International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Diabetes Diagnosis using Machine Learning Algorithms

Mitesh Warke[1], Vikalp Kumar[2], Swapnil Tarale[3], Payal Galgat[4], D.J Chaudhari[5]

[1,2,3,4]Dept of Computer Science and Engineering, Government College of Engineering Chandrapur, Chandrapur - 442403, Maharashtra, India
[5]Dept. of Computer Science and Engineering, Government College of Engineering Chandrapur, Chandrapur - 442403, Maharashtra, India

---

Abstract - *Diabetes is a chronic disease and one of deadliest diseases and also a major public health challenge worldwide. Diabetes diseases commonly stated by health professionals or doctors as diabetes mellitus (DM), which describes a set of metabolic diseases in which the person has blood sugar, either insulin production inefficient, or because of the body cell do not return correctly to insulin, or by both reason. The day is now to prevent and diagnose diabetes in the early stages. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. The discovery of knowledge from medical datasets is important in order to make effective medical diagnosis. Furthermore, predicting the disease early leads to treating the patients before it becomes critical.*

Key Words: **Diabetes, Machine Learning, Decision Tree, KNN, Naïve Bayes, Support Vector Machine**

1. INTRODUCTION

Diabetes is the third leading cause of death following diseases of heart and cancer. But with the rise of Machine Learning approaches we have the ability to find a solution to this issue. The aim of Machine Learning and Data Mining is to extract knowledge from information stored in dataset and generate clear and understandable description of patterns. We are going to develop a Diabetes Diagnosis system using Machine Learning which has the ability to predict whether the patient has diabetes or not. Furthermore, predicting the disease early leads to treating the patients before it becomes critical. Machine Learning and Data mining has the ability to extract hidden knowledge from a huge amount of diabetes-related data. This paper reviewed and analyzed the current studies on classification of Diabetes. Furthermore, the study has developed a classification model for diabetes using decision tree, Naïve Bayes, Support vector machine and k nearest neighbour Algorithm. The classification model is based on a dataset of 15000 cases collected from different National Institute of Diabetes and Digestive and Kidney Diseases. The results of the Naïve Bayes can be used by medical specialist to classify and diagnose diabetic patients. These results help the medical doctors in the classification process of diabetes.

This study follows different machine learning algorithms to predict diabetes disease at an early stage. Such as, KNN, Naïve Bayes, Decision Tree, and Support Vector machine to predict this chronic disease at an early stage for safe human life.

1.1 Problem Statement

Doctors rely on common knowledge for treatment. When common knowledge is lacking, studies are summarized after some number of cases have been studied. But this process takes time, whereas if machine learning is used, the patterns can be identified earlier. (Perner, 2006)

For using machine learning, a huge amount of data is required. There is very limited amount of data available depending on the disease. Also, the number of samples having no diseases is very high compared to number of samples actually having the disease.

1.2 Aims and Objectives

The primary aim of this project is to analyse the Diabetes Dataset and use Logistic Regression, Support Vector Machine, Naïve Bayes, K-Nearest Neighbours algorithms for prediction and to develop a prediction engine. The secondary aim is to develop a web application with following feature.

- Allow users to predict diabetes utilizing the prediction engine.

The objective is set to achieve the aims of the project through a Research on statistical models in machine learning and to understand how the algorithms works

2. RELATED WORK

Orabi et al.[4] in designed a system for diabetes prediction, whose main aim is the prediction of diabetes a candidate is suffering at a particular age. The proposed system is designed based on the concept of machine learning, by applying decision tree. Obtained results were satisfactory as the designed system works well in predicting the diabetes incidents at a particular age, with higher accuracy using Decision tree.

Pradhan et al in [15] used Genetic programming (GP) for the training and testing of the database for prediction of diabetes by employing Diabetes data set which is sourced from UCI repository. Results achieved using Genetic Programming [16], [17] gives optimal accuracy as compared to other implemented techniques. There can be significant improve in accuracy by taking less time for classifier generation. It proves to be useful for diabetes prediction at low cost.

3. METHODOLOGY

This study aims to propose a new model for diabetics classification. Numerous algorithms and different approaches have been applied, such as traditional machine learning algorithms, ensemble learning approaches and association rule learning in order to achieve the best classification accuracy.

The methods employed in this research are split by the four main phases of the research work, which are the problem formulation phase, the dataset collection phase, and the experimentation phase and the results summarizing.

This research started with formulating the research problem that is reviewing of the literature and formulating of the research problem. After the research problem formulation, this research identified the scope of the research, the objectives, and limitations of the research procedure.

The second phase of the study is the dataset collection. The dataset items were collected from National Institute of Diabetes and Digestive and Kidney Diseases.

The third phase of the study was the data preparation which included:

- Converting Data to Appropriate format
- Data Preprocessing
- Use Machine Learning to manipulate Data

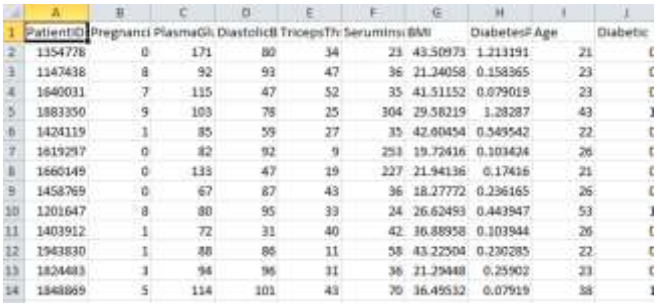
In the experimentation phase several experiments were conducted and results were collected.

3.1 Dataset

Dataset has been obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. Diabetes dataset has 9 attributes in total. All the person in records are females and the number of pregnancies they have had has been recorded as the first attribute of the dataset. The dataset contains 15000 records of female patients.

Diabetes dataset has 9 attributes in total. All the person in records are females and the number of pregnancies they have had has been recorded as the first attribute of the dataset. Second is the value of Plasma glucose concentration a 2 hours in an oral glucose tolerance test and then is the Diastolic blood pressure (mm Hg), fourth in line is the Triceps skin fold thickness (mm), then is the 2-Hour serum insulin (mu U/ml), sixth is Body mass index (weight in kg/ (height in m) ^2) and then seventh is the Diabetes pedigree function and the second last value is the that of the Age (years). The ninth column is that of the Class variable (0 or 1), 0 for no diabetes and 1 for the presence.

The Software used to visualize the entire dataset is “Jupyter Notebook” from Anaconda Navigator and the programming language used id Python 3.6.



	A	B	C	D	E	F	G	H	I	J
1	PatientID	Pregnanci	PlasmaGl	DiastolicB	TricepsTh	SerumIns	BMI	DiabetesP	Age	Diabetic
2	1354778	0	171	80	34	23	43.50973	1.213191	21	0
3	1147438	8	92	93	47	36	21.24058	0.158365	23	0
4	1640031	7	115	47	52	35	41.51152	0.079019	23	0
5	1883350	9	103	78	25	304	29.58219	1.28287	43	1
6	1424119	1	85	59	27	35	42.60454	0.549542	22	0
7	1619297	0	82	92	9	253	19.72416	0.103424	26	0
8	1660149	0	133	47	19	227	21.94136	0.17416	21	0
9	1458769	0	67	87	43	36	18.27772	0.236165	26	0
10	1201647	8	80	95	33	24	26.62493	0.443947	53	1
11	1403912	1	72	31	40	42	36.88958	0.103944	26	0
12	1943830	1	88	86	11	58	43.22504	0.230285	22	0
13	1824483	1	94	96	31	36	21.29448	0.25902	23	0
14	1848869	5	114	101	43	70	36.49532	0.07919	38	1

Figure: Representation of data used for solving problems

Table 3.1 The Diabetes Dataset Features.

Sr.No.	Attribute	Description	Type
1.	Pregnancies	Number of times pregnant	Numeric
2.	PlasmaGlucose	Plasma glucose concentration of 2 hours in an oral glucose tolerance test	Numeric
3.	DiastolicBloodPressure	Diastolic Blood Pressure in mmHg	Numeric
4.	TricepsThickness	Triceps Skin Fold Thickness measured in mm.	Numeric
5.	SerumInsulin	2-Hour serum insulin measured in µU/ml	Numeric
6.	BMI	Body Mass Calculated using: $Weight\ in\ kg(height\ in\ meter)2$	Numeric
7.	DiabetesPedigree	Diabetic Pedigree function – how likely the person is to have given their family history and other factors.	Numeric
8.	Age	Age of the Patient in years.	Numeric
9.	Diabetic	Presence of diabetes. 1 – Yes, 0 – No	Numeric

3.2 Development

The prediction engine was developed in four increments. The libraries used for the implementation of the prediction engine are provided in below Table. The increments are discussed below.

Table 10.1 Libraries used in Python

Library	Version	Project Website
numpy	1.12.1	https://docs.scipy.org/doc/numpy-dev/user/quickstart.html
scipy	0.19.0	https://www.scipy.org/install.html
pandas	0.19.2	http://pandas.pydata.org/
scikit-learn	0.18.1	https://github.com/scikit-learn/scikit-learn
matplotlib	2.0.0	https://matplotlib.org
seaborn	0.7.1	https://github.com/mwaskom/seaborn
flask	0.12.1	http://flask.pocoo.org/
wtforms	2.1	https://wtforms.readthedocs.io/en/latest/

3.2.1 First Increment

The first iteration was conducted using "Jupyter Notebook". The objectives of the first increment were.

- To visualize the dataset
- To find any correlation between features
- To find the accuracy of the prediction with different algorithms
- To create the general workflow of the prediction task

Major code snippets of the increment

Decision Tree Classifier

```
In [45]: # Decision Tree Classifier Overview

X_train, X_test, y_train, y_test = train_test_split(X,Y, stratify=Y, random_state=42)
tree = DecisionTreeClassifier(random_state=0)
tree.fit(X_train, y_train)

print('Accuracy on the training subset: {:.3f}'.format(tree.score(X_train, y_train)))
print('Accuracy on the test subset: {:.3f}'.format(tree.score(X_test, y_test)))

Accuracy on the training subset: 1.000
Accuracy on the test subset: 0.700

In [46]: tree = DecisionTreeClassifier(max_depth=4, random_state=0)
tree.fit(X_train, y_train)

print('Accuracy on the training subset: {:.3f}'.format(tree.score(X_train, y_train)))
print('Accuracy on the test subset: {:.3f}'.format(tree.score(X_test, y_test)))

Accuracy on the training subset: 0.973
Accuracy on the test subset: 0.688
```

Major visualizations from first increments

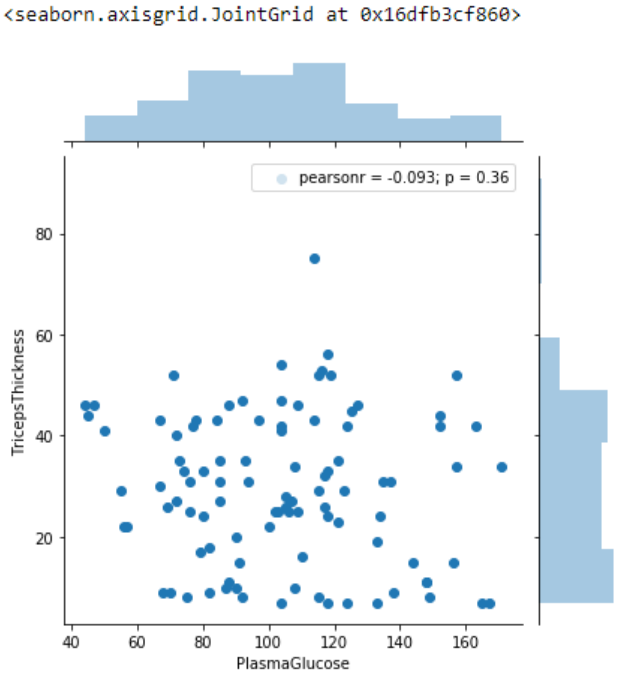


Figure 3.2.1 Scatter plot of Diabetes Disease Dataset

Figure 3.2.1 shows the scatter plot of all the features present in the "Diabetes Disease Dataset". The visualization was created in order to study the distribution of data and find any outliers.

3.2.2 Second Increment

The achievements of this increment are given below.

Different algorithms were extracted to functions.

- Functionality to store and load trained models was developed.
- Functionality to show different types of scores for the trained models was developed.

Figure 3.2.2 Prediction using Naive Bayes (Second Increment)

3.2.3 Third Increment

The third increment introduced web API for predicting diabetes. This increment utilized Flask for web framework.

Major code snippets of third increment

```
from flask import Flask,render_template,url_for,request
import pickle
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import GaussianNB
from sklearn.externals import joblib

app = Flask(__name__)

@app.route('/')
def home():
    return render_template('home.html')

@app.route('/predict',methods=["POST","GET"])
def predict():
    dataset = pd.read_csv('data/Diabetes.csv')
    x = dataset.iloc[:, :-1].values
    y = dataset.iloc[:, 8].values
```

Results of third increment

Figure 3.2.3 Running Prediction Engine

Figure 3.2.3 Prediction Response of Heart Disease

4. EXPERIMENTATION RESULTS

Table-4 represents different performance values of all classification algorithms calculated on various measures. From Table-4 it is analyzed that Naive Bayes showing the maximum accuracy. So the Naive Bayes machine learning classifier can predict the chances of diabetes with more accuracy as compared to other classifiers.

Table 4. Accuracy Measures

Classification Algorithms	Precision
Naive Bayes	0.72
Decision Tree	0.68
SVM	0.62
KNN	0.66

5. CONCLUSIONS

This project presented a comparison of Naïve Bayes classifier with other linear classifiers such as Logistic Regression, Support Vector Machines and K-Nearest Neighbours. Overall Naïve Bayes outperformed every other classifier but at the cost of being computationally expensive. K-Nearest Neighbours performed as good as Naïve Bayes with far less computational requirement.

Classification with Naïve Bayes shows the best accuracy of 0.72%.

The solution (web application) provided is a workable solution for the data problem. Currently there is a static prediction engine that serves prediction results for one disease. There is a possibility of extending the system, to allow end-users to write their own prediction engine, execute it and publish it.

In summary, the primary and secondary aims of the project have been achieved but there is still room for improvement and further enhancement.

ACKNOWLEDGEMENT

First of all I would like to thank all my co-authors for all the support. And also I would like to express attitude to the reviewer of the paper and their value able suggestion. Finally I would like to thank to my friends and parents for their support.

REFERENCES

[1]Kumari,V.A., Chitra,R., 2013. Classification of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications(IJERA)www.ijera.com3,1797–1801.

[2]NaiArun,N.,Moungmai,R.,2015.ComparisonofClassifiersfortheRiskofDiabetesPrediction.ProcediaComputerScience69,132–142.doi:10.1016/j.procs.2015.10.014.

[3]NaiArun,N.,Sittidech,P.,2014.EnsembleLearningModelforDiabetesClassification.AdvancedMaterialsResearch931932,14271431.doi:10.4028/www.scientific.net/AMR.931-932.1427.

[4]Orabi,K.M.,Kamal,Y.M.,Rabah,T.M.,2016.EarlyPredictiveSystemforDiabetesMellitusDisease,in:IndustrialConferenceonDataMining,Springer.Springer.pp.420–427.

[5]Perveen,S.,Shahbaz,M.,Guergachi,A.,Keshavjee,K.,2016.PerformanceAnalysisofDataMiningClassificationTechniquestoPredictDiabetes.ProcediaComputerScience82,115121.doi:10.1016/j.procs.2016.04.016.

[6]Pradhan,P.M.A.,Bamnote,G.R.,Tribhuvan,V.,Jadhav,K.,Chabukswar,V.,Dhobale,V.,2012.AGeneticProgrammingApproachforDetectionofDiabetes.InternationalJournalOfComputationalEngineeringResearch2,91–94.

[7]Priyam,A.,Gupta,R.,Rathee,A.,Srivastava,S.,2013.ComparativeAnalysisofDecisionTreeClassificationAlgorithms.InternationalJournalofCurrentEngineeringandTechnologyVol.3,334–337.doi:JUNE2013,arXiv:ISSN2277-4106.

[8]Ray,S.,2017.6EasyStepstoLearnNaiveBayesAlgorithm(withcodeinPython).

[9]Rish,I.,2001.AnempiricalstudyofthenaiveBayesclassifier,in:IJCAI2001workshoponempiricalmethodsinartificialintelligence,IBM.pp.41–46.

[10]Sharief,A.A.,Sheta,A.,2014.DevelopingaMathematicalModeltoDetectDiabetesUsingMultigeneGeneticProgramming.InternationalJournalofAdvancedResearchinArtificialIntelligence(IJARAI)3,54–59.doi:doi:10.14569/IJARAI.2014.031007.

[11]Sisodia,D., Shrivastava,S.K., Jain,R.C., 2010. ISVM for face recognition. Proceedings-2010 International Conference on Computational Intelligence and Communication Networks, CICN2010, 554–559 doi:10.1109/CICN.2010.109.

[12]Sisodia,D.,Singh,L.,Sisodia,S.,2014.FastandAccurateFaceRecognitionUsingSVMandDCT,in:ProceedingsoftheSecondInternationalConferenceonSoftComputingforProblemSolving(SocProS2012),December28-30,2012,Springer.pp.1027–1038.

[13] Yang Guo , Guohua Bai , Yan Hu School of computing Blekinge Institute of Technology Karlskrona, Sweden, "Using Bayes Network for Prediction of Type-2 Diabetes"

[14] Beckles GLA, Thompson-Reid PE, editors. Diabetes and Women's Health Across the Life Stages: A Public Health Perspective. Atlanta:

[15] Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1,763–770. doi:10.1007/978-3-319-11933-5.

[16]Sharief,A.A.,Sheta,A.,2014.DevelopingaMathematicalModeltoDetectDiabetesUsingMultigeneGeneticProgramming.InternationalJournalofAdvancedResearchinArtificialIntelligence(IJARAI)3,54–59.doi:doi:10.14569/IJARAI.2014.031007.

[17]Pradhan,P.M.A.,Bamnote,G.R.,Tribhuvan,V.,Jadhav,K.,Chabukswar,V.,Dhobale,V.,2012.AGeneticProgrammingApproachforDetectionofDiabetes.InternationalJournalOfComputationalEngineeringResearch2,91–94.