Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Diabetes Prediction Using KNN and Decision Tree: Leveraging Machine Learning for Early Diagnosis and Personalized Care

Tanguturu SP Madhuri[1], Dr. G. S Raghavendra[2]

[1]*Research scholar CSE at Koneru Lakshmaiah Education Foundation, Deemed to be University, India, 2312031034@kluniversity.in*
[2]*Assistant Professor CSE at Koneru Lakshmaiah Education Foundation, Deemed to be University, India, g.raghavendra@klh.edu..in*

This study focuses on predicting diabetes using Machine learning algorithms. Machine Learning offers advantage of early diagnosis and individual treatment plans, improves patient out comes. In this research, KNN (K nearest neighbors) is applied to predict diabetes by using key health metrics like glucose levels, blood pressure, BMI. The data is processed to separate features from target variable. Additionally a decision tree classifier is developed. Decision tree can be applied for unseen data also. The performance is measured using metrics like accuracy precision, recall and f1 score to show its effectiveness. The combination of these methods demonstrates their potential in early diagnosis, aiding in clinical decision making and personalized treatment planning. Diabetes prediction using KNN and Decision Tree algorithms can be applied to small data sets such as 10 or 20 data entries to explore their effectiveness. Limited data may impact prediction accuracy but provides insight to KNN's ability to handle small sample sizes. With 20 data entries Decision Tree is trained and tested allowing for more complex decision boundaries and improved accuracy. This analysis highlights both algorithm's potential in early diabetes prediction emphasizing the adaptability of Machine learning to various data sizes for personalized health care.

Keywords: accuracy, Decision Tree, features, KNN.

1. Introduction

DIABETES is considered a chronic disorder or chronic disease that is identified by abnormal blood glucose levels caused by ineffective utilization or insufficient production of insulin. This is considered chronic in nature because this disease requires continuous monitoring and, if not monitored, this can even develop into more complex disease. Uncontrolled diabetes may affect heart and blood vessels, eyes, kidneys, nerves, gastro intestinal track, gums and teeth. Uncontrolled diabetes also results in long-term damage to several parts of the body and can even cause stroke, hypertension, and cardiovascular disease. Diabetes is often classified into type 1 diabetes mellitus, type 2 diabetes mellitus, and gestational diabetes. Type 2 diabetes mellitus requires monitoring of the daily activity of the person as this type of disease occurs mostly because of the changes in lifestyle/activity, whereas type 1 diabetes mellitus is insulin-dependent diabetes and occurs mainly because of the variations in the insulin levels. Type 2 diabetes is often linked with low physical activity levels and increasing age. So, there is a need for continuous monitoring to avoid the complications due to diabetes.

In this paper two Machine learning algorithms are used namely, KNN and Decision Tree(DT). Combining two algorithms is a powerful approach, both strategies can be used. KNN's approach is non parametric, Decision Tree (DT) approach is interpretable and intuitive. KNN algorithms finds k most similar instances or neighbors in the given training data. Whereas Decision Tree (DT) splits the data set into subset based on feature values. It (DT) creates tree like structure of decisions. DT can handle both numerical and categorical data. By combing two approaches (KNN and DT) a balance between local pattern recognition of KNN and global interpretable structure of decision tree can be achieved. This kind of hybrid approach enhances predictive accuracy, reduce overfitting.

2. Related Work

In building a predictive model for diabetes, you need to select features that are informative and likely to contribute to predict accurately. Here's a list of commonly used features for diabetes prediction:

1. Glucose Levels: Blood samples collected during fasting, along with glucose levels measured during that period, are key indicators of diabetes. Elevated glucose levels indicate impaired glucose metabolism, which is a hallmark of diabetes.
2. Body Mass Index (BMI): BMI is a measure of body fat based on height and weight. Higher BMI values are strongly associated with an increased risk of type 2 diabetes.
3. Age: Advancing age is a major risk factor for diabetes. As age progresses, insulin sensitivity and pancreatic function may decline.
4. Family History: A family history of diabetes, especially in first-degree relatives, increases the risk of developing the condition due to genetic factors.
5. Waist Circumference: Abdominal obesity, indicated by an increased waist circumference. Larger waist circumference is a major risk factor due to it's strong link with insulin resistance.

6. Blood Pressure: Hypertension (high blood pressure) is both a risk factor for and a complication of diabetes. Monitoring blood pressure levels can help assess cardiovascular risk in individuals with diabetes.
7. Physical Activity: Regular physical activity is protective against diabetes, as it improves insulin sensitivity and helps maintain a healthy weight.
8. Dietary Habits: Poor dietary habits, characterized by a high intake of processed foods, sugars, and saturated fats, increase the risk of diabetes. Conversely, a diet rich in fruits, vegetables, whole grains, and lean proteins can lower the risk.
9. Cholesterol Levels: Abnormal lipid profiles, including high LDL cholesterol and triglycerides, and low HDL cholesterol, are associated with insulin resistance and an increased risk of diabetes.
10. Gestational Diabetes History: Women who have had gestational diabetes during pregnancy are at increased risk of developing type 2 diabetes later in life.
11. Ethnicity: Certain ethnic groups, such as African Americans, Hispanic Americans, Native Americans, and Asian Americans, have a higher prevalence of diabetes compared to others.
12. Sleep Patterns: Poor sleep quality, insufficient sleep, and sleep disorders like sleep apnea have been linked to an increased risk of diabetes due to their effects on metabolism and insulin sensitivity.

3. History

Diabetes is an old enemy, being mentioned in ancient Indian, Greek and Roman medical texts. About 2,500 years ago ayurvedic authorities Sushruta and charaka noted that victims frequently passed large amounts of sugar laden urine which attracted ants. Hence, they called the disease madhu meha. The history of diabetes prediction has evolved significantly over the years. Early recognition of diabetes relied on clinical symptoms and urine tests for glucose, dating back to ancient civilizations. The 20th century saw the development of blood glucose testing and the discovery of insulin, revolutionizing diabetes management. In the 1970s, the introduction of glycated hemoglobin (HbA1c) testing provided a longer-term measure of blood sugar control. With the advent of computers and data analysis in the late 20th century, predictive models using demographic, lifestyle, and genetic data emerged. The early 2000s saw the incorporation of machine learning and artificial intelligence, enabling more accurate and individualized risk assessments. Recent advances include the use of continuous glucose monitors (CGMs) and real-time data analytics. Today, diabetes prediction involves complex algorithms integrating genetic, biomarker, and lifestyle information to identify at-risk individuals and guide preventive measures.

4. Literature Review

4.1 Machine-Learning-Based Diabetes Prediction Using Multisensor Data

Diabetes is a chronic disease that requires ongoing monitoring. Wearable sensor devices have made this process easier by integrating sensors capable of tracking various physiological signals, such as electrocardiograms (ECG) and accelerometer (ACC) data. Several datasets have been made available for diabetes detection and prediction, including MIMIC I from the PhysioNet2 platform and datasets from the University of California Irvine (UCI) machine learning repository. The D1NAMO dataset is also widely used. Machine learning algorithms are applied to these datasets to determine the best-performing model when using individual sensors for diabetes prediction. The data from different sensors are given to a machine learning engine where the individual data is processed and analyzed separately. The dataset used for this work consisted of four types of health data namely glucose data, ECG data, ACC data, and breathing data. A multisensor combination using glucose, ECG, and ACC sensors gives the highest prediction accuracy of 98.2% with the XGBoost algorithm and using a 5-min window size. Multisensor combinations showed an increase of nearly 4%–5% in the diabetes prediction rates as compared to single-sensor predictions. The breathing-sensor-related data have very little influence on the prediction of diabetes.

4.2 Diabetes Prediction Using Machine Learning Algorithms and Ontology

Early detection of diabetes plays a crucial role in managing the condition effectively. Machine learning algorithms are valuable for extracting insights from large datasets, particularly in the medical field. Classification algorithms are commonly applied to categorize datasets into predefined groups and predict future outcomes based on their high accuracy and performance. Popular machine learning classification techniques such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Naive Bayes, Logistic Regression, and Decision Trees are widely used for identifying diabetes patients at an early stage.Ontology, a data representation method, is one of the most commonly adopted approaches for organizing, managing, and extracting data. In evaluating classification methods, performance metrics like precision, accuracy, and recall are employed. Notably, without the application of feature selection, ontology-based classification has demonstrated the highest accuracy.

4.3 Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers

In recent years, plenty of methods have been proposed and published for diabetes prediction. A new pipeline for diabetes prediction from the PIMA Indians Diabetes dataset is proposed. It consists of outlier rejection, filling missing values, data standardization, feature selection, and K-fold cross-validation. a robust framework for diabetes prediction where the outlier rejection, filling the missing values, data standardization, feature selection, K-fold cross-validation, and different Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron (MLP) were employed.

4.4 Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review

Data mining based forecasting techniques for data analysis of diabetes can help in the early detection and prediction of the disease and the related critical events such as hypo/hyperglycemia. The performance of four well known methods, namely J48 decision

tree (DT) classifier, KNN, random forests algorithm, and support vector machine (SVM), is evaluated in terms of prediction of diabetes using data samples with and without noise from the University of California Irvine (UCI) machine learning data repository. WEKA, TANAGRA, and MATLAB. A comparative study of nine different techniques is conducted for diabetes prediction using the Pima Indian Diabetes Dataset (PIDD) from the UCI Machine Learning Repository.

4.5 Prediction of Diabetes Empowered With Fused Machine Learning

Machine Learning techniques are used for prediction of chronic disease . Fused machine learning techniques, incorporating Support Vector Machines (SVM) and Artificial Neural Networks (ANN), are employed in a proposed model for diabetes prediction. The model consists of two layers: a training layer and a testing layer. The dataset used is sourced from the UCI Machine Learning Repository. After preprocessing, the data is utilized to train both SVMs and ANNs for prediction tasks. The proposed fuzzy decision system achieves an accuracy of 94.87%, surpassing the performance of existing systems.

4.6 Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction

Cloud Computing, Internet-of-Things (IoT), Artificial Intelligence (AI) and Machine Learning have increased the number of internet-connected smart devices, such as wearable sensors, and revolutionized the way the medical industry operates. y. IoMT(Internet of Medical Things with smart sensors, smart devices and smart communication protocols facilitated the development of various smart systems in the field of healthcare. The ELSA dataset is used. Diabetes Risk Scores to apply it in parallel to the training and test dataset. A multi-objective optimization based technique since it is more robust compared to the single objective one and constructs more efficiently the classifier ensemble (WeightedVotingLRRFs), ensemble methods constitute a useful tool for predicting type 2 diabetes. The ELSA-based constructed dataset. In addition, the comparison of state of the art techniques such as XGBoost, AdaBoost or high layer DNNs would probably provide better insights regarding the predictive limitations of the constructed dataset.

5 Data Compression

For K-Nearest Neighbors (KNN), data compression can enhance efficiency by applying several techniques: normalizing data to maintain a uniform scale, using dimensionality reduction methods like Principal Component Analysis (PCA) to lower the number of features while preserving essential variance. For prediction of chronic diseases like diabetes only important features are taken into consideration. Data compression is important in decision tree algorithms because it reduces the size of the dataset, which can significantly improve the efficiency of the model training and prediction processes. By compressing data, we can lower the memory usage and computational overhead, enabling faster processing times and the ability to handle larger datasets. Compressed data can also help in reducing the complexity of the decision tree, potentially leading to simpler, more interpretable models. Additionally, it may help in mitigating overfitting by eliminating redundant or less important features, thus improving the overall generalization capability of the decision tree.

6 Calculations And Results

Sample Diabetes Data Set

ID	Glucose	BMI	Age	OutCome
1	85	29.0	31	0
2	89	31.2	33	0
3	80	25.5	29	0
4	133	33.6	50	1
5	150	30.0	45	1

Table 1

Knn Query Features:

Glucose :90

BMI: 28.0

Age:30

Calculating the Euclidean Distance:

$D = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$ (1)

1. Calculating distance to ID1

$$d = \sqrt{(90-85)^2 + (28-29.0)^2 + (35-31)^2}$$

$=\sqrt{25+1+16} \quad = \sqrt{42} \cong 6.48$

2. Calculating Distance to ID2:

$$\sqrt{(90-89)^2 + (28.0-31.2)^2 + (35-33)^2}$$

$=\sqrt{1+10.24+4}$

$=\sqrt{15.24} \cong 3.90$

3. Distance to ID3:

$$\sqrt{(90-80)^2 + (28.0-25.5)^2 + (35-29)^2}$$

$=\sqrt{100+6.25+36}$

$=\sqrt{142.25} \cong 11.92$

4. Calculating Distance to ID4:

$$\sqrt{(90-133)^2 + (28.0-33.6)^2 + (35-50)^2}$$

$= \sqrt{1849+31.6+225}$

$=\sqrt{2105.36} \cong 45.88$

5. Calculating Distance to ID5:

$\sqrt{(90-150)^2+(28.0-30.0)^2+(35-45)^2}$

$=\sqrt{3600+4+100}$ $=\sqrt{3704}$

$\cong 60.86$

Sorting Distances

ID	Distance
2	3.90
1	6.48
3	11.92
4	45.88
5	60.86

Table 2

Calculations for Decision Tree

Gini Impurity Calculation:

Gini = 1-($\sum$(pi^2) (2)

P^i is proportion of instances of class i in the set

Initial Gini Impurity

Number of instances :5

Number of class 0: 3

Number of class 1: 2

Gini = 1- $((\frac{3}{5})^2+(\frac{2}{5})^2)$

=1-(0.36+0.16)

=1-0.52

=0.48

Splitting by Glucose

Left split(Glucose≤ 100)

Instances 1,2,3 call class 0

Gini : 1-$(1-\left(\frac{3}{3}\right)^2)-1$

=1-1 =0

Right split (glucose>100)

Instances 4,5 both class 1

Gini: 1-$(\frac{2}{2})^2$)

=1-1 =0

Weighted Gini impurity for Glucose split

$\text{Gini}_{\text{Glucose}} = \frac{3}{5} * 0 + \frac{2}{5} * 0$

= 0

Splitting by BMI

Split at BMI =30

Left split BMI≤ 30

Instances 1,3,5(class 0,0,1)

Gini: 1- $((\frac{2}{3})^2 + \left(\frac{1}{3}\right)^2)$

= 1-$(\left(\frac{4}{9} + \frac{1}{9}\right)$

= 1-$\frac{5}{9} = \frac{4}{9}$

$\cong 0.44$

Right split (BMI≥ 30)

Instances 2,4 (class 0,1)

	Precision	recall	F1 score	support
0	0.73	0.89	0.80	9
1	0.86	0.67	0.75	9

Table 3

Gini: $1 - ((\frac{1}{2})^2 + (\frac{1}{2})^2)$

= 1-$(\frac{1}{2})$

= 0.5

Weighted Gini Impurity for BMI Split

$\text{Gini}_{\text{BMI}} = \frac{3}{5} \times 0.44 + \frac{2}{5} \times 0.5$

=0.26+0.2

0.464

Splitting by Age

Split Age =40

Left split(Age≤ 40)

Instances 1,2,3 (all classes 0)

Gini : $1 - ((\frac{3}{3})^2$

= 1-1= 0

Right split (Age> 40)

Instances 4,5 (both class 1)

Gini :$1 - ((\frac{2}{2})^2$

= 1-1 =0

Weighted gini Impurity for Age Split

$Gini_{Age} = \frac{3}{5} \times 0 + \frac{2}{5} \times 0$

=0

Decision Tree Structure:

1. Root Node : Glucose≤ 100:
2. if true(left branch), outcome =0 (non diabetic)
3. if false(right branch, outcome= 1(diabetic)

Results

1. Knn algorithm Applied on 100 data entries

Accuracy: 0.78

Confusion Matrix:

[[8 1]

[3 6]]

Classification Report:

Accuracy	0.78		18	
Macro avg	0.79	0.78	0.78	18
Weighted avg	0.79	0.78	0.77	18

Table 4

2. Knn algorithm applied on 20 data entries

Accuracy: 0.50

Confusion Matrix:

[[0 2]

[0 2]]

Classification Report:

	precision	Recall	f1-sore	Support
0	0.00	0.00	0.00	2
1	0.50	1.00	0.67	2

Table 5

3. Knn algorithm applied on 10 data entries

Accuracy	0.50		4	
Macro avg	0.25	0.50	0.33	4
Weighted avg	0.25	0.50	0.33	4

Table 6

Accuracy: 0.50

Confusion Matrix:

[[0 1]

[0 1]]

Classification Report:

	Precision	recall	F1-score	Support
0	0.00	0.00	0.00	1
1	0.50	1.00	0.67	1

Table 7

Accuracy	0.50		2	
Macro avg	0.25	0.50	0.33	2
Weighted avg	0.25	0.50	0.33	2

Table 8

4. Decision Tree applied on 100 data entries

Accuracy: 0.56

Classification Report:

	Precision	recall	F1-score	support
0	0.55	0.67	0.60	9
1	0.57	0.44	0.50	9

Table 9

Accuracy	0.56		18	
Macro avg	0.56	0.56	0.55	18
Weighted avg	0.56	0.56	0.55	18

Table 10

Confusion Matrix:

[[6 3]

[5 4]]

5. Decision Tree applied on 20 data entries

Accuracy: 0.50

Classification Report:

	precision	recall	F1-score	support
0	0.50	0.50	0.50	2
1	0.50	0.50	0.50	2

Table 11

Accuracy	0.50		4	
Macro avg	0.50	0.50	0.50	4
Weighted avg	0.50	0.50	0.50	4

Table 12

Confusion Matrix:

[[1 1]

[1 1]]

6. Decision Tree applied on 10 data entries

Accuracy: 0.50

Classification Report:

	precision	recall	F1-score	support
0	0.50	1.00	0.67	1
1	0.00	0.00	0.00	1

Table 13

Accuracy	0.50		2	
Macro avg	0.25	0.50	0.33	2
Weighted avg	0.25	0.50	0.33	2

Table 14

Confusion Matrix:

[[1 0]

[1 0]]

Algorithm	No. of Data Entries	Accuracy	Precision	Recall
KNN	100	0.78	0.73	0.89
KNN	20	0.50	0.50	1
KNN	10	0.50	0.50	1
DT	100	0.56	0.57	0.44
DT	20	0.50	0.50	0.50
DT	10	0.50	0	0

Table 15

7. Conclusion

Supervised learning algorithms in machine learning are extensively used for diabetes prediction. The process begins with data collection and identifying key features. In this study, the features considered include glucose levels, number of pregnancies, blood pressure,

skin thickness, insulin levels, BMI, diabetes pedigree function, and age. KNN and Decision Tree algorithms are applied to datasets containing 10, 20, and 100 entries. Both algorithms are widely used for classification tasks in healthcare due to their interpretability and accuracy. With smaller datasets, performance may be limited. However, with 100 entries, the algorithms generally show improved predictions. When new patient data is introduced, KNN calculates the distance between this data point and all other points in the dataset, identifies the k nearest neighbors, and classifies the patient based on the majority class of those neighbors. In contrast, the Decision Tree (DT) builds a tree where each node represents a feature, and each branch represents a decision rule, leading to terminal nodes (leaves) that indicate the outcome (diabetic or non-diabetic). KNN performs well with small datasets, especially when there are no irrelevant or redundant features. DT is efficient with larger datasets, and its hierarchical decision structure offers an interpretable model that is easy to visualize and understand. Both KNN and Decision Tree provide valuable insights into diabetes prediction, especially as dataset size increases.

References

1. FARRUKH ASLAM KHAN 1 , (Senior Member, IEEE), KHAN ZEB 2 , MABROOK AL-RAKHAMI 3,4, (Member, IEEE), ABDELOUAHID DERHAB 1 , AND SYED AHMAD CHAN BUKHARI5 , (Senior Member, IEEE) "Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review" date of publication February 12, 2021, date of current version March 24, 2021. Digital Object Identifier 10.1109/ACCESS.2021.3059343.
2. MD. KAMRUL HASAN 1 , MD. ASHRAFUL ALAM1 , DOLA DAS2 , EKLAS HOSSAIN 3 , (Senior Member, IEEE), AND MAHMUDUL HASAN 2 "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers" date of publication April 23, 2020, date of current version May 7, 2020. Digital Object Identifier 10.1109/ACCESS.2020.2989857.
3. Hakim El Massari, Zineb Sabouri, Sajida Mhammedi and Noreddine Gherabi " Diabetes Prediction Using Machine Learning Algorithms and Ontology" Received 11 February 2022; Accepted 12 March 2022; Publication 11 May 2022.
4. NIKOS FAZAKIS , OTILIA KOCSIS , ELIAS DRITSAS , SOTIRIS ALEXIOU, NIKOS FAKOTAKIS, (Member, IEEE), AND KONSTANTINOS MOUSTAKAS , (Senior Member, IEEE) "Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction" date of publication July 20, 2021, date of current version July 29, 2021. Digital Object Identifier 10.1109/ACCESS.2021.3098691.
5. Aditi Site , Jari Nurmi , Senior Member, IEEE, and Elena Simona Lohan , Senior Member, IEEE " Machine-Learning-Based Diabetes Prediction Using Multisensor Data" IEEE SENSORS JOURNAL, VOL. 23, NO. 22, 15 NOVEMBER 2023.
6. USAMA AHMED1,2, GHASSAN F. ISSA3 , MUHAMMAD ADNAN KHAN 1,4 , SHABIB AFTAB 2,5, (Member, IEEE), MUHAMMAD FARHAN KHAN6 , RAED A. T. SAID7 , TAHER M. GHAZAL 3,8, (Member, IEEE), AND MUNIR AHMAD 5 , (Member, IEEE) " Prediction of Diabetes Empowered With Fused Machine Learning" date of publication January 11, 2022, date of current version January 24, 2022. Digital Object Identifier 10.1109/ACCESS.2022.314209.