ISSN 2455 – 9091

IJCSET MARCH Volume 9 Issue 3

International Journal of Computer Science Engineering and Technology (IJCSET)

https://www.doi.org/ 10.5281/zenodo.8402152

An Effective Diabetes Prediction System using Random Forest Algorithm

[1]P. Krishna Reddy
[1]Assistant Professor,
[1] Department of Computer Science Engineering,
[1] St. Martins Engineering College, Secunderabad, Telangana.

ABSTRACT: Diabetes is a prevalent metabolic disorderthat affects a significant number of people globally.Timelydetectionandtreatmentofdiabetescanprevent complications and improve health outcomes.Thehealthcareindustryisfacinganincreasing demand for better patient care and disease predictionsystems.ThisstudyproposesaDiseasePredictionSystem that integrates various features, including anAIChatBot,DiabetesPredictionSystem,Chat andAppointmentBookingSystem,toimprovediseaseprediction accuracy. The Random Forest algorithm isutilizedintheDiabetesPredictionSystem,whichenhancestheoverallaccuracyofthesystem.Withmultipleinputs,thesystembecomesproficientinaccurately classifying diseases and predicting outputs.The system's accuracy was evaluated using a patientinformationdataset,resultinginanoverallaccuracyof90.4%.TheseresultsdemonstratetheDiseasePrediction System's potential to improve healthcareoutcomes by providingtimely and accurate diseaseprediction.Inconclusion,thisstudy'sproposedsystemhasthepotentialtosignificantlybenefithealthcareproviders and the medical field. With its high diseasepredictionaccuracy,efficientdiseaseclassification,anduser-friendlyfeatures,thissystemcanassisthealthcare professionals in making precise diagnoses,providing effective treatments, and enhancing patientoutcomes.

Keyword: [Diabetes Prediction, Machine Learning, Manual Information, Random Forest, Decision Trees, Adaptive Boosting.]

1. INTRODUCTION

The condition or disease which is permanent to thehumanbodyformorethanthreemonthsisknownasa chronic condition, Thereare different types ofchronicdiseasessuchascancer.Alzheimer's.Arthritis. Asthma.Heartdisease.andDiabetes.These types of diseases force people to change thedailylifestyleoftheaffectedpeople.Diabetesoccurs when the level of blood sugar is too high.Diabetesoccurswhenenoughinsulincan'tbeproduced by the pancreas, or the produced insulincan't be utilized by our body [1]. Insulin regulatesbloodsugarinthehumanbody.Uncontrolleddiabetes occurs due to the abnormal rise of sugar inthe blood. In the long run, it increases the chance ofsevere damage to organs such as nerves and bloodvessels. As a result, this disease increases the perilof several fatal diseases [1]. According to WHO(World Health Organization), diabetes is of threetypes:(i)Juvenilediabetes(Type1)thatoccurswhen enough insulin can't be produced by the body[2].Generally,childrenandyoungadultsareaffectedbyit[2].(ii)TypeIIdiabeteshappenswhenthe body can't utilize insulin effectively. Middle-aged people who are more than 45 years old aremostlyaffectedbyTypeIfdiabetes[3].Butnowadays, it is also developed in children and youngadults. At present, 95% of all diabetes is of Type II[3]. (iii) Gestational diabetes (Type III) occurs dueto a high glucose level in the blood. Usually, it isdiagnosed in women during pregnancy and had noexperienceofdiabetesbefore.Accordingtothelatestreport releasedbyWHO,totaldiabeticpatientshaveincreasedfrom108millionto422millionfrom1980to2014[1].Accordingto WHO,diabetesmellituswillbetheseventhleadingcauseofmortality by 2030 [1]. A recent study has stated that642 million young adults will be affected by diabetesby 2040 [3]. In 2016, diabetes directly affected thedeathsofapproximately1.6millionpeople[1].Unfortunately,thisdisease cannotbeeradicated. Butwecancontrolitbyrestrainingtheglucoselevelinblood. Whendiabetesisdetected,itseffectcanbeminimized. But this is not an easy task. To identifythedisease,dataaretakenfrompatientslikeinsulin,age, body mass index, family history of diabetes, etc.and then consulted to a doctor [4]. Then the doctordecidesusing his/ herknowledgeandexperience. Butthisidentificationprocessisverytime consuming and sometimes most costly. Sometimes, italsomisleads thediagnosisprocessduetothelackof experience of the doctors. Computer automateddiagnosis can playan important characterinthedetection of diabetic patients. A lot of research hasbeen done to identify diabetic patients at an earlystage.Severalmedicaldatasetshavebeendevelopedto accelerate the research in this field. Due to thenon-linear, non-normal, and complex nature of themedical data, classification of the diabetic patientsaccurately is a challenging task for the researchers.That'swhyMachineLearning(ML)anddeeplearningmethodsarewidelyutilizedtoextractvaluable knowledge from the dataset and predictdiabetesdisease.Variousmachinelearningmethodshavebeenappliedtoclassifydiabeticpatients.Theirclassification accuracy was not significant enoughbecause most of them ignored the importance ofhandlingmissingvalues,removalofirrelevantfeatures , and handling outliers. While interpretingthe medical datasets, classification accuracy mostlydepends on the pre-processing of the data. Afterconsidering all of these factors, the classification ofdiabetic patients remains an arduous task for theearly diagnosisof patients[5].

Thisstudyemphasizesonhowtoimprovetheclassificatio naccuracy for the earlydiagnosis ofdiabeticpatients.Toclassifydiabeticpatients,aTree-Basedpredictionmodelhasbeenproposedbasedonmach inelearningtechniques.Atfirst,missingvaluesare handled bytheir groupmeanvalue.Afterhandlingthemissingvalues.outli ersaredetectedandhandledusingtheRobustscalingnorm alization technique. An oversampling methodisadoptedtoavoidtheimbalancedclassdistributi onproblem.Thenaneffectivefeatureselectiontechnique is applied to build a Tree-Based predictionmodelforearlydiagnosisof diabetespatients.

2. LITERATUREREVIEW

Over the last few years, various researchers haveworkedonpredictingdiabetesusingdifferentmachi ne learning approaches. Jeevan, Rajesh andVijay.[1]employedthreeclassifiers,namelyDecisio n Tree, Support Vector Machine, and NaiveBayes, to detect diabetes at an early stage using thePimaIndiansDiabetesDataset(PIDD).Theirresultsi ndicated that Naive Bayes outperformed the othertwo classifierswithan accuracyof81.21%.

AnotherstudybyVignanaJyothi.[2]usednormalization andrandomforesttoclassifydiabetes patients.Theypre-processedthedatausingnormalization and achieved an accuracy of 80.6%.Inadifferentstudy,ElifNur,BelgiumErkalandTu lin[3]utilizedK-Nearest NeighborandNaiveBayesclassifierstopredictdiabetes. Theirmodelachieved an accuracy of 84% on the PIDD dataset.Inasimilarwork,Vinay,Vasu,Shreya,JayandBh ushan[4]usedK-foldcross-validationtopredictdiabetes.Theycomparedtheperfor manceoffourclassifiers,namelyK-NearestNeighbor,DecisionTree,Naive Bayes,and SupportVectorMachine.TheirresultsshowedthatK-NearestNeighborperformedthebestwithan accuracy of82.68%.

Inconclusion,thesestudiesshowthatmachinelearning algorithms can accurately predict diabetes,and different algorithms and techniques can be usedto obtainhighaccuracyrates.

3. PROPOSEDMETHODOLOGY

Dataset Description

ThePimaIndiansDiabetesDataset(PIDD)wasutilizedin thisresearch,obtainedfromtheUniversityofCalifornia,I rvine(UCI)machinelearningrepository.Thedatasetcom prises768instances, with a majority of female subjects at 21years of age. There are two classes in the datasetbased on the binary class attribute. A value of '0'indicatesnodiabetes(-ve),while'1'denotesdiabetes(+ve).Thedatasetcontainse ightindependentattributes, with 500 instances (65.1%) indicating nodiabetes (-ve) and 268 instances (34.1%) indicatingdiabetes(+ve).

The details of the PIDD dataset is given in Table

S. N.	Name ofAttribute	AttributeDescri ption	Percentageof MissingValues
1	Pregnancies	No. of timespregnant	0
2	Glucose	Glucoseconcent rationfor 120mins	0.65%
3	BloodPress ure	Diastolic bloodpressure	4.55%
4	SkinThickn ess	Skin foldthickness	29.55%
5	Insulin	Serum-insulin(2hours)	48.69%
6	BMI	Body massindex	1.43%
7	PedigreeFu nction	Diabetespedigr eefunction	0
8	Age	Ageinyears	0
9	Class	1fordiabetes positiveand0	0

TABLEI -1 DESCRIPTIONOFPIDDDATASET

Data Pre-processing

Real-world medical data is often complex in natureand may contain missing values, inconsistent data,and outliers, making it non-linear and non-normal.Datapre-processingplaysacrucialroleintheperformance of any classification algorithm. In thisresearch,ourobjectiveistoincreasetheaccuracyofdiabeti c patient prediction, and thus we have takenstepstohandlethedataefficiently.Datapre-processingtechniquesutilizedinthisresearchincludemissin gvalueimputation,removalofirrelevant features, outlier detection, oversampling,andfeaturescaling.Tofacilitatethesesteps,we haveproposedaflowgraphmodel,illustrated inFig.1.

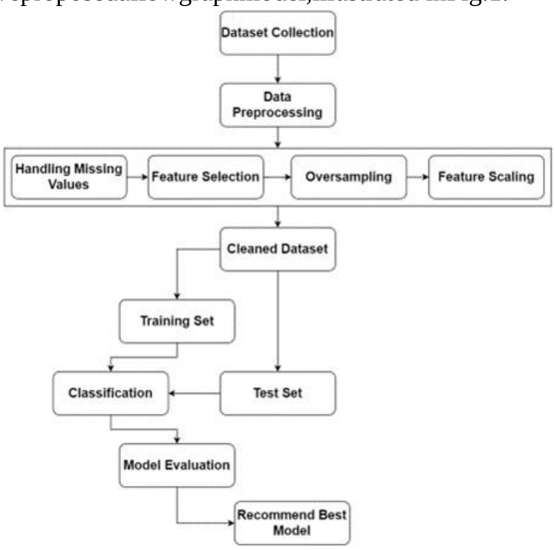

Figure.1 WorkflowofProposedModel

1. Missing Values Imputation

ThePimaIndiansDiabetesDatasetcontainsmissingvaluesfo rglucose(0.65%),bloodpressure(4.55%),skinthickness(29. 55%),insulin(48.69%),andBMI(1.43%).Tohandlethesemi ssingvalues,meanimputationwasperformed, wherebythemissingvaluesforeachattributewerefilled in with themean valueofthatattribute.

Eliminating Outliers

Outliersinadatasetcansignificantlyaffectthecalculated parameters; therefore, it is important toidentifyand remove the mfromthe data. In thisresearch,theInterQuartileRange(IQR)methodwasusedtodetectandeliminatetheoutliersinthedataset. TheIQRmethodidentifiesoutliersbyfirstcalculatingthe firstquartile(Q1)andthethirdquartile(Q3)points.TheIQRisfurthercalculatedasQ3-Q1. Thenormaldatarangeisdefinedwithalowerlimitof Q1-1.5IQRandanupperlimitofQ3+1.5IQR.Toensureaccurateanalysis,anydatapointthatfallsoutsideofthespecifiedrangeisdeemedanoutlierandshould beexcludedfrom further analysis. TheconceptofquartilesandIQRcanbebestvisualized using a boxplot, where the minimum andmaximumpointsare definedas Q1-1.5IQRandQ3+1.5IQR,respectively, and any point outside this range is considered anoutlier. It is important to note that extreme datapointsdonotalwaysnecessarilymeantheyareoutliers.

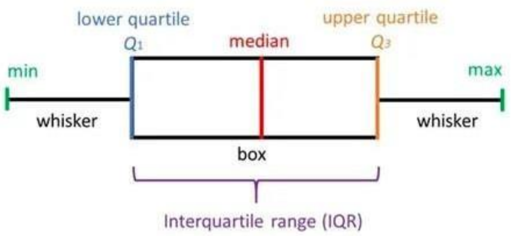

Figure. 2 BoxplotshowingQuartiledistribution

Oversampling

Oversampling is a technique used to address class imbalance in a dataset. Class imbalance occurswhenone class has significantly more samplesthantheother(s), whichcanresultinbiasedclassificationmodelsthatfavourthemajorityclass. Over sampling involvesr and omlyduplicating samples from the minority class until the datasetis balanced. This technique ensures that the classification model is not biased towards the majority class.

Feature Scaling

Feature scaling is the process of standardizing therange of independent feature values in a dataset. Itisimportanttoscalefeaturesbeforecreatinga machine learning model because some algorithmsare sensitive to the scale ofthe input features.Feature scaling ensures that the features are on asimilarscale,whichhelpsthemodelconvergefasterandreducestheimpactofoutliers.TheRobust Scaler is a technique used to scale the databy using the interquartile range, which makes itrobustto thepresenceofoutliersin thedataset.

Classification

DecisionTree(DT)

ThePimaIndiansDiabetesDatasethasbeenclassified using Decision Tree (DT), a supervisedMLtechniquethatconstructsamodelforclassification and regression in the structure of atree.DThasbeenwidelyusedfortheclassificationofdisease diagnosisasitiseasilyexplainable[16].Moreover, it requires less data cleaning as outliersand missing values have less significance on themodel'sdata.InDT,theclassispredictedbasedondecision rules taken from input data. It splits thedataintosubsetsofdataandrepresentsthosesubsets in a tree structure wherein each node adecisionismade.Thefinalclassificationisextracted fromleafnodes.

RandomForest(RF)

Random Forest is an ensemble ML algorithm thatgenerates numerous classification models (decisiontrees)whereeachmodelisconstructedusingafeatureselectorsuchasGiniIndex,InformationGain,andGainRatio.Thesemodelslearnandmakecontributions to the prediction in a discrete manner.Thefinalresultismadefromthoseobtainedpredictions.

AdaptiveBoosting(AB)

Adaptive Boosting or AdaBoost integrates manyweakclassifierstogenerateastrongclassifier.AdaBoost sets weights to each weak classifier andensurescorrectclassificationbytrainingthesampledata in each iteration while predicting outliers orunusualobservations.Theintuitionbehindthisclassification technique is that a single classifiercanaccuratelypredictaportionofthedatasetgivingincorrectresultsforotherportions,butincorrect portions canbe correctly predicted byother weak classifiers. The combination of weakclassifiersisrepresentedby

$f(x)=\sum_{t=1}^{T}\alpha_t h_t(x)$ (1)

Where $h_t(x)$ is a weak classifier. The result of thefinalclassifierisH(x) = sign(f(x)).

Evaluation Measures

Toevaluatetheeffectivenessofourproposedmodel,weutilizedK-foldcross-validationtechniquestorandomlysplitthedatasetintoksubsetstoconstructthetrainingandtestsets.During each iteration, k-1 subsets were used to trainthemodel,whiletheremainingsubsetwasusedfortesting.Byrepeatingthisprocessktimes,weobtainedthemodel'sperformancebyaveragingthetest results of the independent k subsets. In thisstudy, we used 10-fold cross-validation to reducebiasandvariance.Variousstatisticalmeasures,such as F1-score, precision, recall, and ReceiverOperating Characteristic Curve (ROC AUC), wereconsidered to evaluate the model's performance.Typically,theconfusionmatrixsummarizestheoverall performance of any prediction model, andaccuracy, F1-score, recall, and precision can bederivedfromit.Fromtheconfusionmatrixaccuracy,F1score,recall,andprecisionareintended asfollows.

Actual	Predicted	
	Positive	Negative
Positive	TruePositive(TP)	FalseNegative(FN)
Negative	FalsePositive(FP)	TrueNegative(TN)

TABLE. 2 CONFUSIONMATRIX

$$Accuracy = (TN + TP)$$
$$(TN + TP + FN + FP)$$

ISSN 2455 – 9091
IJCSET MARCH Volume 9 Issue 3

$$Precision = (TP)(FP + TP)$$
$$Recall = (TP)$$
$$(FN + TP)$$
$$FI - Score = 2(TP)$$
$$(FN + FP + 2TP)$$

4. RESULTSANDDISCUSSION

TheexperimentwasconductedonGoogleCollab,afreecloud-basedservice.Toevaluatetheperformance of the model, we used the 10-foldcross-validationtechnique.80%ofthedatasetwasusedfortrainingandtheremaining20%wasusedtotesttheaccuracyofthemodel.Theclassification phase was divided into two steps,where all features were used for classification. Theresults,showninFigure2,indicatethattheRandomForestClassifier,DecisionTreesClassifier,andAdaBoostClassifierachievedaccuraciesof90.4%,85.1%,and82.4% , respectively.Thehighestaccuracywasachievedusing theRandom ForestClassifier at90.4%. familyhistoryofdiabetesshouldtakeextraprecautions.OurPredictionmodelwasabletoachieve an accuracy rate of 90.5% using RandomForestAlgorithm.

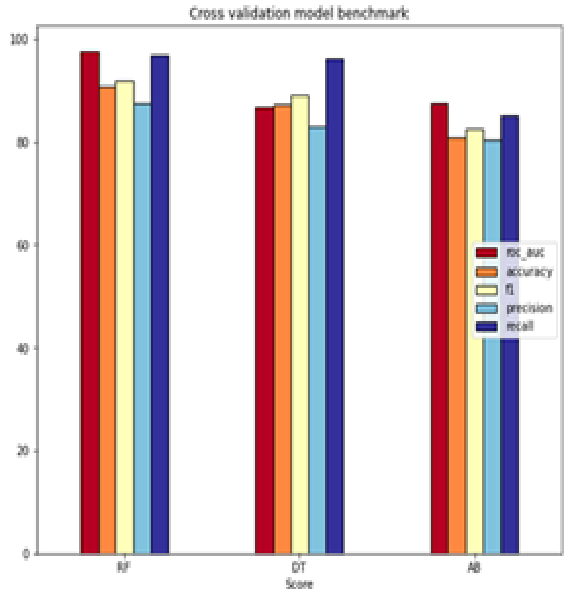

Figure.3PerformanceAnalysisUsingAllFeatures

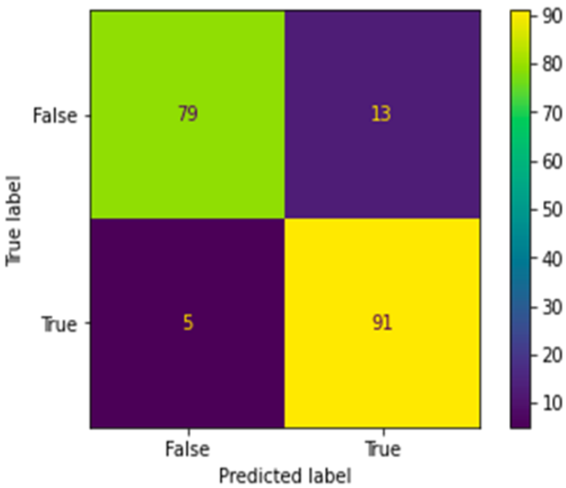

Figure.4 ConfusionMatrix

Evaluation Measure	Random Forest Classifier	Decision Tree Classifier	AdaBoost Classifier
Accuracy	90.4%	85.1%	82.4%
F1-Score	0.91	0.86	0.83
Precision	0.875	0.79	0.79
Recall	0.94	0.95	0.88
ROC AUC	0.90	0.84	0.82

TABLE.3 CLASSIFICATION PERFORMANCE USING ALL FEATURES

CONCLUSION

Based on the findings of the experiment, it can beinferred that the Random Forest algorithm performsbetterthanotheralgorithmsintermsofaccuracy.To preventdiabetes, individualsshouldstrivetomaintain consistent glucose levels, prioritize theirmentalandphysicalhealth,andconsumeabalanceddiettoregulatetheirinsulinlevels. Thosewitha.

REFERENCE

[1]. Kumar, Tiwari, Pande, "Diabetespredictionusingmachinelearningtools","20214thInternationalConferenceonRecentTrendsinComputer Science and Technology (ICRTCST)",27 May2022.

[2]. Elif Nur Haner Kırğıl, Begüm Erkal, TülinErçelebiAyyıldız,"PredictingDiabetesUsingMachineLearningTechniques","2022InternationalConferenceonTheoreticalandAppliedComputerScienceandEngineering(ICTASCE)", 13January2023.

[3]. Khilwani, Gondaliya, Patel, Hemnani Gandhi,KumarBhart,"DiabetesPrediction,usingStackingClassifier","2021InternationalConferenceonArtificialIntelligenceandMachineVision(AIMV)", 10January2022.

[4]. Kumari Shreya, Krishna Prathibha, "MachineLearningbasedDiabetesDetection","20216thInternational Conference on Communication andElectronicsSystems(ICCES)",02August2021.

[5]. Q.Wang,W.Cao,J.Guo,J.Ren,Y.Chengand D. N. Davis, "DMP MI: An Effective DiabetesMellitus Classification Algorithm on ImbalancedData With Missing Values," in IEEE Access, vol.7,pp.102232-102238,2019.

[6]. Deepti Sisodia, Dilip Singh Sisodia, "PredictionofDiabetesusingClassificationAlgorithms,"Procedia Computer Science, vol. 132, pp. 1578-158, 2018.

[7]. Abdullah Caliskan, Mehmet Emin Yuksel, Hasan Badem, Alper Rasturk, "Performanceimprovement of deep neural network classifiers byasimpletrainingstrategy, "Engineering Applications of Artificial Intelligence, vol. 67, pp.14-23,2018.