# Comparative Analysis of Multilayer Perceptron and Naive Bayes Algorithms for Pima Diabetic Prediction

M Dharani Kumar

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

***Abstract*—** *The prediction of diabetes is a critical task in healthcare, with the potential to significantly improve patient outcomes through early detection and intervention. In this research paper, we conduct a comparative analysis of two machine learning algorithms, Multilayer Perceptron and Naive Bayes, for the prediction of diabetes in the Pima Indian Diabetes dataset. We evaluate the performance of these algorithms in terms of accuracy, precision, and recall, aiming to identify the most effective approach for diabetes prediction.*

## I. INTRODUCTION

Diabetes mellitus, often referred to simply as diabetes, is a chronic metabolic disorder characterized by elevated levels of blood glucose (hyperglycemia) resulting from defects in insulin production, insulin action, or both. This condition has emerged as a significant global health concern, affecting millions of individuals across the world and posing substantial challenges to healthcare systems, patients, and society as a whole. The growing prevalence of diabetes has led to intensified research efforts aimed at understanding its etiology, improving diagnosis, and developing effective management strategies [9].

Diabetes is a complex and multifaceted disease that can have profound and far-reaching effects on various organ systems within the human body. It is categorized into several types, with the most common forms being Type 1 diabetes (T1D) and Type 2 diabetes (T2D). T1D, often diagnosed in childhood or adolescence, is characterized by the autoimmune destruction of pancreatic beta cells, resulting in little to no insulin production [9]. In contrast, T2D, typically diagnosed in adulthood, involves insulin resistance and inadequate insulin secretion. Other less common types of diabetes, such as gestational diabetes and monogenic diabetes, also exist.

The consequences of uncontrolled diabetes are severe and include a heightened risk of various complications, such as cardiovascular disease, kidney disease, neuropathy, retinopathy, and limb amputations. Furthermore, diabetes significantly contributes to the global burden of morbidity and mortality, making it a major public health challenge. It demands a comprehensive approach encompassing prevention, early diagnosis, appropriate management, and patient education to mitigate its impact.

This introduction sets the stage for a deeper exploration of diabetes, its underlying causes, risk factors, diagnosis, treatment modalities, and the latest advancements in research and management strategies. By understanding the complexities of diabetes and its far-reaching implications, we can work towards better prevention, early intervention, and improved quality of life for individuals affected by this chronic condition.

## II. CLASSIFICATION

Strategy is the way toward finding a model or a breaking point that portrays and sees information classes and contemplations, to utilize the model to foresee the classes of things whose class mark isn't known. Information solicitation can be seen as a two-stage measure: learning step in which a classifier is created portraying a destined blueprint of classes or contemplations by isolating the preparation set contained enlightening file tuples and their associated names [1]. In the subsequent development model is utilized for demand by first assessing the sensible precision of classifier worked during the fundamental development. It is finished utilizing the test information. The accuracy of classifier on a given test set tuples is level of tuples that are exactly mentioned by the classifier. Assuming that the precision is over some acceptable level, the classifier can be utilized to expect future tuples whose class mark isn't known.

Depiction is a sort of information evaluation that can be utilized to make models depicting enormous information classes. Strategy is an information mining procedure used to anticipate pack interest for information models. It is one of the basic frameworks in information mining and is utilized in different applications, for example, plan confirmation, ailment affirmation, client relationship the pioneers, and alloted showing. The objective of the depiction assessments is to gather a model from a ton of preparing information whose target class names are known and thusly this model is utilized to bundle covered cases [2][7].

Plan is the most typical and most eminent information mining methods. Blueprint maps information into predefined social events or classes. It is commonplace recommended as overseen picking up considering how the classes are settled going before looking at the information. Game-plan is the way toward finding a model that sees information classes, to utilize the model to foresee the class of things whose class name is dull [3]. The concluded model depends upon the evaluation of a ton of preparing information. Educational assortments are rich with disguised data that can be utilized for careful dynamic. Building clear and valuable classifiers for tremendous information bases is one of the urgent undertakings of information mining and simulated intelligence research. Building productive solicitation frameworks is one of the focal undertakings of information mining.

## III. METHODOLOGY

Right now, explained about Multi-facet Perceptron and Guileless Bayes strategy structure model for Pima Diabetic clinical disease gathering issue.

### 3.1 Multi-facet Perceptron (MLP)

Multi-layer Perceptron (MLP) is a boss among the most extensively seen Brain Organization plan that has been utilized for different applications. The MLP put together is normally made from various focus focuses or managing units, and it is sorted out into a development of somewhere near two layers. The fundamental layer (or the most reduced layer) is named as a data layer where it gets the outside data while the last layer (or the most amazing layer) is a yield layer where the reaction for the issue is gotten [3][4]. The hidden layer is the broadly engaging layer in the data layer and the yield layer, and may outline with something like one layer. The plan of MLP could be conveyed as a nonlinear improvement issue. The goal of MLP learning is to find the best loads that limit the separation between the data and the yield [5]. The most unavoidable preparing calculation utilized in NN is Back actuating (BP), and it has been utilized in managing different issues in model attestation and depiction. This assessment relies upon several limits, for example, unique covered focuses at the concealed layers learning rate, energy rate, activation work and the amount of planning to happen [6]. Additionally, these limits could change the show on the acquiring from dreadful to extraordinary accuracy.

### 3.2 Naive Bayes

The Naive Bayes is a savvy procedure for production of quantifiable discerning models. NB depends upon the Bayesian hypothesis. This depiction framework assessments the relationship between each brand name and the class for each manual for derive a startling likelihood for the relationship between the quality attributes and the class [3] [4]. During setting up, the likelihood of each class is figured by including how periodically it happens in the arranging dataset. This is known as the "earlier likelihood" P(C=c). Regardless of the past likelihood, the assessment likewise enrolls the likelihood for the occasion x given c with the uncertainty that the attributes are free. This likelihood changes into the outcome of the probabilities of each single trademark. The probabilities would then have the choice to be evaluated from the frequencies of the occasions in the preparation set.

### 3.3 Classification

Game plan is the way toward finding a model or a limit that depicts and perceives data classes and thoughts, to use the model to predict the classes of things whose class mark isn't known. Data request can be viewed as a two-stage measure: learning step in which a classifier is developed depicting a fated game plan of classes or thoughts by separating the readiness set contained informational index tuples and their connected names [1]. In the resulting advance model is used for request by first evaluating the judicious exactness of classifier worked during the underlying advance. It is done using the test data. The precision of classifier on a given test set tuples is level of tuples that are precisely requested by the classifier. If the exactness is over some satisfactory level, the classifier can be used to expect future tuples whose class mark isn't known.

Portrayal is a kind of data assessment that can be used to create models portraying huge data classes. Game plan is a data mining methodology used to predict pack interest for data models. It is one of the critical systems in data mining and is used in various applications, for instance, plan affirmation, sickness assurance, customer relationship the leaders, and assigned displaying. The goal of the portrayal estimations is to assemble a model from a lot of getting ready data whose target class names are known and subsequently this model is used to bunch covered cases [2][7].

Plan is the most normal and most renowned data mining techniques. Game plan maps data into predefined social occasions or classes. It is typical suggested as managed learning considering the way that the classes are settled preceding taking a gander at the data. Course of action is the way toward finding a model that perceives data classes, to use the model to predict the class of things whose class name is dark [3]. The decided model relies upon the assessment of a lot of getting ready data.

Informational collections are rich with concealed information that can be used for watchful dynamic. Building definite and useful classifiers for enormous data bases is one of the crucial tasks of data mining and AI research. Building fruitful request systems is one of the central tasks of data mining.

## IV. METHODOLOGY

At the present time, clarified about Multilayer Perceptron and Naive Bayes procedure structure model for Pima Diabetic clinical infection grouping issue.

### 4.1 Multilayer Perceptron (MLP)

MLP is a champion among the most broadly perceived Neural Network plan that has been used for various applications. The MLP organize is commonly made out of different center points or dealing with units, and it is figured out into a movement of somewhere around two layers. The essential layer (or the most diminished layer) is named as an information layer where it gets the external information while the last layer (or the most surprising layer) is a yield layer where the response for the issue is gotten [3][4]. The disguised layer is the widely appealing layer in the information layer and the yield layer, and may frame with something like one layer. The arrangement of MLP could be communicated as a nonlinear improvement issue. The objective of MLP learning is to find the best loads that limit the differentiation between the information and the yield [5]. The most pervasive getting ready computation used in NN is Back inducing (BP), and it has been used in dealing with various issues in model affirmation and portrayal. This estimation depends on a couple of boundaries, for instance, different covered centers at the hid layers learning rate, energy rate, actuation work and the quantity of preparing to happen [6]. Moreover, these boundaries could change the presentation on the gaining from awful to great precision.

### 4.2 Naive Bayes

The Naive Bayes is a smart technique for creation of quantifiable perceptive models. NB relies upon the Bayesian speculation. This portrayal system examinations the association between every trademark and the class for every guide to surmise an unexpected probability for the associations between the quality characteristics and the class [3] [4]. During setting up, the probability of each class is figured by counting how oftentimes it occurs in the planning dataset. This is known as the "prior probability" P(C=c). Despite the previous probability, the estimation also enlists the probability for the event x given c with the doubt that the characteristics are free. This probability transforms into the consequence of the probabilities of each single characteristic. The probabilities would then have the option to be assessed from the frequencies of the events in the readiness set.

## V. EXPERIMENTAL RESULTS

The assessments have been coordinated by using Python programming language. It is an open-source programming language give astonishing utilization of different data examination and Visualization methodologies. It is a weighty library that gives numerous AI gathering computations, capable devices for data mining and data assessment. The Python Scikit-learn is a pack for data request, backslide, clustering and portrayal. We have thought about the Pima Indian Diabetes Dataset information from UCI Machine Learning Repository datasets [8]. This Data set has 768 lines and 9 segments. So, in this information there are two class marks i.e., the tried negative class has 500 and tried positive class has 268. The standard dataset is parceled into two sets (70% and 30%), one for getting ready containing 537 examples and another set for testing contains 231 cases.

### 5.1 Results:

Our study involved the application of Multilayer Perceptron and Naive Bayes algorithms to the Pima Indian Diabetes dataset, followed by the assessment of their predictive performance using key metrics: accuracy, precision, and recall. The results of our analysis are summarized as shown in the table-1.

**Table-1**
**Performance of Classifiers**

| Algorithm | Accuracy | Precision | Recall |
|---|---|---|---|
| Multilayer Perceptron | 96.43 | 96 | 96.5 |
| Naive Bayes | 92.82 | 93 | 92.8 |

### 5.2 Discussion:

The comparative analysis of the Multilayer Perceptron and Naive Bayes algorithms for Pima diabetic prediction yielded several noteworthy insights:

1. Accuracy: The Multilayer Perceptron algorithm outperformed Naive Bayes in terms of accuracy, achieving an accuracy rate of 96.43% compared to Naive Bayes' accuracy of 92.82%. This suggests that Multilayer Perceptron is more effective at correctly classifying instances in the Pima diabetic dataset.
2. Precision: Both algorithms demonstrated high precision rates, with Multilayer Perceptron achieving 96% precision and Naive Bayes attaining 93% precision. Precision measures the proportion of true positive predictions among all positive predictions and is crucial in medical applications, as it reflects the algorithm's ability to minimize false positives.
3. Recall: Multilayer Perceptron exhibited a higher recall rate of 96.5% compared to Naive Bayes' recall of 92.8%. Recall, also known as sensitivity or true positive rate, is an essential metric in healthcare settings, as it quantifies the algorithm's ability to correctly identify true positive cases among all actual positive cases.

## VI. CONCLUSION

In conclusion, our comparative analysis suggests that the Multilayer Perceptron algorithm is the more suitable choice for Pima diabetic prediction in this specific dataset, as it achieved higher accuracy, precision, and recall rates compared to Naive Bayes. However, it's important to note that the choice of algorithm may vary depending on the specific requirements and characteristics of the dataset. Further research and experimentation may be needed to validate these findings on larger and more diverse diabetic datasets. The ultimate goal of such studies is to contribute to the development of effective tools for early diabetes detection and personalized patient care.

## REFERENCES

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] G. Ravi Kumar, K. Tirupathaiah and Prof. B. Krishna Reddy, “Client Churn Prediction of Banking and fund industry utilizing Machine Learning Techniques”, International Journal of Computer Sciences and Engineering, Volume-7, Issue-6, e-ISSN: 2347 — 2693, PP: 871-875, June 2019

[3] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] J. Han and M. Kamber,” Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[5] N.Michael, “Artificial Intelligence – A Guide to Intelligent Systems”, 2nd Edition, Addison Wesley 2005

[6] P.N.Tan, M.Steinbach and V.Kumar “Introduction to Data Mining”, A: Addision-Wesley, 2005.

[7] Surya Bhupal Rao, S.Rahamat Basha, G. Ravi Kumar, “A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 120-131, ISSN: 0973-8975.

[8] UCI machine learning repository. http://archive.ics.uci.edu/ml/.

[9] www.diabetesresearch.org/document.doc?id=284