Turkish Journal of Computer and Mathematics Education *Vol.12 No.3(2021), 5638-5651*

Research Article

A review on prediction of diabetes type 2 by machine learning techniques

Alka Singh [a]

[a] Professor Anjna Jayant Deen, Professor Sanjay Silakari,Department of Computer Science and Engineering, University Institute of Technology, RGPV, Bhopal, India
Email:[a]alkasingh929@yahoo.com , anjnadeen@rgtu.net , ssilakari@yahoo.com

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: Machine learning is considered to be one of the most promising tools when it comes to working with heterogeneous data. It provides a new dimension which enables one to extract relevant data and take decision for the effective functioning of the network, making use of network generated data. Every sphere of our life is now dependent on machine learning. It has flourished in every dimension. Making it versatile and ever demanding.

Department of healthcare contains very abundant and sensitive information which is needed to be carefully handled. Diabetes mellitus is increasing exponentially and is spreading like anything in the world. A reliable prediction system should be present for diagnosing diabetes. Variety of machine learning techniques find their use in the examination of data from variant perspectives and summarizing it into effective information. Usage of new patterns is done to elucidate these patterns in order to deliver relevant information for their users. By making use of techniques such as SVM, random forest, logistic regression, naïve bayes etc the prediction of diabetes can be done easily and accurately. In this study we will make use of different machine learning techniques and try to find accurate prediction regarding the same.

Keywords: Machine learning, diabetes type 2, supervised, unsupervised, reinforcement, training and algorithm

1. Introduction

Machinelearninghasthepotentialwhichenablesitto learn from previous data to generate futuristictrends in behavior. It has the capability to learn byits own. Machine learning can be applied onnumerous data making it very integral to thetelecommunication world today (Hang Lai et al2019) [1]. Machine learning methods detectlinearities/non linearitiesintherelationship

between dependent and independent variables(Geofrrey et al 2019) [2]. They can be used formaking predictions in case of continuous outcomes,known as regression type problems or can be usedfor making predictions in case of levels ofcategorical variable, which is known asclassification problems. It gives solution from theproblemsandlearnshowtotacklewiththeproblem thatmayormaynotbe sameby making

useoftrainingdatasetprovidedtothealgorithmearlier.

Diabetes is such a prolonged disease that canhappen when body cannot efficiently make use ofthe insulin it generates. As a result, diabetes affectsorgans which include heart diseases which could beheart stroke, high blood pressure andatherosclerosis, nerve damage that could lead tonumbness, gradually losing all sense of feelingespecially in the limbs, kidney failure is verycommon in diabetic patients, and hearingimpairment is also seen in diabetic patients, the riskof Alzheimer's disease increases with type 2diabetes.

Diabetescan becategorized into threetypes:-

(a) ChildhoodorJuvenilediabetes

(b) AdultorType2diabetes

(c) Type3orGestationaldiabetes

Generally, type 1 diabetes occur because of thedeficiency in insulin production and is commonlyfound in children. Diabetes type 2 is a chronicdisease which affects how the human bodymetabolizes glucose. In case of diabetes type 2, thehuman body behaves in either of the 2 ways; firstlyit resists the effect of insulin which is a hormoneresponsibleforregulating themovementofsugarinthe cells. Secondly it doesn't produce ample insulinforthemaintenanceof normalglucoselevel.

Diabetes type 2 was known to be adult onsetdisease but nowadays much of the younger agegroupisbeingdiagnosedwith thesame, becauseoftheriseinobesity in children.Thereisnocure

available for the same but person can switch fromsedentary life style, follow balanced diet and canexercise well to manage the disease, as depicted infigure1. If this would not suffice then the personshould go for medications and insulin therapy. Theinsulin is secreted into the bloodstream by thepancreas. This insulin then circulates, enabling thesugartoenter thebodycells. Theamountofglucose in the bloodstream is lowered by theinsulin. Glucose i.e. sugar, is a major source ofenergy for cells that make up muscles and othertissues and it comes from food and liver. In case oflower glucose level the liver breaks down glycogeninto glucose in order to keep the glucose levelnormal. When it comes to type 2 diabetes, the sugarstarts to build up in the bloodstream instead ofmoving into the cells which lead to more release ofinsulinbybetacellsinthepancreas,graduallythese cells become impaired and become incapableof releasing more insulin to fulfill the requirementof body whereas in case of type 1 diabetes theimmune system by mistake destroys beta cellswhich leavethebody withlittleor no insulin.

Gestational diabetes is hyperglycemia whichhappens due to the change in hormones duringpregnancy.

Fot the past few decades we have seen that themachine learning discipline is assisting us to solvedifferent relevant biomedical problems. Themachinelearningtechniquesarefoundtooperateinboth real-life and scientific problems. In this study,we will be evaluating the performance of variousmachinelearningtechniquesfortheclassificationofpeople whetherthey arediabetic ornot.

1.2.Generalized Architecture

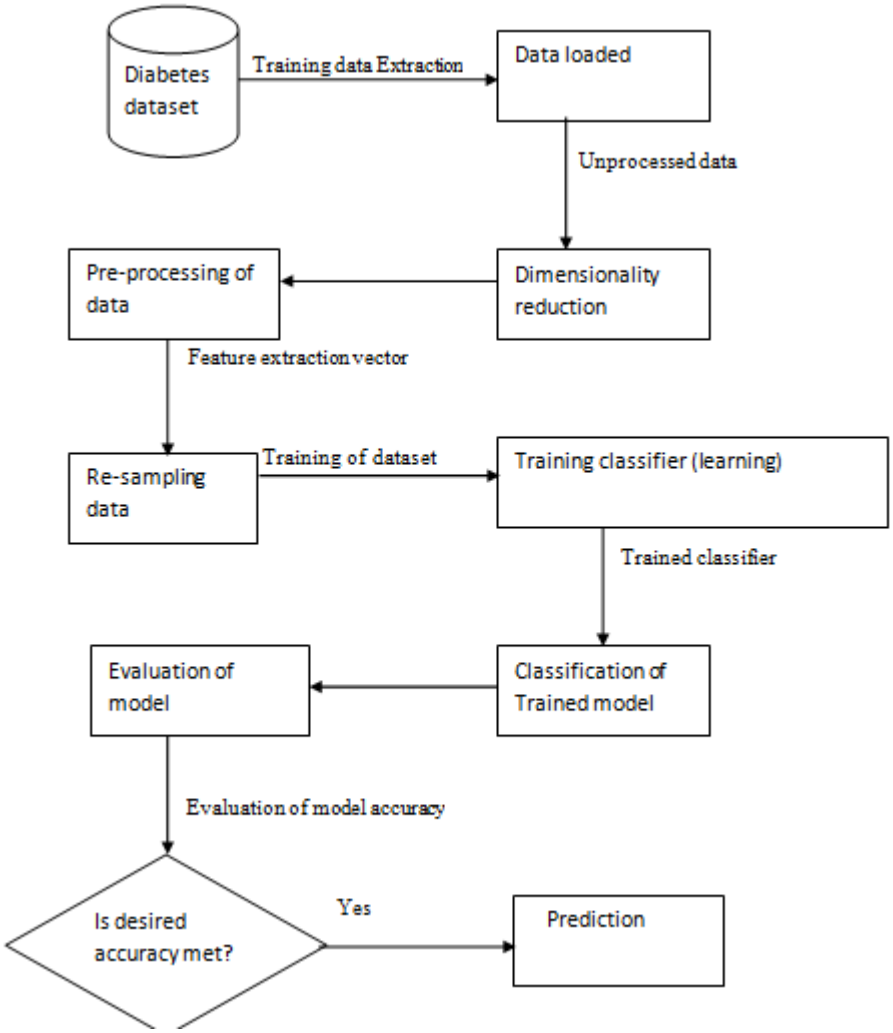

Figure1

2.Literature survey

Thissectionreviewsvariousresearch worksthatarerelatedtoourproposed work.

Arianna Dagliati et al [3] designed a machinelearningpredictionmodelfortype2diabeteswheredemographic data (i.e age, gender, time todiagnosis), clinical data from the EHR (BMI,Hba1c,lipid

profile,smokinghabit),Administrative data (antihypertensive therapy) arecollected based on that predictive model formicrovascular complications in the population wasdesigned which focused on issues such asnephropathy, neuropathy and retinopathy. Themodel showed higher AUC values in case of SVMand Random Forest. The final model is based onlogistic regression with rebalanced classes whichsupportsnomograms.

D. Asir Antony Gnana Singh et al [4] designed amachinelearningmodelforpredictionofdiabetes.Three different type of supervised learningalgorithm namely probabilisticbasednaïvebayes(NB),functionbasedmultilayerperceptron

(MLP), decision tree based random forests(RF) areused. Test methods such as 10 fold cross validation(FCV), makes use of percentage split (PS) with66%andtrainingdataset(UTD).Thepreprocessingtechnique is used to increase the accuracy of themodel. In case of pre-processing technique averageaccuracy for NB is increased as compared tomachinelearningalgorithm.

K. Srinivas et al [5] developed data miningapplication techniques that can be used in case ofhealth care and prediction of heart attacks. In theresearch they made use of medical profiles such asblood pressure, age, blood sugar and sex and usedthis to predict the likeliness of getting kindneyproblemsandheartattack.

IdemudiaChristianUwaet el[6]designedamachinelearningpredictionmodelforpredictionof diabetes. On applying univariate selectionmethod with chi squared statistical test in case ofnon negative feature we obtain following attributeslikeplasma,bloodpressure,age,pedigreefunction,

B.M.I. Here the algorithms that were been appliedarenaïvebayes,logisticregression,SVM, XG

Boost, KNN. The dataset were of 2 types one frompimaIndiandatasetandtheotherwasdr.schorlingdataset derived from hospital. We found that onboth the dataset, naïve bayes model showedconsistency and after naïve bayes logisticregression proved to be better wih accuracies of83%and81%respectively.

V. Ranjani et al [7] emphasized on the potentialuse of classification based data mining techniquesthat incluses aritifical neural network (ANN), rule-based methods, Naïve Bayes and decision treealgorithm to huge volume of data of health care. Intheresearch,medcalproblemshavebeenanalysedand evaluated which include blood pressure andheartdisease.

M.Durairaj et el [8] demonstrates a hybridprediction system consisting of Rough Set Theoryand Artificial Neural Network for depictingmedicaldata.Thisprocessof developmentofanew data mining technique and a software to helpcompetent answers in case of analysis of medicaldata is been explained. A hybrid tool is beenproposed that incorporates RST and ANN to makeefficient data analysis and indicative predictions.The experiments' on spermatological data set thatisbeenusedfortheprecictingexcellenceofanimalsemen. The hybrid prediction system is beenapplied in case of pre-processing medical databaseand for the purpose of training the ANN for theprediction of production. The accuracy in case ofprediction is obtained in case of comparison that isbeen made between the observed and predictedcleavagerate.

S.M Hasan Mahmud et al [30] designed a machinelearningmodelfortheprediction ofdiabeteswhere

the comparison is been based on the performanceevaluation by 10-fold validation technique. Aframework is also been generated for diabetesprediction, monitoring and application (DPMA).Here the basic concept is that multiple machinelearningclassifiersaresupposedtoperformbetterthan asinglemachine learningclassifier.

Akm Ashiquzzaman,AbdulKawsar Tushar etal

designed a diabetes prediction model bymaking use of the application of the drop outmethod.Novelformofdeepneuralnetworkfortheprognosis of diabetes with increased accuracy isbeen discussed.

Ioannis Kavakiotis et al [32] designed machinelearning and data mining approaches that wereappliedonalltheaspectsofDMresearchand thatwere applied on biomarker identification andpredictiondiagnosis.

Muhammad Azeem Sarwar, Nasir Kamal et al [33]designed a model for the prediction of diabetes bymaking use of various machine learning learningtechniques where the data is divided into trainingdata and testing data. Enthought Canaopy methodand thus the result is been obtained. SVM andKNNhaveshown higher accuracy in themodel.

2.1.Literature review:-

S.No.	AUTHOR	DESCRIPTION/WORK	TITLE	METHODS/TOOLS	RESEARCHGAP
1.	AriannaDagliati,Simone Marini,LuciaSacchi,GiuliaCognai	Dataminingandcomputational methodsare adopted in to derivepatient specificinformation to predictoutcomeof interest.	Machinelearningmethods topredictdiabetescomplications	SVM, RandomForest, LogisticRegression	Development ofpredictivemodel for theonset ofmicrovascularcomplicationsin case ofT2DM.
2.	Dr. D. AsirAntonyGnanaSingh Dr. E. Jebamalarleavline, B. ShahawazBaig	Supervised learningalgorithm is used for thediagnosis and predictionof diabetes and theaccuracy is increased bypre-processingtechnique.	Diabetesprediction usingmedicaldata	Probabilistic basednaïve bayes(NB),function basedmultilayer perceptron(MLP), decision treebased randomforests(RF)	Thepreprocessingincreasedtheaccuracy of allthe models thatinclude naïvebayes , MLP,RF except the10 fold cross validationmethod.
3.	K.Srinivas, B. KavihtaRani, Dr.A. Govrdhan	An effective approach forthe extraction ofsignificantpatternsisbeen established, on thebasis of calculated weight,the value greater than thethreshold ischosen.	Applications ofdata miningtechniques inhealthcare andprediction ofheart attacks.	Naïve bayes,Artificial NeuralNetwork, DecisionTree	The role of textmining can beused inexpanded so towiden its rolein-case ofunstructureddata.
4.	IdemudiaChristian Uwa,NehikhareEfehi	Data mining processesthat are been used in caseof medical diagnosis andthe usage of variousmachine learningtechniques for predictingdiabetes.	Evaluating theperformance ofmachinelearning algorithms fordiagnosingdiabetes inindividuals	Logistic regression,naïve bayes, supportvector machine,XGBoost,kNN	To gather newdata and finetuningtechniquestobe used, meansof handlingimbalance classdatacanbeexplored.
5.	M.Durairaj, V.Ranjani	Combination of more thanone data mining techniquefor diabetes predictionyielding better accuracycomparatively.	Data miningapplications inhealthcaresector: AStudy	Rough Set, Artificialneural network, ANNand HybridTechnique	Hybridtechniques when appliedfor variousdiseases canyield betteraccuracy.

6.	M.Durairaj, K.Meena	Hybridization of two MLtechniques such as ANNand RST is used as aalternativetotheconventional methods fortheprediction.	A hybridpredictionsystem usingrough sets andartificial neuralnetworks.	Rough Set Theory(RST), ArtificialNeural Network(ANN)	Incorporationof biologicalinformation,systematiccomparison ofdifferentmachinelearningalgorithms,hybridization ofrough sets andneural networkensembles tobuild predictorsfor improvingperformance.
7.	SMHasanMahmud,Md AltabHossin,Md. RazuAhmed	Aframeworkisdesignedfor real time diabetesprediction, monitoringand application. Anoptimized and efficientmachine learningapplication is developedwhich could predictdiabetes. Differentclassification criteria areused for investing theperformance of differentclassificationtechniques.	Machinelearning basedunifiedframework fordiabetespredicti on.	10 foldvalidationtechnique along withnaïve bayes, ANN,Logistic Regression,Decision TreeRandom Forest andSVM	Mobile basedapplication forprediction ofdiabetes basedon multiplemachinelearningclassifiers thatwould performbetter thanasingle learningclassifier.
8.	AkmAshiquzzaman,Abdul KawsarTushar, Md. RashedulIslam, Jong-MyonKim	A reliable predictionsystem that aims tominimize over-fittingissue by using dropoutmethod.	Reduction ofOver-fitting inDiabetesPrediti onUsing DeepLearning NeuralNetwork.	Novel form of deepneural network withthe application ofdropoutmethod	Performance incase ofpredictivemodels ofdiabetes canhave betterpredictionscores whichwill pave wayforbreakthrough inhealthprognostication.

9.	IoannisKavakiotis, Olga Tsave,Athanasios, NicosMaglaveras,IoannisVlahavas,IoannaChouvarda	Asystematicreviewofthe applications ofmachine learning, datamining techniques andtools with respect todiagnosis, prediction,genetic background andenvironment, and healthcareand managementisbeen done based ondifferentmachinelearningtechniques.	Machinelearning anddata miningmethods indiabetesresearch.	Logistic regression,SVM,ANN, NB, Linear DiscriminantAnalysis, KNN,fuzzy c-mean,Random forest,CART, MultifactorDimensionalityReduction.	With the adventof bio-technology, andwith the hugeamount of dataproduced, andwith theeverincreasingamount ofEHRs thediagnosis, andtreatment ofdiseases can beenriched.
10.	MuhammadAzeem,Sarwar,Nasir Kamal,Wajeeha Hamid,MunamAli Shah	Datasetsofpatient'srecord is obtained andvarious machine learningalgorithmsareapplied andbasedonthataccuracyand prediction isdone.	Prediction ofDiabetes UsingMachineLearningAlgorithm inhealthcare.	Naïve Bayes,KNN,SVM,LR,DT, andRandomForest	Theadvancementcan be made interms ofapplyingvarious techniques suchas big data,cloudcomputing withmachinelearningtools.

3 Machine Learning Algorithms:-

The significance of machine learning algorithmsdependsinthedevelopmentofmodelsthatisbasedontheexistingdataandconsequently,classification or prediction by making use of noveldata. Machine learning methods have been widelyusedin variousapplicationsindiversifieddomains

like system biology, genomics. Specificallyspeaking, supervised machine learning techniqueshave been finding immense importance in anumber of bioinformatics prediction techniques.The aim here is to showcase an overview of themachinelearningalgorithmsaswellasapplicationmethodsbasedonsame.

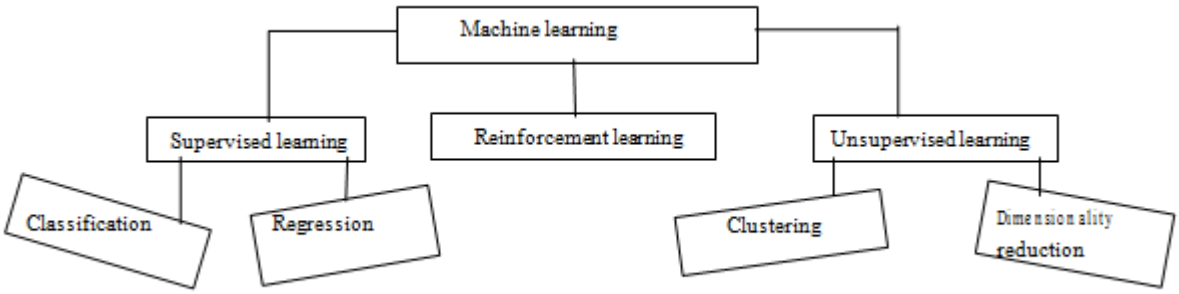

Figure 2: Types of machine learning algorithms:-

Machine learning techniques can be broadlycategorized as:-

Supervised learningUnsupervised learningReinforcementlearning

2..3.Supervised learning:

Supervised learning has the involvement ofsupervisor which works in the same way as ateacher in real life. It

is such a type of learning inwhichweteachortrainamachinebymaking useof data which has already been tagged with thecorrectanswer (PaulAkangah et el,2018) [9].

Further, we experiment the machine with new setsof data so that the supervised learning algorithmcananalyzethetrainingdataand cangiveacorrectoutcome on the basis of the previous labeled data(R.Sathya et al 2013) [10].

Supervised learning is classified into twocategories:

- **Regression:**IFtheoutputvariableincludesa real value, for example "dollars" or"weight"then wecallitaregression

problem. We see use of regression algorithmwhile dealing with decision trees, linearregression, logistic regression etc (GiovanniGrano et el, 2018) [11]

- **Classification** : IF the output variableincludes a category, for example "Red" or"blue" or "disease" and "no disease" we callit a classification. We see use ofclassification algorithm while dealing withnaive bayes classifier, support vectormachine, K- Nearest Neighbor(Kondi SrujanKumarret el, 2019)[12]

3.1.1.Artificial Neural Network:-

This algorithm is conceptualized on basis ofbiological neurons. We can see that in case ofbiological learning process the process of learningisthoughttobebasedon minoradjustmentstothesynaptic connections between neurons whereas inANN the learning process is totally based on theinterconnections between the processing elementswhich combineto formnetwork topology.

Basically, ANN consists of 3 layers i.e. input layer,hidden layer,and theoutputlayer.Weseethat

incase of ANN, the training of hidden layercontainingnetworkand makesuseofitsconnectedstructures for the purpose of pattern recognitionand classification. In case of bioinformaticsapplications of ANN, we employ different types ofarchitectures with perceptron and multi layeredperceptron being thesimplestinthecategory.

Radial basis function networks and Kohonen selforganizing mapsarealsofound useful.

ThemajorstepswhichareinvolvedinanANNalgorithm areasfollows:

- Byprocessingavailabledatatrainingandtestdatasetsare generated.
- Data is encoded into digital format bymakinguseofencodingsystems,suchasbinary systems.
- ANN architecture is designed anddevelopedbymakinguseof3layersforthepurpose ofprediction.
- Ann is trained by making use ofappropriateinputdataandparameters.
- ANN model is such selected which givesthevalidoutput.
- Ann model is thus validated by using testdataset for the purpose of estimation ofefficacy for prediction.

The biggest advantage we observe in case of ANNis its ability to analyze and process over largecomplex datasets, having non-linear relationships.This model includes more benefits like having theability to handle noisy data and the caliber ofgeneralization. The limitation of the methodobserved is in the amount of time that would betakenincaseofprocessingcomplexdatasets.ANNhas extensively been used in case of geneprediction,sequencefeature analysisetc.

3.1.2.Support Vector Machine

Support Vector Machine is a supervised learningmethod that is based on statistical learning theory.For linearly separable illustrations, SVM creates amaximum margin hyper-plane that separates thedata points into 2 different classes. The hyper-planeworks as a decision surface between two classes(Affsan Abbrar et al, 2018) [22]. In case of non-linearly separable data, firstly SVM changes datainto higher dimensional feature space andconsequently makes use of a linear maximummarginhyper-plane.Thisleadsto theintroduction

of computational intractability that requires atransformationto ahigherdimensionalspace(An

T. Nguyen et al, 2018) [23]. SVM resolves this bydefining most appropriate kernel functions by thehelp of which the computations can be taken intoconsideration in the original space itself. The threepopular kernel functions that are used generally arelinear, polynomial and radial basis function (SandraVieira et al, 2019) [24]. In case of bioinformatics,we see many domain specific kernel functions suchasgraph

kernel,stringkernel(JesseH.Krijtheet

al, 2017)[25].Thisconcept canalsobe usedincase of multiclass classification. The two mostcommon multiclass classification methods that findtheirusehereareviz.,oneagainalland oneagainstone (Konstantinos Sechidis et al, 2017) [26]. Thesteps that are employed in SVM algorithm aregiven below:

- Feature vector is constructed in-order torepresent positive and negative dataset:this feature vector contains properties ofthe input data that could be amino acid,physio chemicalpropertiesetc.
- Appropriate kernel function is chosen soasto fitforthepredictiontaskby makinguseof classifiertraining.
- The model is selected with bestperformancetomakepredictions.
- Theapplicationofchosenmodelfordoing predictions on the unknown inputdata set, the most robust classifier is SVM,ithasthebestgeneralizationabilityincase of unseen data in comparison to othermethods.

SVM is the most commonly used machinelearning method that is used in case ofcomputational biology and bioinformatics.It is also been used for secondary structureprediction, gene finding, fold recognitionaswellasbindingsiteprediction.

Support vector machine is a distinguishingclassifier which is previously defined byseludinghyper-planewhichmeans,onthegivenlabeledtrainingdata,heresupervisedlearning , the algorithm gives output in theform of a hyper-lane which will categoriesnew examples. The hyper-lane is a linewhich divides a plane into two parts, incase of the two dimensional space whereeachoftheclasslieontheeitherside.

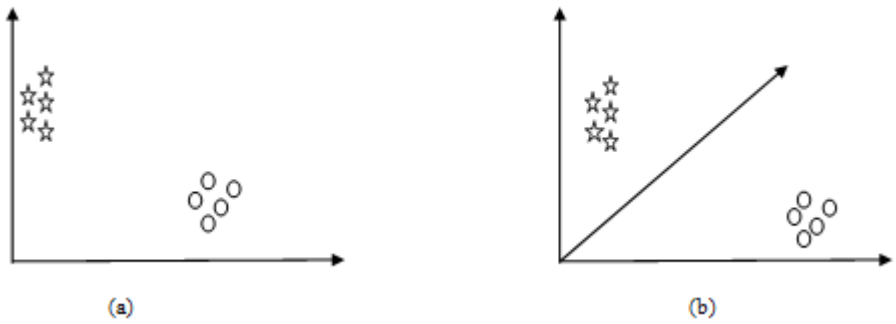

Figure3

Figure 2, in the above example diagram “b” shows that a line in this case separates the two different classes asdepictedin example“a”.Hereweusetheequation oflineasy=x.wemay alsousethefollowing y=mx+c.

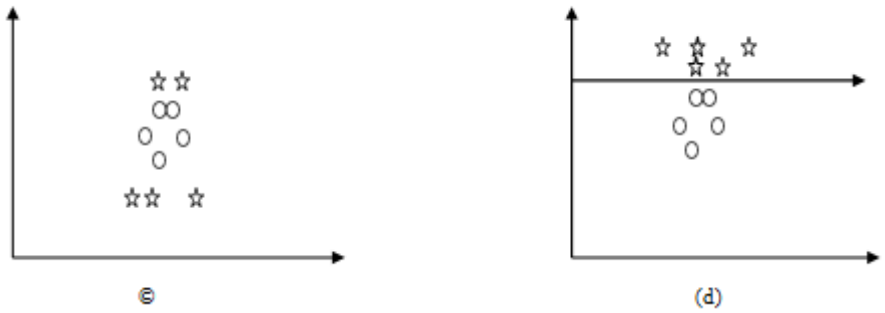

Figure4

Figure3,intheaboveexampleweseedifferentpropertyofSVMwherewearemaking useofz^2=x^2+y^2,herebymakinguse ofsquarevalueswecanseparate “c” as“d.

3.1.3.K-Nearest Neighbor

KNN classifier finds the k nearest examples in thereference set, and considering majority vote

fromtheclassesofthesekexamplestoallocateaclasstoa query (P.Yasodhaetal2014)[18].Assignmentof classes in decision boundaries are implicitlyderived in case of KNN. Below are the relevantsteps that are involved in the development of KNNclassifier:

- Featuresetisconstructedandadistancemetric is used to compute distancesbetween features.
- Numberofnearestneighborsisdeterminedforthetrainingset.
- Euclidian distances or any other distancemeasure such as Mahalanobis distance arecalculated between the query instance andthetrainingsamples.
- Distances are sorted and nearest neighborsare determined on the basis of the k-thminimum distance.
- Class label are predicted in case of new orunknown instance by making use of theclasslabel ofnearestneighbors.

The most significant advantage of KNN method isthat it has higher efficiency on large datasets androbustness while processing noisy data (GopiBattineni et al, 2019) [19]. The drawback of KNNisitshighcomputationcost, which deducesits

speed. In case of bioinformatics, we observe thatKNN model is been employed successfully (Yun-leiCai et al, 2010) [20].

3.1.4Decision Tree

Decision tree are considered to be a branch testbased classifier. The construction of the sameinvolves the analysis of the set of trainingexamples, class labels are known for them. Newand unseen examples are classified by thisinformation. A leaf node symbolizes a specificclass and every branch represents a group ofclasses (Mikolas Janota et al, 2018) [21]. A test ona single attribute value is been represented by thedecision node, with its one main branch and thesubsequent classes are represented as possibleoutcomes (Sullivan hue et al, 2018) [29]. Themajor steps that are to be considered in case ofdecision tree algorithmisgivenbelow:-

- Training dataset is prepared in such anappropriate form in case of the classifierby the method of feature extraction frominputdata.
- Decisiontreeisconstructedbyputtingtheinstancesintrainingsetattheinitialnode.
- The instances are divided into twodistinguishable classes i.e. child nodesbased on their chosen testvalue.
- By the recursive application of the laststep it is checked that the fulfillment oftermination or pre- pruning condition ismet.
- The resultant tree is pruned with itsapplications for performing predictions.Decisiontreesaresimpleclassifiersandhence have better interpretability ascompared to other machine learningmethods. They are widely used inbioinformatics for predicting geneticinteractionsand related applications.

3.1.5.Random Forests

Random Forests is a group of randomly createdindependent classifiers and decision trees (PaulAkangah et al, 2018) [9]. It generally depictssubstantialperformanceimprovisationsoversingletree classifiers such as C4.5, CART. Randomnessor Feasibility can be introduced in the RFalgorithm intwoways:

1. .1.Bootstrapping,abootstrapset is created from the originaltraining data set by making useof random sampling by doingreplacementtogenerateeachtree (Marcus Muller et al,2018)[28].

1.2. Construction of bootstrapset is done by making use oforiginal training dataset by thehelp of random sampling by theprocess of replacement in orderto generate each tree.

2. Node Splitting: Here the selectionof subset of attributes is carried out.On splitting a node, where there areM input attributes, then the number‘m’, where m<<M and is beenspecified in such a way that at eachnode, m attributes are randomlyselected and the best split isconsidered on them. A value that isgood of ‘m’ is by default selected bymaking use of variousimplementations, considering ‘m’ assqrt (M) for the very purpose ofclassification. On the basis of theCART algorithm the classificationtree is induced by making use of ‘inbag’ data. After that an out of bagdata, that is been formed after leavingout the in-bag samples from those ofthe original data is used in crossvalidationwork.The stepsinvolvedincaseofrandomforestalgorithmaregivenbelow:

- CART algorithm is beenemployed on data for thegrowth of randomclassification trees.

- Bootstrap data is beenused which is known asin-bag set that is used totraintheCARTalgorithm.
- On the basis of the bestcondition on a randomsubset of 'm' attributesnodesplittingisdone.
- By making use of majorityvote strategy in order todecideclassaffiliationincaseofeach OOBsample.
- Variable importance (VI)ranking,thatcanbeusedlater to retrain random forestby using a smaller subset ofthemostrelevantvariables.
- Resistance to over fitting ofdata random forest and itsvariants are been applied tosolve a huge amount ofbioinformatics problemswhichincludesclassification of geneexpression, analysis of massspectroscopy data fordiabetes prediction,sequence annotation andprediction of diabetes 2mellitus.

3.1.6.Ensemble Classifiers

In case of ensemble classifiers, the individualdecisionsincaseofasetofclassifiersarejoinedwith weighted or un-weighted voting for thepurposeofclassification ofnewinstances.

Ensemble classifiers are also called as multi-classifier systems. These classifiers are found to beefficient in prediction tasks because of the fact thatthey find use of a combined classifier and cancapture features that cannot even be captured bymaking use of any single model alone. Thesemethods are been applied in differentbioinformatics problems because of their highprediction accuracy.

3.1.7.Unsupervised learning:

Unsupervised learning is that type of training inmachine where we make use of information that isneither labeled nor is classified and so it lets thealgorithm to work on this information without anyprior guidance as in case of supervised learning(Nagdev Amruthnath et al, 2018) [13]. The task ofthe machine here is to group unsorted informationinto patterns or on the basis of differences andsimilaritieswithoutthepriortrainingbeingdoneonthedata (MemoonaKhanametal, 2015)[14].

Unsupervisedlearningisclassifiedintotwocategories:

- **Association**:Dimensionalityreductionisthe other name of association rule learningproblem.Anassociationlearningproblemisone where one needs to find rules that couldbe applied to large data sets that may includefor example people who wish to buy A andarealso intendedtobuyB.
- **Clustering**: A clustering problem is onewhere we want to find the inherentgroupingswithinthedata,whichincludesgrouping various customers by theirpurchasing behavior (Oyelade et al 2010)[15].

Wefinduseofclusteringalgorithmwhiledealingwith K-means;meanshift,K-medoids.

3.1.8.Artificial Hidden Markov Models (HMM)

Hidden Markov Models have found their use invery popular machine learning approaches such asin case of bioinformatics. They are probabilisticmodelsthataregenerallyimpliedintimeseriesandlinear sequences. It can be used to describe theevolution of those events which are observable andthese depend on internal factors, which themselvesarenotobservable.Hereweseethattheobserved

events arecalledassymbol andtheinvisiblefactors that are underlying the observations that arereferred to as a state. An HMM comprises ofseveral states, that are connected by means oftransition probabilities, which leads to theformation of a Markov process. Every state herehas anobservablesymbol that is beenattachedtoit. An HMM comprises of visible process withobservable events and a hidden process whichincludes internal states with their movement intandem. The goal here is to find the optimal pathfrom the states, which leads to maximization of theoccurrence of observed sequence of symbols. Therelevant steps that associated in the algorithm forthegenerationofHMM are given below:

- HMM architecture is been developed bymaking use of various states whichultimately represent the given set offeatures.
- Assignment is been done of the hiddenstates to the features and so is theconstructionofHMMmodelisbeendone.
- The HMM is thus trained using supervisedtechnique or unsupervised technique inorder to let the model sufficiently fit theproblem that isunderstudy.

- Emission probabilities are derived thatinfluence the distribution of observedsymbols, which implies that theprobability of a symbol being observedprovidedthatHMMisinaspecificstate.
- HMM is decoded for the prediction ofhidden statesfromthe data.
- The benefits associated with HMMs arethe ease of their use, need of smallerdatasets and precise comprehension of theprocess.

Among the major drawbacks associated withHMMs is their higher computational cost. HMMsare found to be most effective in case of biologicalsequence analysis and so they are periodicallyapplied for multiple sequence alignments, genefinding,etc

3.1.9K-Means clustering

The k mean clusteringalgorithmprovides ageneralized methodto implement approximatesolution. The reason why k mean clusteringalgorithm is very popular is because of the ease andsimplicity. Kmean canbeconsideredtobea

gradient descent procedure, where the initiation inthe algorithm is done at starting cluster centroidsand it iteratively decreases the objective function.The convergenceof the k meangenerallytakesplace at the local minimum. It basically performsthe updation work unless the local minimum isfound.Theproblem tofindtheglobalminimumis

NP- complete. The time complexity of the k-means clustering algorithm is O(nkl) where, therequired number of clusters is denoted by “k”, thetotal number of objects in the dataset is denoted by“n” and the number of iterations is denoted by “I”,k<=n,I<=n.

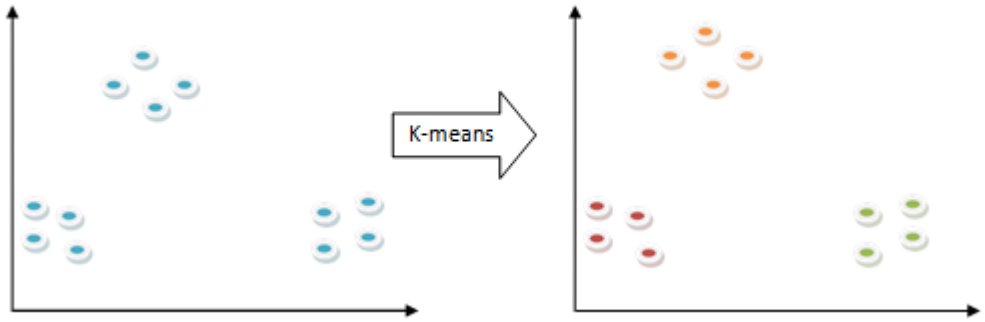

Figure5: Diagram depictingK-meansclustering

Reinforcement learning

Reinforcement learning belongs to that area ofMachine Learning where the actions are takenpurely to achieve maximize rewards in a specificsituation. It can be used on different types ofmachinesandevenonsoftwareforfindingthebestpath possible or behavior it is supposed to take inany specific situation (Jiachi Xie et al 15) [16]. Itdistinguishesitself fromsupervised learningin a

way that in case of supervised learning the trainingdata has the answer key with it and the model istrained with the correct answer by its own on theother hand, in case of reinforcement learning,answer key is not available but here we can see thatthe reinforcement agent decides what is to be donein order to perform the given task (Nicolas Bougieet al, 2019) [17]. In the absence of training data set,it is bound to learn from its own previousexperiences.

Table1: Theadvantagesanddrawbacksrelated tovariousmachinelearning algorithmare:-

Algorithm	Advantages	Drawbacks
ARTIFICIAL NEURALNETWORK	• Good performance in caseofnonlinearrelationships. • Capacity to handle noisydata.	• Computationalburdenisgreater. • Over fitting isaproblem.
HIDDENMARKOVMODEL	• Precise comprehension ofbackground process. • Easy to use and ispowerful.	• Intensivecomputation • Slower than othermethods. • Pronetoover-fitting.

SUPPORTVECTORMACHINE	• Bestgeneralizationabilityisprovided. • Robustto noisy database. • Susceptibilityislesstoover-fitting.	• Computationisexpensivein some cases such as incase of non-linearlyseparableproblems.
K-NEARESTNEIGHBOR	• Simpleand easy to learn. • It is found efficient whenthetrainingdataislarge.	• Complexincomputation • As the number ofattributesgetsincreased,
	• Fasttraining.	performance becomesinconsistent.
DECISIONTREE	• Capability to handle bothcontinuous and discreteattributes. • Interpretabilityisbetter. • Resultsarebetterincaseof redundantattributes.	• In the presence of largenumberofclasses,thedataisprone toerrors. • Sensitivitytowardssmall variationsindata.
RANDOM FOREST	• Highspeedandaccuracy. • Less prone towards over-fitting. • Able to evaluate everyattributeforprediction.	• Tendency of over-fitting incaseof noisydata.
ENSEMBLECLASSIFIERS	• Incaseofprediction,greater efficiency isobtained. • Utilization ofdataismore.	• Computational complexityismore.

4.Machine Learning Advancements in diabetes prediction:-

Machine learning can be used in case of digitaldiagnosisofanydisease. Itcan detectpatternsofcertain diseases and help in providing a broaderperspective.

4.1.Diabots:

Itisfoundthatthischatbotiscapableofinteractingwith patients seamlessly based on the symptoms.There are many generic text-to-text diabot i.e.diagnostic chatbot which makes use of NaturalLanguage Understanding (NLU) for the providingpersonalized prediction by making use ofgeneralized health dataset and also on the basis ofvarioussymptomssoughtfrom the patient.

4.2.Oncology:

Here the researchers are making use of deeplearning techniques for the purpose of training thealgorithm and to make it recognize carcinogenictissue (but at the same time it is taken intoconsiderationthatthebloodsugarlevelisnormal)at such a level that is comparable to evenphysicians.

4.3.Better Radiotherapy:

As the machine learning algorithms have thepotential to learn from the multitude of varioussamplesthatarebeenavailablein hand,itbecomeshighly effective to diagnose and find the variablesif any. The example includes Google's DeepMindHealthwhich isassisting the healthcare

professional to distinguish between the healthy andunhealthy people. Here the advancement is beenmade in terms of diagnosing eye damage done byvariousdiseaseswhich includesdiabetestoo.

4.4.Outbreak Prediction:

Machine learning is used in monitoring andpredictingepidemicsaroundtheglobe.ANNcanbe used to collect information from differentwebsites and predict information from dengueoutbreak to severe chronic infectious diseases. Thiscan also assist in knowing the world wide increaseinthediabetes patients roundtheglobewhichledus to the conclusion that India is the diabetic capitalofthe world.

4.5.Crowd sourced Data Collection:

Crowd sourcing has helped researchers andpractitioners to get access to huge amount ofinformation that are been uploaded by people basedon their consent. This helps in collected data that isbeen collected by the consent of the patients and isassisting inthe research.

The various applications available for thepredication of diabetes includes Diabetik byUglyApps,DiabetesinCheckbyEverydayHealth,iCookbook Diabetic by Publications International,mySugr Junior by mySugr GmbH, HealthyOut byHealthyOutandmanymore.

5.Conclusion/ Future work:-

The applications of machine learning could beappliedforthediagnosisofvariousdiseases,theirsymptoms, their cause, their treatment. The suddendeaths occurring due to kidney failure, heart attack,strokes etc. accompanied with diabetes can beprevented through early treatment and diagnosis. Inthe study we saw various algorithms such as SVM,decision tree, KNN, naïve bayes, etc making theiruse in the prediction of incidence of diabetes. Theclassificationtechniquesgivedifferentresultswhenapplied to different dataset. We found that variousclassification techniques are useful for differentdata sets. The variation in the model performancecan be noticed for different datasets and the causecould bepredictedaccordingly.

Future study can be focused on acquiring newdataset that would lead to new insight andknowledge to improving the prediction of diabetesusing machine learning techniques. Based on theparameters like age, body mass index, obesitylevel, history of chronic disease, etc whenaccompanied by various machine learningtechniques will lead to better prediction levels. Thenew dimension which is extending is usage is deeplearning whichwhenassisted withmachinelearning can give tremendous results in terms ofpattern recognition andbetterpredicted values

References

Hang Lai, Huaxiong Huang, “Predicting models for diabetes milletus using machine learning techniques, BMC publication, 2019

Geoffrey Charles Fox, James Alexander Glazier, Jcs Kadupitiya, James Sluka,” Learning Everywhere: Pervasive Machine Learning for effective High Performance computation: Application Background” research gate, 2019

Arianna Dagliati,” Machine Learning methods to predict diabetes complications”, Journal of diabetes Science and technology, 2018

D. Asir Antony Gnana Singh, “Diabetes prediction using medical data”, Journal of Computational Intelligence in Bioinformatics, pp. 1-8, 2017

Srinivas, K., Kavihta, R.B., Govrdhan, A.,”Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks” International Journal on
Computer Science and Engineering, 2(2), pp. 250-255, 2010

Idemudia Christian Uwa, Nehikhare Efehi, “Evaluating the performance of machine learning algorithms for diagnosing diabetes in individuals” International Journal of Science and Research (IJSR), 2018

Durairaj, M., Ranjani, V., “Data Mining Applications In Healthcare Sector: A Study”, International Journal of Scientific & Technology Research, 2(10), pp. 31-35, 2013

M.Durairaj, K. Meena, “A hybrid prediction system using rough sets and artificial neural networks”, International Journal of Interactive Technology and Creative, VOL.1.No. 7, 2011

Paul Akangah, Leotis Parrish, Andrea Nana Ofori-Boadu, Francis Davis,”Predicting academic achievement in fundamentals of thermodynamics using supervised learning technique” American Society For Engineering Education (ASEE) Southeastern Section conference,2018

R. Sathya, Annamma Abraham,” Comparison of supervised and unsupervised Learning Algorithms for Pattern Classification” International journal of Advanced Research in Artificial Intelligence, vol2, 2013

Giovanni Grano, Timofey V. Titov, Sebastiano Panichella, Harald C. Gall,” How High Will It Be? Using Machine Learning Models to Predict Branch Coverage in Automated Testing” MALTESQUE@ SANER, 2018

Kondi Srujan Kumarr, M Ashish Naidu, K Radha,”Pattern Discovery with Web usage Mining using Apriori and FP-Growth Algorithms” International Journal of Computer Trends and Technology, 2019

Nagdev Amruthnath, Tarun Gupta, “ A research study on Unsupervised Machine Learning Algorithms for Fault detection in Predictive Maintenance” Research Gate, 2018

Memoona Khanam, Tahira Mahboob,” A Survey on Unsupervised Learning Algorithms for Automation, Classification, and Maintenance” International Journal of Computer Applications, 2015

Oyelade, Oladipupo, Obagbuwa, ”Application of k-means clustering algorithm for prediction of students’ academic performance” IJCSIS 2010

Jiachi Xie, Chelsea M. Myers, Jichen Zhu, “Interactive Visualizer to facilitate Game

Designers in Understanding Machine learning" CHI 2019
Nicolas Bougie, Ryuntaro Ichise, "Skill- based curiosity for intrinsically motivated reinforcement learning," Springer 2019
P. Yasodha and N.R. Ananthanarayanan, "comparative study of diabetic patient data's using classification algorithm in weka tool" international journal of computer applications technology and research volume 3, 2014
Gopi Battineni, Getu Gamo Sagro, Chintalapudi Nalini, Francesco Amenta, Seyed Khosrow Tayebati,"Comparative machine learning approach: follow up study on type 2 diabetes prediction by cross validation methods", 2019
Yun-lei Cai, Duo Ji ,Dong-feng Cai , "A kNN research paper classification method based on shared nearest neighbor" NTCIR-8 Workshop Meeting, 2010
Mikolas Janota, "Towards generalization in QBF Solving via Machine Learning" AAA-I, 2018
Affsan Abbrar, Dr. Pranam Paul,"Development of Application to Recognise Hand Written Digit at rum time" International Journal of Innovative Research in Computer and Communication Engineering, 2018
An T. Nguyen, Matthew Lease, Byron C. Wallace, "Mash: software tools for developing interactive and transparent machine learning systems, IUI Workshops 2019
Sandra Vieira, Qi-yong Gong, Walter H.L. Pinaya, Cristina Scarpazza," Using Machine Learning and Structural Neuroimaging to Detect First Episode Psychosis: Reconsidering the Evidence, Advance Access Publications, 2019
Jesse H. Krijthe, Marco loog, "Projected estimators for robust semi- supervised classification," Springer, 2017
Konstantinos Sechidis, Gavin Brown, "Simple strategies for semi- supervised feature selection," Springer, 2017
Paul Akangah, Leotis Parrish, Andrea Nana Ofori-Boadu, Francis Davis,"Predicting academic achievement in fundamentals of thermodynamics using supervised learning technique" American Society For Engineering Education (ASEE) Southeastern Section conference,2018
Marcus Muller and Michael Botsch, Dennis Bohml ander, Wolfgang Utschick, "Machine learning based prediction of crash severity distributions for mitigation
strategies" Journal of Advances in Information Technology, vol. 9, 2018.
Sullivan hue, Christophe Hurlin, Sessi Tokpavi," Machine Learning for Credit Scoring: Improving Logistic Regression with Non Linear Decision Tree Effects" Research Gate, 2018
S M Hasan Mahmud, Md Altab Hossin, Md. Razu Ahmed, Sheak Rashed Haider Noori, Md Nazirul Islam Sarkar,"Machine Learning Based Unified Framework for Diabetes Prediction" International Conference on Big Data Engineering and Technology, BDET, 2018
Akm Ashiquzzaman, Abdul Kawsar Tushar,Md. Rashedul Islam,Jong-Myon Kim," Reduction of Over-fitting in Diabetes Prediction Using Deep Learning Neural Network" IT Convergence and Security, 2017
Ioannis Kavakiotis,Olga Tsave, Athanasios, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda," Machine learning and data mining methods in diabetes research" Computational and Structural Biotechnology Journal, 2017
Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah," Prediction of Diabetes Using Machine Learning Algorithm in healthcare" Proceedings of the 24th International Conference on Automation and Computing, 2018