*Turkish Journal of Computer and Mathematics Education Vol.12 No.6(2021), 225-230*

*Research Article*

# Performance Analysis of Diabetes Mellitus Using Machine Learning Techniques

**Kandala Srujana Kumari, K.Bhargavi**

PG Scholar,Department of Computer Science and Engineering,
Associate Professor, Department of Computer Science and Engineering,
Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad, Telangana 500097

**Article History**:Received:11 november 2020; Accepted: 27 December 2020; Published online: 05 April 2021

**ABSTRACT :** Diabetes is a common disease in the human body caused by a set of metabolic disorders in which blood sugar levels are very long. It affects various organs in the human body and destroys many-body systems, especially the kidneys and kidneys. Early detection can save lives. To achieve this goal, this study focuses specifically on the use of machine learning techniques for many risk factors associated with this disease. Technical training methods achieve effective results by creating predictive models based on medical diagnostic data collected on Indian sugar. Learning from such data can help in predicting diabetics. In this study, we used four popular machine learning algorithms, namely Support Vector Machine (SVM), Naive Bayes (NB), Near Neighbor K (KNN), and Decision Tree C4.5 (DT), based on statistical data. people. adults in sugar. , preview. The results of our experiments show that the C4.5 solution tree has greater accuracy compared to other machine learning methods.

**Keywords:** Support Vector Machine, Naive Bayes, Near Neighbor K and Decision Tree C4.5

## 1. INTRODUCTION

Diabetes mellitus, also known as sugar, affects the hormone insulin, which causes abnormal carbohydrate metabolism and increases blood sugar. These blood sugar levels affect the human body in many organs, which interferes with many causes of the body, especially the blood vessels and kidneys. The causes of diabetes have not been fully elucidated and many researchers believe that heredity and environmental factors play a role. In any case, diabetes can be common in adults, as it is called adult diabetes. It is now believed that sugar is associated with aging. According to the Diabetes Association of Canada (CDA), the number of people with diabetes in Canada will increase from 2.5 million to 3.7 million between 2010 and 2020. The current world situation is no different. According to the International Diabetes Federation, the number of people with diabetes in 2013 was 382 million or 6.6% of adults worldwide. According to the World Health Organization, the number of people with diabetes is expected to increase from 376 billion to 490 billion by 2030. Besides, sugar can be an independent factor that causes mild inflammation. Diabetics are unable to cope with the risk of damage to small cells, and chronic complications of heart and heart disease are the leading causes of death. It damages small cells and cardiovascular disease rapidly leading to mental illness, nephropathy, and neuropathy. Early detection of the disease can be controlled and saved. To achieve this goal, this study focuses primarily on early diabetes prognosis, taking into account several risk factors associated with the disease. For research purposes, we collected observational data on 16 characteristics in 200 patients with diabetes. These characteristics include age, nutrition, blood pressure, vision problems, genetics, and so on. We will discuss these services and their value later. Based on these properties, we created a sugar forecasting model using different machine learning methods. Machine learning techniques to advance science by creating predictive models based on available medical diagnostic data from diabetics to achieve optimal results. Learning from such data can help in predicting diabetics. Different machine learning methods allow you to predict sugar. However, it is very difficult to choose the best forecast method based on such services. Therefore, we examined four well-known machine learning algorithms, namely Vector Machine (SVM), Naive Bayes (NB), Neighbors K-Enest (KNN), and C4.5 Declic Tree (DT) for the adult population.

## 2. LITERATURE SURVEY

### Estimating continuous distributions in Bayesian classifiers

When we created the potential distribution with the Bayesian network, we solved the problem as it evolved. Much previous work solved problems with models or assumed data created by Gaussian. In this document, we leave out conventional assumptions rather than using statistical methods to measure unmeasurable violence. For simple Baesi mediation, we present the results of experiments in a natural and artificial environment by experimenting with two methods of saturation estimation: standard calculations and models for each country distribution and Gausia; and calculate the level of language equipment without measurement. We see error



*Corresponding author: Kandala Srujana Kumari
PG Scholar,Department of Computer Science and Engineering,
Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad, Telangana 500097

reduction in a variety of natural and artificial data, and this nuclear assessment is an important tool for learning Bayesian models.

In recent years, important alternatives have emerged to create the possibility of interpreting learning data, going through machine learning methods, such as cutting down trees, and making decisions. For example, Cooper and Herskowitz (1992) began statistical calculations that determine the network structure of Bayesi results from data, while Heckerman, Geiger and Chickering (1994), Prowan and Singh (1995) and others reported the results of these fundamental methods. , Bayesi Torpedoes represents machine learning for practical reasons in the work of research: a direct relationship with errors and noise, which is a major problem in a variety of early studies. The most impressive result to date, on a much simpler and larger scale, is the intelligent robotic laboratory from Pat Langley, Stanford University, CA 94305 LangleyCCS.Stanford.EDU HTTP: // robot .stanford. Bayes division. Despite the simple opinion of the Bayesi predecessors, experiments with real data show that it is more complicated than the infectious algorithm. For example, when Clark and Niblett (1989) reported that ports were believed to be designed based on infectious methods in medicine, Langley, LBA, and Thompson (1992) found that there were four of the five domains. This surprising result has led some researchers to read about the expansion of Naive Bayes, to reduce belief in its theory, but maintain simplicity and semantic clarity. Langley and Sage (1994) involve changes that facilitate increased freedom and eliminate predictable behavior from others. Kononenko (1991) and Pazzani (1995) proposed an alternative response to this theory that integrates work records into model processes. These and similar methods form a critical line for machine learning, the purpose of which is not only to work with real data but also to find semantically clear teaching methods. One way to achieve this is to read and comment on the modern system in Bayesi, but other research programs will try to improve it by creating alternatives and eliminating ideas that will hinder their work. In this essay, we use the latter method, started by the Baye division, which assumes that digital assets are created by the Gaussian distribution. Gauss may be a reasonable estimate for most global distributions, but this is not the best way. This approach proposes another area in which we can increase and increase profits: evaluating conventional methods of moisture assessment.

## 3. OVERVIEW OF THE SYSTEM

**Existing System:**

Kavakiotis et al. Lacking tools, 10 checks work as evaluation methods in three different algorithms, and with Naive Bayes and SVM, SVM offers 84% better performance and accuracy than other algorithms. Zhen et al. Unconventional forest use, KNN, SVM, Naive Bayes, which defines the return of wood and materials, improves filtering parameters in the pre-sugar phase. Similar to .KNN, J48, ANN, ZeroR and NB are not used in different sugar groups. Pradeep et al. Distribution arbitration has not been tested. For sugar prediction and management, Huang et al. During 2000-2004, the Ulster Council and Hospital Insurance (UCHT) discussed three ways to extract data from IB1, Naive Bayes and C4.5. With the help of asset selection methods, IB1 scores and Naive Bayes produce the best results. Xue Hui Meng et al. Diabetes diagnosis uses three AR data query techniques, equipment removal and J48 user queries using real-time data collection and query data collection..

**Disadvantages:**

Distribution arbitration has not been tested.

**Proposed system:**

Machine learning techniques achieve good results in the acquisition of science by creating predictable diagnostic data models from diabetic patients. Learning from such data can help in predicting diabetics. Different machine learning methods allow you to predict sugar. However, overcoming such services is very difficult to choose the best forecasting technology. So, we used four popular machine learning algorithms for research, namely Vector Machine (SVM), Naive Bayes (NB), Neighbors K-Enest (KNN) and C4.5 Declic Tree (DT) for the population. adults wear.

**Advantages:**

We collected diagnostic data from 200 patients in health facilities, who were diagnosed with diabetes behaviors and risk factors.

We compare the effectiveness of different machine learning techniques and evaluate the expected effects on risk management.

**Modules:**

**1) Support Vector Machines:** This is one of the most popular classification methods. ???? SVM distinguishes individuals from certain classes. Recognize and distribute unused data. SVM does not ignore the noise distribution for each class. The only drawback of this algorithm is that we can do backlink analysis to find the appropriate function, and another extension is to learn subdivisions to create different parts and objects.

**2) Naive Bayes***:* Naif Dams is a popular method that is recommended for possible classification. Naive Bayes, also known as Byles Theorem, is a simple and efficient algorithm that calculates potential outcomes by calculating the frequency and data of a set of data. In real use, the idea of low conditional freedom and provides a large number of pure distribution results.

**3) K-Nearest Neighbor Algorithm***:* Neighboring K is a simple division with a retrieval algorithm that uses non-parametric methods. The algorithm registers all valid assets and assigns new assets according to fairness. They use the communication structure as a tree to determine the distance from the point of interest to the mark in the training data. The property has been introduced by neighbors. In the classification method, the value is a good number for a close neighbor. Neighbors are selected from multiple classes or property values for that issue.
**4) Decision Tree:**Decision tree is a tree that provides the same distribution method for sugar forecast. Most statistics show several areas, separated by a function called "classification". Every personal domain and service in the field is called a class. The asset properties of the class assets are determined by the internal nodes in the decision tree. The tree file value is indicated by all the attributes and attributes associated with the target value. Maximum data recovery for all properties is calculated for each node in the tree structure.
To achieve our goals, the research approach includes several steps, including data collection on diabetes and disease properties, the development of early numerical properties, machine learning of different distribution methods and initial analysis for the use of this data. We will discuss these steps a little below.
A. Dataset and Attributes
In this study, we collected data on sugar at Chittagong Medical Center (PKS) in Bangladesh. The database contains 200 patients who have behavioral or risk factors for diabetes. Table 1 shows the properties and values of each.

**Table 1: Dataset Description**

| SI No. | Attributes | Type | Values |
|---|---|---|---|
| 1 | Age (Years) | Numeric | {1 to 100} |
| 2 | Sex | Nominal | {Male, Female} |
| 3 | Weight (Kg's) | Numeric | {5 to 120} |
| 4 | Diet | Nominal | {Vegetarian, Non-Vegetarian} |
| 5 | Polyuria | Nominal | {Yes, No} |
| 6 | Water Consumption | Nominal | {Yes, No} |
| 7 | Excessive Thirst | Nominal | {Yes, No} |
| 8 | Blood Pressure (mmHg) | Numeric | {50 to 200} |
| 9 | Hyper Tension | Nominal | {Yes, No} |
| 10 | Tiredness | Nominal | {Yes, No} |
| 11 | Problem in Vision | Nominal | {Yes, No} |
| 12 | Kidney Problem | Nominal | {Yes, No} |
| 13 | Hearing Loss | Nominal | {Yes, No} |
| 14 | Itchy Skin | Nominal | {Yes, No} |
| 15 | Genetic | Nominal | {Yes, No} |
| 16 | Diabetic | Nominal | {Yes, No} |

B. Data Preprocessing
To achieve the objectives of this study, several preparation procedures were performed on diabetic data. For example, the actual value for a property number is not important for predicting sugar. So, we shift the value of the attribute number to the denomination. For example, patient age is divided into three groups: youth (10-25), adults (26-50), and older (over 50). Similarly, the patient weight falls into three groups, very low (less than 40 kg), normal (41-60 kg), and high (more than 60 kg). Finally, blood pressure is divided into three groups: normal (120/80 mm Hg), low (80 mm Hg), and high (120 mm Hg).
C. Apply Machine Learning Techniques
When the data is ready to be copied, we use the popular four-machine classification to predict sugar. That is why we provide general information about this technique.
1) Support for vector machines: This is one of the most popular classification methods proposed by J. Platt and others. elemyûn. Support Vector (SVM) is a classification of official data deletion and hyperplane isolation. SVM distinguishes individuals from certain classes. They can also identify and classify examples that are not

supported by data. SVM does not distribute written data to each class. The only extension of this algorithm is to perform repetitive analysis to find the corresponding function, and another extension is to learn subdivisions to form divisions with different objects.

2) BayiveBayive: Naive Bayes is a popular distribution method suggested by John et al. elemyûn. Bayive Bayes, also known as Bayes Theorem, is a classification of learning with these simple, useful, and versatile machines. The algorithm calculates the possible outcome by calculating the frequency and comparing the values obtained from the data set. Bayesi theorem assumes that all assets are not independent and change according to class value. The idea of conditional freedom rarely works realistically and provides good and accurate classification results.

3) The closest algorithm K: the closest neighbor K is the division and the simple conversion algorithm, namely Aha et. elemyûn. The algorithm registers all valid assets and assigns new assets according to equity. They use the communication structure as a tree to determine the distance from the point of interest to the mark in the training contact data. Properties are shared by neighbors. In the classification method, the value is a good number for a close neighbor. Neighbors are selected from multiple classes or property values for that issue.

4) Decision tree: The decision tree is a tree that provides an effective distribution channel for sugar forecasting. Most statistics indicate multiple areas, separated by a function called "classification". Every personal domain and service in the field is called a class. The input properties of the class properties are defined by the internal nodes in the decision tree. The tree file value is indicated by all the attributes and attributes associated with the target value. For all functions, return the above information for each node in the calculated tree structure. Machine learning technologies such as ID3, J48, C4.5, C5, CHAID, and QART have many popular solutions for sugar data distribution trees. In our study, the C4.5 tree solution algorithm was selected to measure the effect of sugar data. C4.5 provides the complex functions of the ID3 solution tree algorithm proposed by Ross Quinlan et al. elemyûn. The C4.5 solution tree uses training data as ID3, which is part of the learning task. Training methods can be used to find medical data to predict the meaning of a specific decision. In each branch of the tree, C4.5 selects the characteristic values of the data and effectively distributes the read data up to the data level that enriches the class. A tree is created by receiving general information. Conventional data uses are selected to determine the largest warranty based on the characteristics of the owner and evaluated according to the decision of tree C4.5

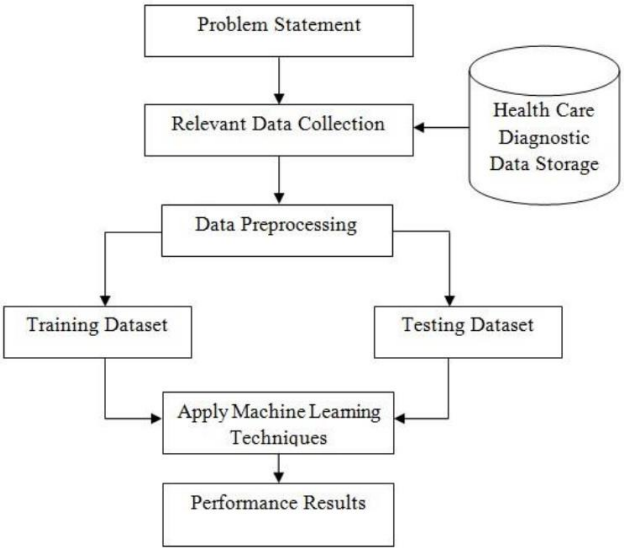



**Figure 1: An overview of the overall process**

Figure 1 provides a summary of all our manufacturing processes. As shown in Figure 1, if any problems are detected, we can download the correct data from Storage Identification. Then we process the data to create a predictable model. Then we use many of the machine learning techniques discussed above in the training package. Finally, a set of test data measures the effectiveness of the methods used to select diabetics.

**FUNCTIONAL REQUIREMENTS**

**User**

- Load data
- Data analysis
- Data preprocessing
- Model building
- Prediction

## 4. OUTPUT SCREEN SHOTS

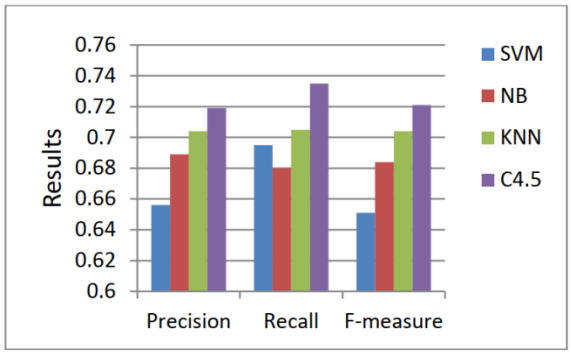



**Figure 2: Predictions Results of Various Machine Learning Techniques**

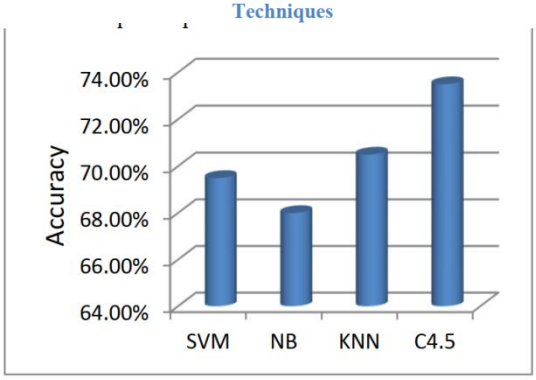


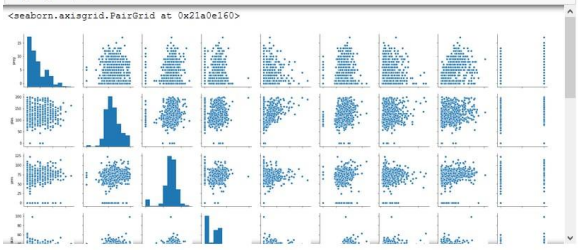

**Figure 3: Accuracy Results of Various Machine Learning Techniques**

```
df.describe()
```

| | preg | plas | pres | skin | test | mass | pedi | age | class |
|---|---|---|---|---|---|---|---|---|---|
| count | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 |
| mean | 3.845052 | 120.894531 | 69.105469 | 20.536458 | 79.799479 | 31.992578 | 0.471876 | 33.240885 | 0.348958 |
| std | 3.369578 | 31.972618 | 19.355807 | 15.952218 | 115.244002 | 7.884160 | 0.331329 | 11.760232 | 0.476951 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.078000 | 21.000000 | 0.000000 |
| 25% | 1.000000 | 99.000000 | 62.000000 | 0.000000 | 0.000000 | 27.300000 | 0.243750 | 24.000000 | 0.000000 |
| 50% | 3.000000 | 117.000000 | 72.000000 | 23.000000 | 30.500000 | 32.000000 | 0.372500 | 29.000000 | 0.000000 |
| 75% | 6.000000 | 140.250000 | 80.000000 | 32.000000 | 127.250000 | 36.600000 | 0.626250 | 41.000000 | 1.000000 |
| max | 17.000000 | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 | 2.420000 | 81.000000 | 1.000000 |

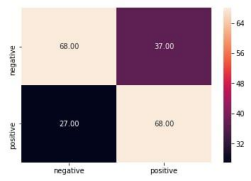
```
sb.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x21a0e160>
```



```
sb.heatmap(cm1,annot=True,fmt='.2f',xticklabels=['negative','positive'],yticklabels=['negative','pos
```

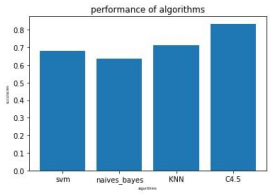
```
<matplotlib.axes._subplots.AxesSubplot at 0x25a954e0>
```




```
sb.heatmap(cm2,annot=True,fmt='.2f',xticklabels=['negative','positive'],yticklabels=['negative','pos
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x25b48a90>
```

```
plt.bar(index,acc)
plt.xlabel('algorithms', fontsize=5)
plt.ylabel('accuracies', fontsize=5)
plt.xticks(index, label, fontsize=10)
plt.title('performance of algorithms')
plt.show()
```




## 5. CONCLUSION

In this study, we analyzed the prognosis of early diabetes using machine learning techniques and calculated some of the risk factors associated with the disease. Learning from accurate health data can help diabetics. To prevent sugar, we use four popular machine learning algorithms in our experiments, namely Auxiliary Vector Machine (SVM), Naive Bayes (NB), near K (KNN), and C4.5, so that we can add sugar estimated based on the number of adults.

## 6. REFERENCES

Platt, John C. "12 fast training of support vector machines using sequential minimal optimization." Advances in kernel methods (1999): 185-208.

John, George H., and Pat Langley. "Estimating continuous distributions in Bayesian classifiers."Proceedings of the Eleventh Conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1995.

Aha, David W., Dennis Kibler, and Marc K. Albert. "Instance-based learning algorithms." Machine learning 6.1 (1991): 37-66.

Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Witten, I. H. et al. (1999). Weka: Practical machine learning tools and techniques with Java implementations.

Morteza, M., Franklyn, P., Bharat, S., Linying, D., Karim, K., and Aziz G. 2015. Evaluating the Performance of the Framingham Diabetes Risk Scoring Model in Canadian Electronic Medical Records.Canadian Journal of diabetes 39, 30(April. 2015), 152-156.