2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)

DIABETES PREDICTION USING GAUSSIAN NAÏVE BAYES AND ARTIFICIAL NEURAL NETWORK

T Bharath Chandra
Department of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
thiruveedulabharathchandra@gmail.com

A Sujan Reddy
Department of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
sujanreddy298@gmail.com

A Adarsh
Department of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
adarshpersonalmail@gmail.com

M A Jabbar
Department of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
jabbar.meerja@gmail.com

B N Jyothi
Department of CSE(AI&ML)
Vardhaman College of Engineering
Hyderabad, India
jyothi.jo515@gmail.com

2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT) | 979-8-3503-8295-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDCOT61034.2024.10516226

***Abstract*—Early detection of diabetes can help prevent possible complications such as heart problems, kidney problems, nerve damage, and vision loss. Early detection not only improves health status but also improves overall quality of life by allowing timely control of blood sugar levels and important lifestyle changes. The main goal is to develop improved cohort-based diabetes prediction models. The main goal is to allow early diagnosis and early initiation of treatment. The strategy incorporates key steps such as comprehensive data preprocessing, exploratory data examination, and application of dimensionality reducing methods such as vital component investigation (PCA). This involves building predictive models using both Artificial Neural Networks (ANN) and Gaussian Naive Bayes (GNB). Additionally, a cluster model has been developed that allows comparative analysis between the different techniques used. The proposed model has a promising accuracy of approximately 94% for predicting diabetes onset, which has attracted attention and highlights its possible importance in real-world errors. Emphasizing detailed data analysis, data preprocessing optimization, rigorous model training, and comprehensive performance analysis. These learning efforts, combined with sophisticated data and advanced analytical concepts, deliver the results you expect from well-trained individual and cluster models.**

***Keywords*—PCA, Ensemble Learning, Gaussian Naïve Bayes, Voting Classifier**

I. Introduction

Diabetes is a serious health problem that requires timely intervention and self-monitoring to prevent complications. The International Diabetes Federation reports that the current global population of individuals with diabetes is 387 million, and this figure is expected to double by the year 2035[1].Apply advanced machine learning techniques to comprehensively analyse diabetes datasets and develop robust predictive models that can provide accurate disease predictions. Beyond the primary goal of early detection, by enabling proactive lifestyle changes and timely medical interventions to significantly improve patient outcomes, application of the model will improve resource management and improve healthcare outcomes. It is expected to advance diabetes management. The aim is not only to contribute to diabetes research, but also to harness the potential of advanced machine learning algorithms to gain new insights into the complexity of this disease. Our approach leverages ensemble learning, leverages Gaussian Naive Base (GNB), and combines artificial neural network (ANN) techniques with principal component analysis (PCA) to develop accurate predictive models. Main areas of focus include the definition of dynamic relationships, where GNB processes probabilistic relationships and reveals micropatterns in ANN data. Additionally, real-time data processing and easy integration into existing diabetes intervention programs highlight our commitment to effective implementation of the model. This paper is organized as follows : In Section 2, we review related work in the field. Section 3 outlines the methodology employed in our study. Results and discussion are presented in Section 4. Finally, we summarized important aspects and suggested areas for future research in Section 5.

II. Related Work

Diabetes Prediction Using Ensemble Techniques includes a comprehensive literature review to examine the application of ensemble techniques and ML algorithms in the environment of diabetes prediction. The survey focused on relevant studies that employed machine learning algorithms and techniques, such as feature importance's, different approaches, and feature selection, for accurate diabetes prediction. The literature survey provided valuable insights that guided the selection of algorithms and methodologies for the diabetes prediction.
Dutta et al.[2] proposed a study focused on analysing the

importance of features in diabetes prediction using ML techniques. The authors considered various algorithms and evaluated their ability to identify the important aspects for accurate predictions. The results of this study can be incorporated into the feature selection process in diabetes prediction to ensure the inclusion of relevant predictors.Sarwar et al.[3] proposed a study aimed at predicting diabetes using ML algorithms in the medical field. The authors evaluated the performance of various ML algorithms, including ensemble models, and compared their predictive accuracy. This finding can be incorporated into the selection of suitable ML algorithms for diabetes detection. Sonar et al.[4] examined the use of different ML approaches for diabetes prediction. The authors compared the performance of different algorithms in accurately predicting diabetes. The results provide insight into the effectiveness of different diabetes prediction algorithms and serve as a guide for selecting an appropriate model.Sivaranjani et al[5] proposed a study focused on diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction techniques. The authors investigated the impact of feature selection and dimensionality reduction on the accuracy of diabetes prediction models. The results can guide the feature engineering process in diabetes prediction to improve model performance.

Hassan et al[6] conducted a study about the usage of different models namely,Logistic Regression model with 88%, XGboost with 86.36% and Random Forest with 86.36% accuracy in diabetes detection. Tirpati et al[7] proposed a study focusing on four machine learning algorithms for diabetes prediction: linear discriminant analysis (LDA), KNN, SVM, and random forest (RF) algorithm. The experimental results of this study show that RF algorithm provides the maximum accuracy of 87%. 66% and outperforms other classification algorithms.

The findings from these studies highlight the significance of feature importance's, different machine learning approaches, and feature selection techniques in diabetes prediction. By incorporating these insights, the diabetes prediction can aim to develop an accurate and reliable model by leveraging suitable machine learning algorithms, identifying important features, and optimizing the feature selection process

III. Proposed Method

The methodology for diabetes prediction through Ensemble Learning, integrating Artificial Neural Networks (ANN), Gaussian Naive Bayes (GNB), and a Voting Classifier, encompasses several essential stages. Initially, the diabetes dataset undergoes preprocessing, involving data cleaning, normalization, and feature engineering to refine it for subsequent modeling. Next, the dataset is divided into subgroups for testing and training. Since the dataset might possess inherent biases or imbalances, techniques such as data resampling (e.g., oversampling) could be applied to the training set to rectify any skewed distributions and ensure balanced representation among classes.

An artificial neural network (ANN) model is then trained on the pre-processed data, using sophisticated algorithms to identify and force patterns of features. At the same time, a Gaussian Navie Bayes (GNB) is built and model is trained, with a probabilistic assumption of partitioning data based on conditional likelihood Based on the above, an ensemble model, especially the vote distribution, is subsequently built by forcing the trained ANN and GNB models by integrating the prophecy. This cluster model cooperatively combines the predictions of the two models to arrive at a final classification decision.

Performance evaluation of individual and ensemble models is done using key metrics such as accuracy, precision, recall, and F1 score. Applying fine-tuning and optimization techniques to improve the performance of individual models. This optimization process involves exploring various parameters and methodologies to improve their predictive accuracy and robustness. The resulting ensemble model combining ANN and GNB models exhibits improved predictive capability by exploiting the strengths of individual algorithms this provides more robust and accurate prediction of diabetes and reduces the probability of error or misclassification. This approach is scalable and can be continuously monitored and updated to adapt to models being developed in the diabetes database, ensuring long-term effectiveness Finally, the ensemble learning approach combines the strengths of ANN and GNB models to provide a comprehensive and accurate predictive approach for diagnosing diabetes.

The methodology adopted is designed to address the critical issue of diabetes prediction through a systematic and approach. In the pursuit of developing a robust and accurate ML -based model for diabetes detection, our methodology can be outlined as follows:

A. Dataset

The dataset used is PIMA Indians diabetes dataset[15]. The source of this is Kaggle. It has a total of 768 instances, each with 8 features. The features capture different aspects such as pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, age, and a diabetes pedigree function. The outcome variable is binary(1 for diabetic, 0 for non-diabetic). TABLE-1 discuss about the breakdown of the counts associated with features and attributes of PIMA dataset.

TABLE I
PIMA DIABETES DATASET

FEATURES	8
INSTANCES	768

B. Data Preprocessing and Feature Engineering

We began by loading the PIMA diabetes dataset[15], conducting exploratory data analysis through histograms and correlation heatmaps. Missing data was imputed using the mean strategy, and the dataset was standardized with Standard Scaler . Dimensionality reduction using PCA was performed for extracting top 5 principal components.

C. Model Building

The dataset was split into training (80%) and testing (20%) sets. An Artificial Neural Network (ANN) model was constructed with two hidden layers, each comprising 10 neurons, and trained on the training data. A Gaussian Naive Bayes (GNB) classifier was also instantiated.
Fig.1 shows the process diagram of diabetes mellitus prediction.

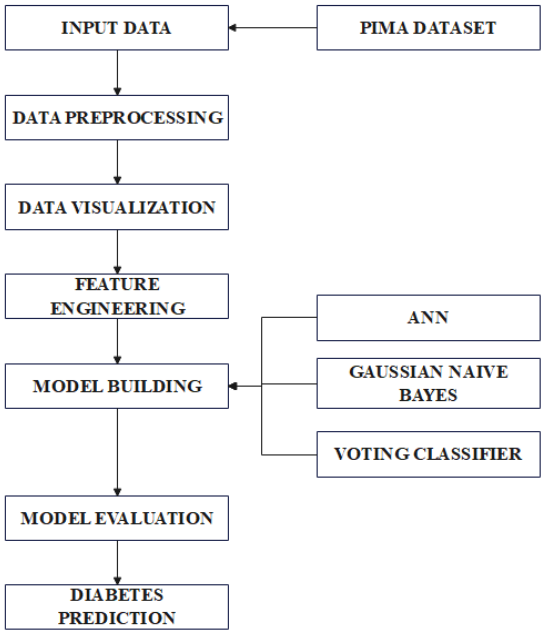


Fig. 1. Process diagram of diabetes mellitus prediction

In our approach,we are using majority voting classifier to combine ANN and Gaussian Naïve Bayes algorithms.

D. Ensemble Model

Ensemble learning is a new way in AI and data mining where we put together a bunch of weak algorithms to make a strong one. [16]. Majority voting classifier is an ensemble technique where multiple models predict an input, and the final decision is based on the majority vote. In our approach, an ensemble model was created using a Majority Voting Classifier, combining the GNB and ANN models. This ensemble model was trained on the training data to leverage the strengths of both models.

E. Model Evaluation

The accuracy of the ensemble model was assessed by making predictions on the test dataset. The ANN model achieved an accuracy of 98%, indicating its proficiency in predicting diabetes outcomes. Similarly, the Gaussian Naive Bayes (GNB) model exhibited a respectable accuracy of 90%, showcasing its competence in the task. However, the ensemble model, which combined the strengths of both ANN and GNB, achieved an impressive accuracy of 94%, highlighting its ability to provide a balanced and robust prediction of diabetes outcomes.

F. Deployment

It was deployed using Flutter, a mobile application development framework. We have developed a mobile interface to accept user input features, sending these features to the deployed model for prediction, and displaying the predicted outcome to users.

IV. Results and Discussion

When evaluating the performance of diabetes mellitus detection, several metrics are commonly used to assess its effectiveness. Here are some important performance metrics:

A. Performance Metrics

1) Accuracy: Accuracy is a measure of how many of the predictions made by the model are correct out of the total number of predictions. Discovering an overall impression of the validity of model's predictions.

2) Precision: Precision is used for measuring the model's capacity to make correct positive predictions(TP) out of all positive predictions (TP + FP)[10]. In the context of diabetes prediction, accuracy is the rate of properly prognosticated diabetes cases to all predicted diabetes cases. It is essential when you want to minimize false positive predictions.

3) Recall: Recall , assesses the model's capacity to accurately distinguish positive cases (TP) out of actual positive cases (TP+ FN)[11].In the context of diabetes prediction, recall would be the rate of properly prognosticated diabetic cases to all actual diabetic cases. It is important when you want to minimize false negative predictions.

4) F1 Score:

$$F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}(1)[8]$$

The results of the diabetes prediction model's performance metrics are presented below:

Artificial Neural Network (ANN): The ANN model demonstrated a commendable accuracy of 98.7%, indicating its proficiency in correctly classifying diabetes outcomes. Additionally, it achieved an impressive accuracy score of 100%, highlighting its ability to accurately identify actual positive cases. His F1 score of 98. 2% for the model reflects a balanced performance in terms of accuracy and repeatability. The ANN model actually identified a significant portion of the actual positive cases with a recall rate of 96. 4

Gaussian Naïve Bayes (GNB): The accuracy of the GNB model was 89. 6%, highlighting that it can accurately predict diabetes outcomes. The accuracy value of 85. 94% highlights its ability to accurately classify TP cases. The F1 score is 85. 95% . The recall rate of 85. 9% indicates that the GNB model actually identified a significant portion of the actual positive cases.

Ensemble Model: This model combines the strengths of both the ANN and GNB models, achieved an accuracy of 94.1%,

showcasing its proficiency in predicting diabetes outcomes. Remarkably, it resulted a precision score of 100%, indicating its ability to precisely identify true positive cases. The ensemble model's F1 score of 91.4% reflects its balanced performance in terms of precision and recall. However, the recall score of 84.2% signifies that it was slightly less effective in identifying actual positive cases compared to the ANN model.

TABLE-2 discuss the performance metrics of proposed models(ANN,GNB and Ensemble model).

TABLE II
PERFORMANCE METRICS

Model	**Accuracy**	**Precision**	**F1-score**	**Recall**
Artificial Neural Networks	98.7	100	98.2	0.964
Gaussian Naïve Bayes	89.6	85.9	85.9	0.859
Ensemble Model	94.1	100	91.4	0.842

TABLE-3 discuss about the accuracies obtained by various models.

TABLE III
COMPARISION WITH OTHER MODELS

S.No	**Models**	**Accuracy**
1	Random Forest[1]	84%
2	SVM[2]	79%
3	Generalized Boosted Regression[7]	90.91%
4	DL-SVM[9]	76.81%
5	KNN[12]	94.5%
6	Naïve Bayes[14]	57%
7	SGD[14]	69%
8	Proposed Model(ANN)	98.7%
9	Proposed Model(GNB)	89.6%
10	Proposed Model(Ensemble model)	94.1%

B. Confusion Matrix

A confusion matrix is a summary chart for assessing how accurately a machine learning model predicts outcomes by comparing its results to the actual data.[13]

Below are the confusion matrix figures of proposed models. Fig.2 shows the confusion matrix for ANN model.Fig.3 shows the confusion matrix for Gaussian Naïve Bayes model.Fig.4 shows the confusion matrix for Ensemble model.

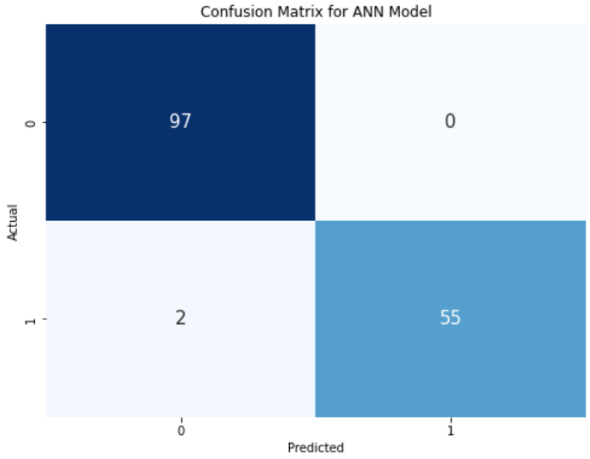


Fig. 2. Confusion Matrix for Artificial Neural Networks

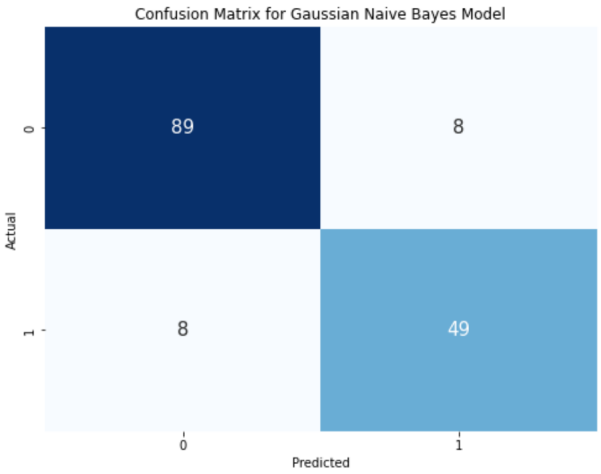


Fig. 3. Confusion matrix for Gaussian Naïve Bayes

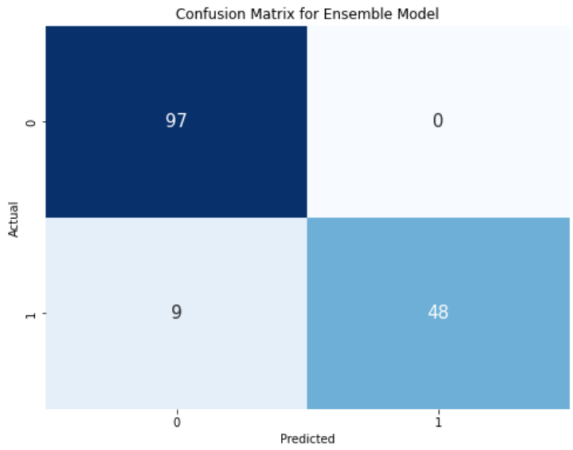
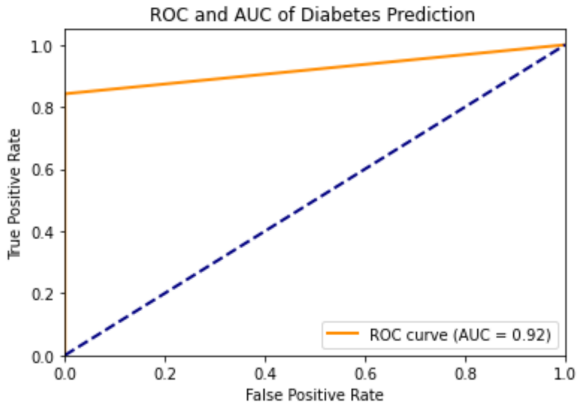


Fig. 4. Confusion matrix for ensemble model

C. ROC and AUC

In our diabetes prediction project utilizing ensemble learning, we assess the performance of our models through Receiver ROC curves and the AUC metric.Fig.5 shows ROC and AUC of diabetes prediction.




```
AUC: 0.9210526315789473
```

Fig. 5. ROC and AUC for diabetes prediction

V. Conclusion

Using an ensemble learning strategy, we developed a model that exhibits good ability to predict diabetic disease. Our efforts have included extensive dataset analysis, model development, and rigorous testing, and are now producing not only useful insights but also impressive results. The most notable feature of our diabetes prediction model was the voting classifier. By combining the strengths of GNB and ANN models, we achieve an impressive 94% accuracy. This is an important medical milestone that allows medical professionals to detect the onset of diabetes early. Early identification of people at risk allows for tailored interventions in the long term, improving outcomes for those affected and reducing strain on healthcare structures in the future. The methodology relies on PIMA Indian diabetes dataset which and its generalizability to other populations may be limited. It may not capture the diversity of diabetes related factors since it only focuses on a specific ethnic group. The proposed model an accuracy 94%, but it is essential to consider the practical implications and potential challenges when applying the model in real-world clinical settings.There is a possibility that Looking to the future, our goal is to continue to develop. We plan to consider additional ensemble techniques to improve the model's predictive ability as well. Additionally, expanding functionality to include more complete patient profiles could open new possibilities for diabetes prediction. Solving statistical challenges and working closely with healthcare organizations is an important part of our future plans. Through continuous research and improvement, our goal is not only to refine the performance and dependability of diabetes prognostic systems. Our goal is to increase ease of use and real-world impact to make a real and positive impact on the lives of people affected by diabetes. Our goal is to contribute to the continuous improvement of medical practice and create a world where proactive interventions lead to better health outcomes.

References

[1] Alghamdi, T., 2023. Prediction of Diabetes Complications Using Computational Intelligence Techniques. Applied Sciences, 13(5), p.3030.

[2] Dutta, D., Paul, D. and Ghosh, P., 2018, November. Analysing feature importances for diabetes prediction using machine learning. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 924-928).

[3] Sarwar, M.A., Kamal, N., Hamid, W. and Shah, M.A., 2018, September. Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) (pp. 1-6).

[4] Sonar, P. and JayaMalini, K., 2019, March. Diabetes prediction using different machine learning approaches. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 367-371).

[5] Sivaranjani, S., Ananya, S., Aravinth, J. and Karthika, R., 2021, March. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 141-146).

[6] Hassan, M.M., Peya, Z.J., Mollick, S., Billah, M.A.M., Shakil, M.M.H. and Dulla, A.U., 2021, July. Diabetes prediction in healthcare at early stage using machine learning approach. In 2021 12th International conference on computing communication and networking technologies (ICCCNT) (pp. 01- 05).

[7] Tripathi, G. and Kumar, R., 2020, June. Early prediction of diabetes mellitus using machine learning. In 2020 8th international conference on reliability, Infocom technologies and optimization (trends and future directions)(ICRITO) (pp. 1009-1014).

[8] Srivastava, Saurabh & Singh, Girdhari. (2016). F1 Score Analysis of Search Engines. S.K.I.T Research Journal. 6. 1-6.

[9] Yahyaoui, A., Jamil, A., Rasheed, J. and Yesiltepe, M., 2019, November. A decision support system for diabetes prediction using machine learning and deep learning techniques. In 2019 1st International informatics and software engineering conference (UBMYK) (pp. 1-4).

[10] Powers, D.M., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

[11] Olson DL, Delen D. Advanced data mining techniques. Springer Science & Business Media; 2008.

[12] Abdulhadi, N. and Al-Mousa, A., 2021, July. Diabetes detection using machine learning classification methods. In 2021 International Conference on Information Technology (ICIT) (pp. 350-354).

[13] Karimi, Zohreh. (2021). Confusion Matrix.

[14] Emon, M.U., Zannat, R., Khatun, T., Rahman, M. and Keya, M.S., 2021, January. Performance analysis of diabetic retinopathy prediction using machine learning models. In 2021 6th International Conference on Inventive Computation Technologies (ICICT) (pp. 1048-1052).

[15] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database Last accessed on 16-02-2024.

[16] Jabbar, M.A. and Aluvalu, R., 2017. RFAODE: A novel ensemble intrusion detection system. Procedia computer science, 115, pp.226-234.