Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776720

PREDICTION AND ANALYSIS OF DIABETES USING MACHINE LEARNING

Avantika Mahadik
Research Scholar, Pacific(PAHER) University Udaipur
avantika_mahadik@rediffmail.com

Dr. Prashant Sharma
Associate Professor, Department of Computer Science and Engineering, Pacific(PAHER) University Udaipur, prashant.sharma@pacific-it.ac.in

Dr. Vaibhav Narawade
Professor, Department of Computer Engineering, Ramrao Adik Institute of Technology, Navi Mumbai, vaibhav.narawade@rait.ac.in

Abstract - Diabetes is a chronic condition that results from the body's resistance to or the pancreas' inability to effectively use the insulin it produces. Insulin, a peptide hormone, was responsible for regulating blood sugar. Repeated episodes of hyperglycemia, also known as high blood glucose or elevated blood sugar, are caused by hysterical diabetes and may cause severe damage to a variety of unique human body systems, including the nervous and cardiovascular systems. Long-term diabetic nerve, eye, renal, vascular, cardiovascular, and visual impairments are real. Adults with diabetes are two to three times more likely to have a heart attack or stroke. The likelihood of a negative result from several viral infections, including COVID-19, is raised in people with diabetes. One in five of the more than 58 million persons who live with diabetes are unaware of their condition. Different diseases are identified using machine learning methods, including Decision Trees (DT), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). The use of machine learning algorithms can result in quick and accurate disease prediction. One of the well-liked machine learning techniques in the medical industry is the decision tree, which has strong categorization capabilities. The most important risk factors for prediabetes were discovered to be age, waist-hip ratio (WHR), BMI, systolic and diastolic blood pressure, and a family history of diabetes. While the classification accuracy of the images produced by both methods is satisfactory, the SVM greatly outperforms the KNN in terms of classification speed and accuracy. SVM offered 98% accuracy, which is higher than DT (92.4%) and KNN (93.94%). Glucose plays a major role in diabetes.
Index Terms - Prediction, Diabetes, Machine Learning.

I. INTRODUCTION

The worldwide healthcare system may face a catastrophe because of diabetes, a long-term ailment. The metabolic disorder known as diabetes [1], or diabetes mellitus, is characterized by persistently high levels of blood sugar. The number of people living with diabetes is expected to rise from the current 522 million in 2022 (up from 108 million in 2000) [2]. As a whole, the rate of increase has been higher in low and medium income

nations than in high income ones. Amputations of the lower limbs, heart attacks, strokes, and even death from kidney failure are all linked to diabetes. Between 2000 and 2022, the age-specific death rate due to diabetes rose by 4 percent. It is predicted that by 2022, 2.5 million individuals will lose their lives to renal disease caused by diabetes [3]. Having a healthy diet, being physically active every day, keeping your weight within a healthy range, and not smoking are all proven ways to reduce your risk of developing diabetes or postpone its onset. Diet, exercise, medication, regular screening, and treatment for complications are all effective ways to control diabetes [4] and delay or avoid its consequences.

Prediabetes is the term when blood sugar levels are higher than normal, but not high enough to be classified as diabetes [5].

MACHINE LEARNING

There are various corporate and professional processes, as well as our everyday lives, that have benefited from the development of modern innovations like machine learning. It's a branch of artificial intelligence (AI) [5-6] that studies how to program smart computers to learn from pre-existing datasets by using statistical methods. A user may provide a large amount of data into a machine learning and deep learning algorithm [7-8], and the system will draw inferences and make recommendations based only on that data.

Decision Tree: It is a form of supervised machine learning that merely states the nature of the input and the expected result. In this, data is continually divided based on a particular parameter [9]. Decision nodes and leaves are the two components that can be used to explain the tree. This subfield of AI employs statistical approaches to provide smart computers with the ability to learn from data they've already been fed. A piecewise constant approximation of a tree can be thought of.

SVM: After training with learning samples, SVM can be used to predict data [10], and it is important to utilize parameter settings for support vector machines (SVMs) using the Particle Swarm Optimization (PSO) technique. SVM attempts to locate a line or hyper-plane in an n-dimensional space that serves as a boundary between the two groups. Then, depending on whether the new point is on the plus or minus side of the hyper-plane, it is classified according to the forecasting classifications. Classification and regression on linear and non-linear data are provided. SVMs are used because they have the potential to reveal complex relationships in your data without needing extensive human modification [11].

K-Nearest Neighbour (KNN) It is a classification algorithm; it should not be confused with k-Means, which have completely different applications. Unsupervised clustering technique K-Means divides data into k groups when given a set of data, where K is a positive integer. In contrast to unsupervised algorithms, supervised algorithms require training data. An illustration of a supervised classification algorithm is K-Nearest Neighbour. In this article, we'll utilise KNN to determine a person's likelihood of having diabetes [12-13] based on a variety of other health indicators. Using KNN, we may infer to which category a given data

point most likely belongs. When pitted against the most exact models, the KNN algorithm holds its own due to the high quality of its predictions. Therefore, the KNN approach might be used in circumstances when high accuracy is required yet a comprehensible model is unnecessary. The precision of the projections is dependent on the distance measurement.

I. DIAGNOSIS AND TREATMENT

Academics have created and used a number of data processing methods to categories and forecast symptoms in medical data. Rapid diagnosis [14] may be achieved at a low cost by blood glucose testing. As well as lowering blood sugar levels, a nutritious diet, frequent exercise, and avoiding other risk factors for blood vessel damage are used [15] to manage diabetes. In order to avoid issues, quitting smoking is essential.

The following initiatives can be implemented with little to no additional cost in poor and middle-income countries:

Regulation of blood sugar, particularly for type 1 diabetes. Foot care, type 2 diabetes treatment choices, oral medications for type 2 diabetic therapy, type 1 diabetes treatment options, blood pressure control, and type 1 diabetes treatment [16-17] options (Patient self-care involves regular medical checkups and foot care (which includes keeping feet clean, donning appropriate footwear, and having any ulcers on the feet treated by a doctor).

Retinal screening and treatment, as well as management of blood lipids (cholesterol) and early detection of diabetic kidney disease are further methods of reducing healthcare expenditures [18-21].

II. **DIABETES PREDICTION USING MACHINE LEARNING**

Based on the attributes, we will be able to determine whether the patient has diabetes or not. It involves the following actions:

1. Data analysis (DA): In this step, one learns how the data analysis phase of a data science life cycle is carried out.

2. Exploratory data analysis (EDA): EDA is one of the most crucial phases of a data science project, and it requires the ability to draw conclusions from visualizations and data analysis.

3. Model building (MB): In this step, we'll use 4 ML models before selecting the one that performs the best overall.

4. Saving model (SM): Using pickle to create predictions from actual data, saving the best model [22-25].

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

sns.set()

from mlxtend.plotting import plot_decision_regions
import missingno as msno
from pandas.plotting import scatter_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import classification_report
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline

diabetes_df = pd.read_csv('diabetes.csv')
      diabetes_df.head()


Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
          dtype='object')
```

TABLE 1 SHOWİNG OUTCOMES WİTH SELECTED PARAMETERS

Pregnancies	Blood Pressure	Skin Thickness	BMI	Outcome
6	72	35	33.6	1
1	66	29	26.6	0
8	64	0	23.3	1
1	66	23	28.1	0
0	40	35	43.1	1
5	74	0	25.6	0
3	50	32	31	1
10	0	0	35.3	0
2	70	45	30.5	1
8	96	0	0	1
4	92	0	37.6	0
10	74	0	38	1

10	80	0	27.1	0
1	60	23	30.1	1
5	72	19	25.8	1
7	0	0	30	1
0	84	47	45.8	1
7	74	0	29.6	1
1	30	38	43.3	0
1	70	30	34.6	1
3	88	41	39.3	0
8	84	0	35.4	0
7	90	0	39.8	1
9	80	35	29	1
11	94	33	36.6	1
10	70	26	31.1	1
7	76	0	39.4	1
1	66	15	23.2	0
13	82	19	22.2	0
5	92	0	34.1	0
5	75	26	36	0
3	76	36	31.6	1
3	58	11	24.8	0
6	92	0	19.9	0
10	78	31	27.6	0
4	60	33	24	0
11	76	0	33.2	0
9	76	37	32.9	1
2	68	42	38.2	1
4	72	47	37.1	1
3	64	25	34	0
7	84	0	40.2	0
7	92	18	22.7	0
9	110	24	45.4	1
7	64	0	27.4	0
0	66	39	42	1
1	56	0	29.7	0
2	70	27	28	0
7	66	32	39.1	1
7	0	0	0	0
1	80	11	19.4	0
1	50	15	24.2	0
5	66	21	24.4	0
8	90	34	33.7	1

7	66	42	34.7	0
1	50	10	23	0
7	68	39	37.7	1
0	88	60	46.8	0
0	82	0	40.5	0
0	64	41	41.5	0
2	0	0	0	0

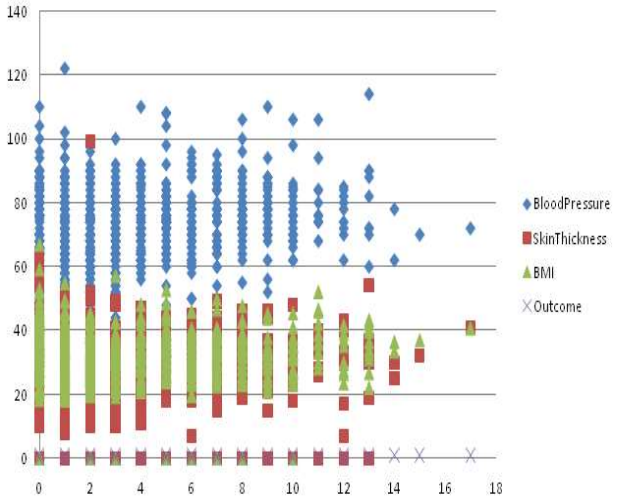

Figure: 1 Outcome with selected parameters

TABLE 2 DIABETES PEDIGREE FUNCTION WITH AGE FACTORS

Outcome	Diabetes Pedigree Function	Age	Glucose	Insulin
1	0.627	50	148	0
0	0.351	31	85	0
1	0.672	32	183	0
0	0.167	21	89	94
1	2.288	33	137	168
0	0.201	30	116	0
1	0.248	26	78	88
0	0.134	29	115	0
1	0.158	53	197	543
1	0.232	54	125	0
0	0.191	30	110	0
1	0.537	34	168	0
0	1.441	57	139	0
1	0.398	59	189	846

1	0.587	51	166	175
1	0.484	32	100	0
1	0.551	31	118	230
1	0.254	31	107	0
0	0.183	33	103	83
1	0.529	32	115	96
0	0.704	27	126	235
0	0.388	50	99	0
1	0.451	41	196	0
1	0.263	29	119	0
1	0.254	51	143	146
1	0.205	41	125	115
1	0.257	43	147	0
0	0.487	22	97	140
0	0.245	57	145	110
0	0.337	38	117	0
0	0.546	60	109	0
1	0.851	28	158	245
0	0.267	22	88	54
0	0.188	28	92	0
0	0.512	45	122	0
0	0.966	33	103	192
0	0.42	35	138	0
1	0.665	46	102	0
1	0.503	27	90	0
1	1.39	56	111	207
0	0.271	26	180	70
0	0.696	37	133	0
0	0.235	48	106	0
1	0.721	54	171	240
0	0.294	40	159	0
1	1.893	25	180	0
0	0.564	29	146	0
0	0.586	22	71	0
1	0.344	31	103	0
0	0.305	24	105	0
0	0.491	22	103	82
0	0.526	26	101	36

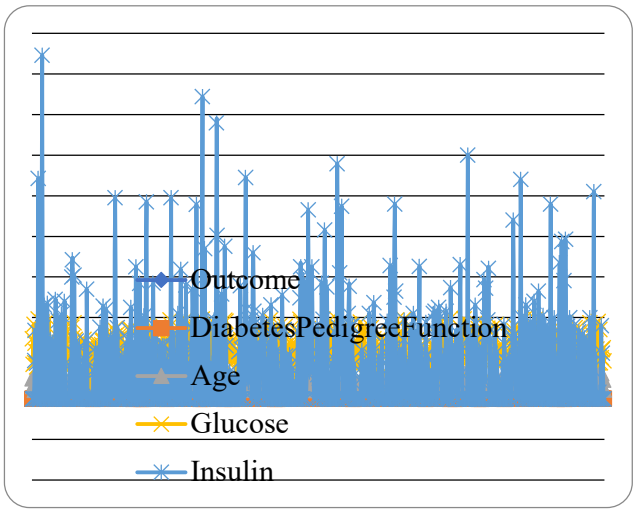

Figure: 2 Outcome of diabetes pedigree function

Null Count Analysis Plot

In a social data set [26], an invalid worth is utilized when a segment's worth is missing or equivocal. A null is not a zero value or an empty string (for character or date time data types) (for numeric data types)

p = msno.bar(diabetes_df)

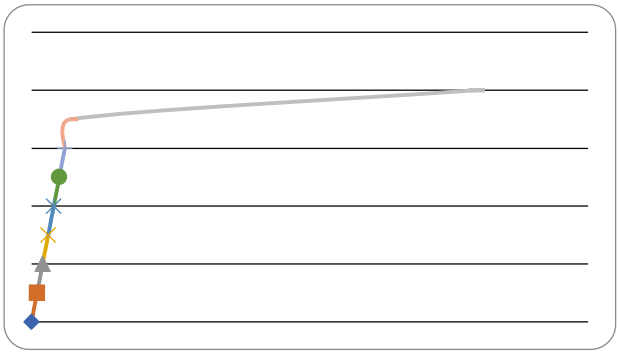

Figure: 3 Null Count Analysis

III. RESULT AND ANALYSIS

Different models, including DT, SVM, and KNN, have been created and examined. The following discussion is also provided:

Building the Model using Decision Tree:

```
From sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier()
dtree.fit(X_train, y_train)
Getting the accuracy score for Decision Tree
From sklearn import metrics
predictions = dtree.predict(X_test)
print("Accuracy Score =", format(metrics.accuracy_score(y_test,predictions)))
Output:
Accuracy Score = 0.92402834645669292
```

Building the model using Support Vector Machine (SVM)

```
from sklearn.svm import SVC
svc_model = SVC()
svc_model.fit(X_train, y_train)
Prediction from support vector machine model on the testing data
svc_pred = svc_model.predict(X_test)
Accuracy score for SVM
from sklearn import metrics
print("Accuracy Score =", format(metrics.accuracy_score(y_test, svc_pred)))
Output:
Accuracy Score = 0.9801574803149606
```

Building the model using kNN

```
from sklearn.ensemble import kNNClassifier
kNNc = kNNClassifier(n_estimators=200)
kNNc.fit(X_train, y_train)
kNNc_train = kNNc.predict(X_train)
from sklearn import metrics
print("Accuracy_Score =", format(metrics.accuracy_score(y_train, kNNc_train)))
Output:
Accuracy = 0.93930249258930
```

Plotting feature importances

```
(pd.Series(rfc.feature_importances_, index=X.columns).plot(kind='barh'))
```

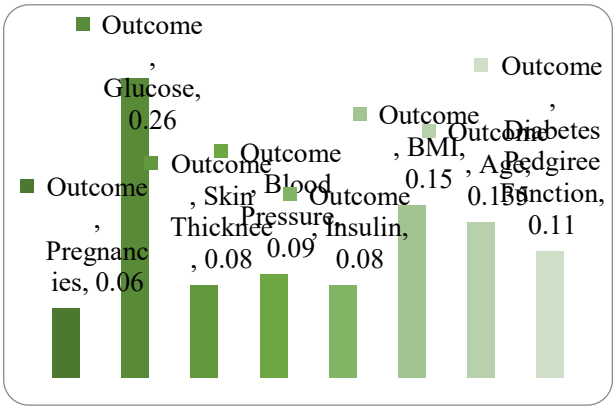

Figure: 4 Plot of various feature

The graph above makes it abundantly evident that the feature of glucose is the most significant in the prediction of diabetes.

CONCLUSİON

Using all of these medical records, we developed three machine learning models: Random Forest, Support Vector Machine, and kNN. Based on the results, it is determined that SVM offers the maximum accuracy. Additionally, it has been noted that people with diabetes are the ones who are most affected by the presence of glucose. SVM offered 98% accuracy, which is higher than DT (92.4%) and KNN (93.94%). Glucose plays a major role in diabetes.

ACKNOWLEDGEMENT

I would lıke to thank Dr. Prashant Sharma and Dr. Vaibhav Narawade for giving me the opportunity and offering me resources to complete this work.

REFERENCES

[1] American Diabetes Association; 2. Classification and Diagnosis of Diabetes: *Standards of Medical Care in Diabetes—2020*. *Diabetes Care* 1 January 2020; 43 (Supplement_1): S14–S31

[2] International Diabetes Federation. Diabetes. Brussels: International Diabetes Federation; 2019.

[3] Gregg E. W., Sattar N., & Ali, M. K. (2016). "The changing face of diabetes complications", *The lancet Diabetes & endocrinology*, *4*(6), pp. 537-547.

[4] Herman W. H., Ye W., Griffin S. J., Simmons R. K., Davies M. J., Khunti K., & Wareham N. J. (2015). "Early detection and treatment of type 2 diabetes reduce cardiovascular morbidity and mortality: a simulation of the results of the Anglo-Danish-Dutch Study of Intensive

Treatment in People With Screen-Detected Diabetes in Primary Care" (ADDITION-Europe). *Diabetes care*, *38*(8), pp.1449-1455.

[5] Kalsch J, Bechmann LP, Heider D, et al. "Normal liver enzymes are correlated with severity of metabolic syndrome in a large population based cohort" Sci Rep. 2015;5(1):pp.1–9

[6] Sanal MG, Paul K, Kumar S, Ganguly NK. "Artificial intelligence and deep learning: the future of medicine and medical practice" J Assoc Physicians India. 2019;67(4):71–3.

[7] Zhang A, Lipton ZC, Li M, Smola AJ., "Dive into deep learning" 2020.

[8] Maniruzzaman M., Kumar N., Abedin M. M., et al. (2017). "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm" *Computer methods and programs in biomedicine*, *152*, pp.23-34.

[9] Muhammad L. J., Algehyne E. A., & Usman S. S. (2020). "Predictive supervised machine learning models for diabetes mellitus" *SN Computer Science*, *1*(5), 240.

[10] Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. "Predicting diabetes mellitus using smote and ensemble machine learning approach: the henry ford exercise testing (fit) project" PLoS ONE. 2017;12(7):e0179805.

[11] Mokarram R, Emadi M. "Classification in non-linear survival models using cox regression and decision tree" Ann Data Sci. 2017;4(3):329–40.

[12] Ivanova MT, Radoukova TI, Dospatliev LK, Lacheva MN. "Ordinary least squared linear regression model for estimation of zinc in wild edible mushroom" (*Suillus luteus* (L.) roussel). Bulg J Agric Sci. 2020;26(4):863–9.

[13] Bernardini M, Morettini M, Romeo L, Frontoni E, Burattini L. "Early temporal prediction of type 2 diabetes risk condition from a general practitioner electronic health record: a multiple instance boosting approach" Artif Intell Med. 2020;105:101847'

[14] Xie, J., Liu, Y., Zeng, X., Zhang, W., & Mei, Z. (2017). "A Bayesian network model for predicting type 2 diabetes risk based on electronic health records" *Modern Physics Letters B*, *31*(19-21), 1740055..

[15] Hertroijs DFL, Elissen AMJ, Brouwers MCGJ, Schaper NC, Kohler S, Popa MC, Asteriadis S, Hendriks SH, Bilo HJ, Ruwaard D, et al. "A risk score including body mass index, glycated haemoglobin and triglycerides predicts future glycaemic control in people with type 2 diabetes" Diabetes Obes Metab. 2017;20(3):681–8.

[16] Cole SR, Chu H, Greenland S. "Maximum likelihood, profile likelihood, and penalized likelihood: a primer" Am J Epidemiol. 2013;179(2):252–60.

[17] Brisimi TS, Xu T, Wang T, Dai W, Paschalidis IC. "Predicting diabetesrelated hospitalizations based on electronic health records" Stat Methods Med Res. 2018;28(12):3667–82.

[18] Moher D, Liberati A, Tetzlaff J, Altman DG. "Preferred reporting items for systematic reviews and meta-analyses: the prisma statement" PLoS Med. 2009;6(7):e1000097

[19] Kitchenham B, Brereton OP, Budgen D, Turner M, Bailey J, Linkman S. "Systematic literature reviews in software engineering—a systematic literature review" Inf Softw Technol. 2009;51(1):7–15

[20] Sambyal N, Saini P, Syal R. "Microvascular complications in type-2 diabetes: a review of statistical techniques and machine learning models." Wirel Pers Commun. 2020;115(1):1–26.

[21] Islam MM, Yang H-C, Poly TN, Jian W-S, Li Y-CJ. "Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: a systematic review and meta-analysis" Computer Methods Programs Biomed. 2020;191:105320.

[22] Chaki J, Ganesh ST, Cidham SK, Theertan SA. "Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review" J King Saud Univ Comput Inf Sci. 2020.

[23] De Silva, K., Lee, W. K., Forbes, A., Demmer, R. T., Barton, C., & Enticott, J. (2020). "Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis" *International journal of medical informatics*, *143*, 104268.

[24] Press, G. (2016). Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. *Forbes, March*, *23*, 15.

[25] Prabhu P, Selvabharathi S. "Deep belief neural network model for prediction of diabetes mellitus" In: 2019 3rd international conference on imaging, signal processing and communication (ICISPC). 2019

[26] Albahli, S. (2020). "Type 2 machine learning: an effective hybrid prediction model for early type 2 diabetes detection" *Journal of Medical Imaging and Health Informatics*, *10*(5), 1069-1075.