve, or ROC curveand it summarize the trade-off between the true positive rateand false positive rate for a predictive model using differentprobability thresholds.Precision-Recall curvessummarizethe trade-off between the true positive rate and the positivepredictivevalueforapredictivemodelusingdifferentprobability thresholds. ROC curves are appropriate when theobservationsarebalancedbetweeneachclass,whereasprecision-recallcurvesareappropriateforimbalanceddatasets.

F.Performanceof EnsembleLearning Method

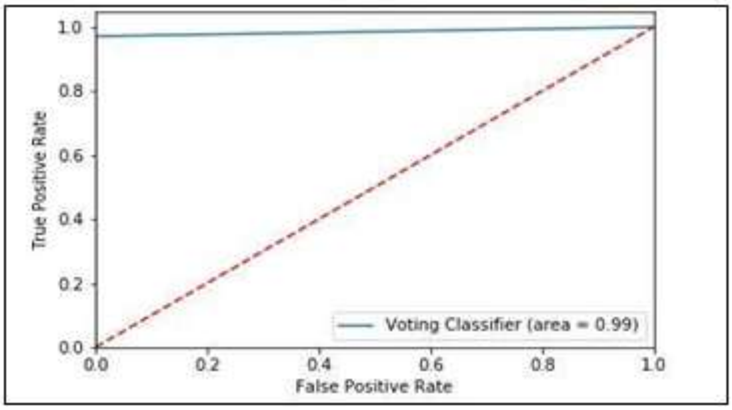


Fig.11:ROCforvotingclassifieralgorithm

E. PerformanceofMachineLearningParameters

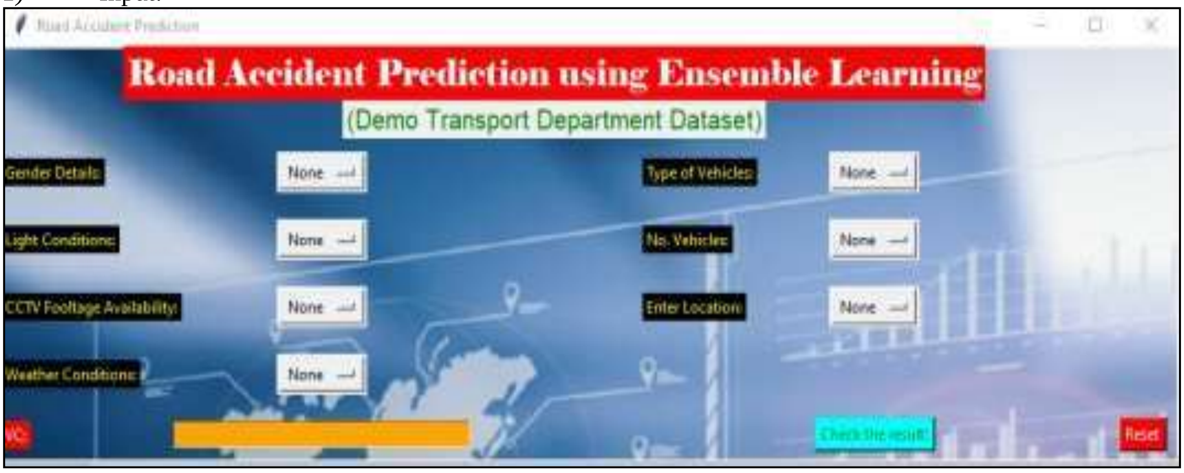
Parameter	LR	DT	RF	SVM	NB	KNN
Precision	0.9	1	1	0	1	0.5
Recall	1	1	1	1	1	0.7
F1-Score	1	1	1	0.7	1	0.6
Sensitivity	1	1	1	1	1	0.7
Specificity	0.9	1	1	0	1	0.2

F.Performanceof EnsembleLearning Method

Parameter	VotingClassifier
Precision	0.99
Recall	1
F1-Score	0.99
Sensitivity	1
Specificity	0.97
TP	200
TN	554
FP	0
FN	6
TPR	0.97
TNR	1
FPR	0
FNR	0.02
PPV	1
NPV	0.98
Accuracy	100

G.TestingResults

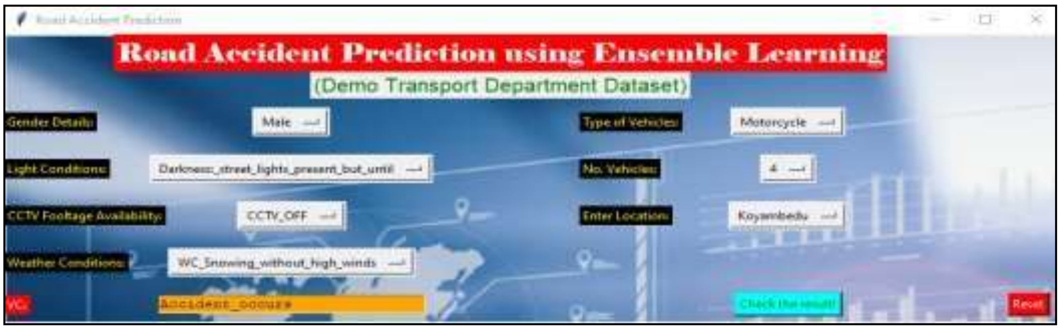
1) Input:



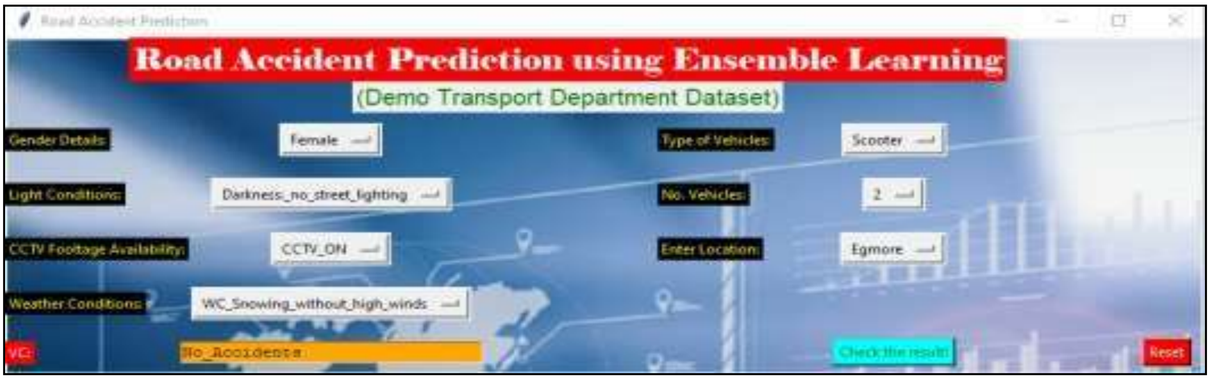
International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 1 Jan. 2023, pp: 758-767 www.ijaem.net ISSN: 2395-5252

2) Output:

Test-01:



Test-02:



IV. CONCLUSION

Theimprovedaccuracyandimplementationismakethe proposed method to help the transport department make adiagnosisbeforetheaccidentsandtheaccuracy resultisvotingclassifieralgorithmbycomparingsuper visedmachinelearningmethod.

V. FUTURE WORK

In future, I would like to discover the automate this processby show the prediction result in web application or desktopapplicationandtooptimizetheworktoimplem entinArtificialIntelligence environment.

REFERENCES

[1] G.ParathasarathyandT.RSoumya(2019),"Usi nghybridDataMiningalgorithmforAnalysingr oadaccidents Data Set", IEEEInternational Conference onComputing and Communication Technologies, 978-1-5386-9371-1.

[2] Noorishta Hashmi and Dr. M. Akheela Khanum (2019),"PreventingRoadAccidentsbyAnalysi ngSpeed,Driving Pattern and Drowsiness Using Deep Learning",InternationalJournalofComputatio nalEngineeringResearch,Vol. 09Iss. 7,ISSN 2250-3005.

[3] Tariq Abdullah and Symon Nyalugwe (2019), "A DataMiningApproachforAnalyzingRoadTraf ficAccidents",UniversityofDerby,IEEE978-1-7281-0108-8.

[4] ThanyawanChanpanitandNarongsakArkama nont(2019),"PredictingtheNumberofPeoplef orRoadTraffic Accident on Highways by Hour of Day", IEEE8th International Conference on Industrial Technologyand Management, 78-1-7281-3268-6.

[5] Christine Maria Sunny, Nithya S and Sinshi k S (2018),"ForecastingofRoadAccidentinKerala :ACaseStudy",IEEECenterforExcellenceinD ataEngineeringand Computational Modeling.

[6] FabioGalatioto,MarioCatalanoandNabeelSha ikh(2018),"Advancedaccidentpredictionmod elsandimpactsassessment",TheInstitutionofE ngineeringand Technology, Vol.12Iss.9, pp.1131-1141.

[7] Ruimin Li, Francisco C. Pereira and Moshe E. Ben-Akiva(2018),"Overview oftraffic incident durationanalysis and prediction", European Transport Research,12544-018-0300-1.

[8] S.PriyaandR.Agalya(2018),"AssociationRul eMiningApproachtoAnalyzeRoadAccidentD

ata",IEEEInternationalConferenceonCurrent TrendstowardConvergingTechnologies,978-1-5386-4349-5.

[9] Suraj D and Sandeep Kumar S (2018), "A Survey onAnalyses of Factors Related to Road Accidents UsingDataMiningTechniques",InternationalJournalofEngineeringDevelopmentandResearch,Volume6,Issue1, ISSN: 2321-9939.

[10] Tadesse Kebede Bahiru and Prof. Dheeraj Kumar Singh(2018),"ComparativeStudyonDataMiningClassification Algorithms for Predicting Road TrafficAccidentSeverity",IEEEProceedingsofthe2nd International Conference on Inventive CommunicationandComputationalTechnologies,978-1-5386-1974-2.

[11] Yongbeom Lee, Eungi Cho and MingyuPark (2018),"AMachineLearningApproachtoPredictionofPassengerInjuriesonRealRoadSituation",IEEEInternationalconferenceonsoftcomputingandInternational systems and International symposium onadvancedintelligentsystems, 978-1-5386-2633-7.

[12] C.Sugetha,L.KarunyaandE.Prabhavathi(2017),"Performance Evaluation of Classifiers for Analysis ofRoadAccidents",IEEEInternationalConferenceonAdvancedComputing(ICoAC),978-1-5386-4349-5.

[13] Gagandeep Kaur and Harpreet Kaur (2017), "PredictionofTheCauseOfAccidentAndAccidentProneLocation On Roads Using Data Mining Techniques",IEEE.

[14] Lu Wenqi, Luo Dongyu and Yan Menghua (2017), "AModelofTrafficAccidentPredictionBasedonConvolutionalNeuralNetwork",IEEEInternationalConference on Intelligent Transportation Engineering,978-1-5090-6273-7.

[15] Prajakta S. kasbe and Apeksha V. Sakhare (2017), "AReview On Road Accident Data Analysis Using DataMining Techniques", IEEE International Conference onInnovationsininformationEmbeddedandCommunicationSystems,978-1-5-90-3294-5