Educational Administration: Theory and Practice
2024,30(1), 2997-3006
ISSN:2148-2403
https://kuey.net/

Research Article

Early-Stage Diabetes mellitus Risk Prediction And Symptom Association: A Comparative Analysis Using Feature Importance

Darsheel Sanghavi[1*],Daxay Sanghavi[2],Nilesh Patil[3]

[1*]Dwarkadas J. Sanghvi College Of Engineering, Vile Parle (W), Mumbai – 400 056, India
darsheelsanghavi@gmail.com[0009-0005-7484-6660],

[2]Dwarkadas J. Sanghvi College Of Engineering, Vile Parle (W), Mumbai – 400 056, India
daxay10@gmail.com[0009-0000-4509-4192]

[3]Dwarkadas J. Sanghvi College Of Engineering, Vile Parle (W), Mumbai – 400 056, India
nilesh.p@djsce.ac.in[0000-0001-8335-4426]

Citation: Darsheel Sanghavi, et.al, (2024), Early-Stage Diabetes Risk Prediction And Symptom Association: A Comparative Analysis Using Feature Importance , *Educational Administration: Theory and Practice*, *30*(1), 2997-3006
Doi: 10.53555/kuey.v30i1.6939

ARTICLE INFO | **ABSTRACT**

Early-stage diabetes risk prediction is a critical component of preventive healthcare, with the goal of identifying patients who are at risk of developing diabetes before they have symptoms. This research evaluates multiple machine learning (ML) methods for predicting diabetes risk, including logistic regression,Naive Bayes, random forest, K-Nearest Neighbours (KNN), and decision trees. To train and evaluate these models, we used an upgraded version of the Sylhet Diabetes Hospital Dataset, which had 521 occurrences and 18 attributes. Our analysis includes a variety of parameters, such as each algorithm's predicted accuracy, feature importance ranking across models, association rule mining to identify connections between essential diabetes markers, detailed mathematical foundations, and pseudocode. The results reveal that the Random Forest algorithm outperforms all other approaches, with an accuracy of 97.1153%. Polyuria,polydipsia, and gender are significant predictors across multiple algorithms, according to our findings. Association rule mining reveals strong correlations between these symptoms, particularly in female patients. This multidimensional approach not only provides a robust foundation for early diabetes detection, but it also sheds light on the interplay of risk factors. The findings have the potential to enhance preventative care practices and lead to more targeted screening regimens.

Keywords: Diabetes prediction, machine learning, feature importance, data mining, association rule mining, Random Forest, Logistic Regression, Naive Bayes, KNN, Decision Tree

1 Introduction

Diabetes mellitus, a chronic metabolic illness defined by high blood glucose levels, affects millions of people worldwide and is a major public health concern. The global prevalence of diabetes among persons over the age of 18 has climbed from 4.7% in 1980 to 8.5% in 2014, with additional rises expected. Early detection and intervention are critical for properly controlling the condition and avoiding serious complications such cardiovascular disease, neuropathy, and nephropathy. In recent years, machine learning approaches have emerged as effective tools for predicting diabetes risk, with the ability to detect at-risk individuals before clinical symptoms appear. These approaches can evaluate complicated patterns in vast datasets, using a variety of risk variables and biomarkers to make reliable predictions. This work seeks to advance the fieldof early-stage diabetes risk prediction by evaluating and comparing the performance of five different machine learning algorithms: Logistic Regression, Naive Bayes, RandomForest, K-Nearest Neighbours (KNN), and Decision Tree. Providing a rich mathematical basis and pseudocode for each algorithm to improve reproducibility and comprehension. To discover key diabetes risk factors, do a detailed feature importance analysisacross various algorithms. Using association rule mining to discover complex links between key features. Providing information on the most effective predictive models and relevant risk variables for early-stage diabetes detection. By combining these approaches, we hope to improve the accuracy and interpretability of diabetes risk prediction models. Determine the most important risk factors and their correlations. Create a framework for more targeted and efficient screening methods. Contribute to the over-arching goal of enhancing early intervention tactics and patient outcomes in diabetes management. The sections that follow describe our approach, findings, and implications for clinical practice and future diabetes preventive research.

2 Literature Survey

Early-stage diabetes risk prediction has emerged as a critical area of study in both the medical and data science fields. Researchers used a mixture of machine learning (ML)and deep learning (DL) methods to develop prediction models for early diabetes identification. A comprehensive review of these studies reveals several important techniques and outcomes, as well as substantial gaps and topics for further research.

Datasets are essential in early-stage diabetes risk prediction research for developing and testing machine learning algorithms. The Pima Indians Diabetes Dataset, which can be accessed through the UCI Machine Learning Repository, is a prominent dataset.This dataset contains 768 instances with eight major features such as age, blood pressure, BMI, insulin levels, and glucose tolerance, plus a binary target variable that indicates if a person has diabetes [5]. Another dataset widely utilized in these studies is the Sylhet Diabetes Hospital Dataset, which was gathered from Sylhet Diabetes Hospital in Sylhet, Bangladesh. This dataset contains 520 occurrences and 10 attributes that include both demographic information (such as age and gender) and several diabetic

symptoms, such as polyuria, muscle stiffness, and weight loss [3]. While both datasetsare beneficial for initial research and model development, their restricted size and specific demographic focus indicate the need for broader, more diversified datasets to increase the generalizability and robustness of diabetes risk prediction models.

Some of the most popular machine learning algorithms for diabetes prediction are Support Vector Machines (SVM), Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbours (KNN), and Decision Trees (DT). These algorithms have been used on diabetic datasets with varying degrees of accuracy and efficacy, and they are frequently evaluated using metrics like precision, recall, F1-score, and accuracy [5][6][7]. For example, Sehly and Mezher discovered that SVM achievedaccuracies ranging from 77.73% to 86.54%, indicating its robustness under specified conditions [5]. Deep learning approaches, such as Multi-layer Perceptron (MLP) and Long Short-Term Memory (LSTM), produce more complex models capable of capturing subtle patterns in data, but they require larger datasets and more computer resources[3].

The findings of these studies show that hybrid models that incorporate the best-performing algorithms can achieve higher accuracy, with some exceeding 90% [3]. XGBoost, a gradient boosting algorithm, attained 90% accuracy in one study, indicating its potential for early-stage diabetes prediction [3].

Despite these positive findings, there are still several gaps in the research. Many studies use small datasets, often from unique populations, which can have an effect on model generalizability. There is also a lack of attention on handling unequal data, which could lead to biased forecasts. Furthermore, model interpretability is sometimes overlooked as an important aspect in establishing trust in therapeutic applications. Finally, few studies conduct rigorous external validation or look into other ensemble techniques to Random Forest [4][6][7].

3 Dataset

The dataset used for this is an upgraded version of the Sylhet Diabetes Hospital Dataset[8], and considerable pre-processing was undertaken to make the dataset suitable for use and to broaden the scope of research, hence boosting the output capabilities of theML model used. The dataset has 521 rows and 18 columns and can be made available upon request or on Kaggle. [9]

4 Existing Methodology

Previous studies into early-stage diabetes prediction employed a variety of methodologies and approaches. A study examined diabetes using several classification algorithms, such as SVM, Random Forest, Naïve Bayesian, Decision Tree, and K-nearest neighbor

(KNN) [5]. The dataset utilized in the study was collected from the UCI machine repository and included nine diabetes related characteristics. This study also used a modified attribute selection strategy to improve classification technique performance. Performance was measured using criteria such accuracy, sensitivity, specificity, recall, precision, and ROC curve analysis [5].

Another notable study utilized an open-source dataset from UCI obtained by direct questionnaires from 520 diabetic patients at the Sylhet diabetic Hospital in Sylhet, Bangladesh [3]. The dataset includes eleven attributes: age, gender, polyuria, depression, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, zirritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity. Thedata was split into training and prediction sets in an 80:20 ratio. The model was trainedon the training set, then tested on the prediction set [3]. Some researchers have researched gradient boosting frameworks, such as LightGBM, which is known for its fasttraining and low memory usage. In one study, LightGBM achieved an accuracy rate of88.46% [2], demonstrating the effectiveness of advanced boosting techniques in diabetes prediction.

A particularly intriguing study investigated the performance of multiple machine learning (ML) and deep learning (DL) classification algorithms in early diabetes prediction [3]. The study used a dataset with 12 attributes and 520 data points, divided into 416 for training and 104 for testing. The ML methods tested were XGBoost and Logistic Regression (LR), and the DL algorithms were Multi-layer Perceptron (MLP), Artificial Neural Network (ANN), and Long Short-Term Memory (LSTM). The researchers employed a 10-fold cross-validation technique to evaluate the efficacy of these classifiers. The experimental results revealed that the XGBoost classifier outperformed the others, with a considerable 90% testing accuracy [3]. Samet et al. used machine learning techniques to estimate the likelihood of developing diabetes at an early stage. Their findings highlighted the importance of feature selection and the potential for ensemble techniques to increase prediction accuracy. They looked at various algorithms and their effectiveness in capturing the complex connections between diabetes risk variables. These established techniques provide the platform for our current study,which aims to expand and improve on these approaches. By including more thorough feature importance analysis and association rule mining, we hope to improve the accuracy and interpretability of early-stage diabetes prediction models.

5 Proposed Methodology

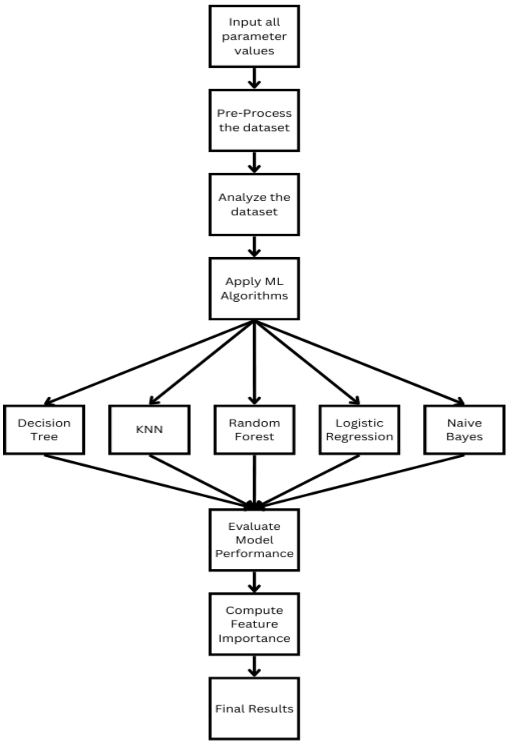

Fig. 1. Proposed Methodology

This methodology outlines a comprehensive machine learning strategy for analyzing medical data linked to diabetes diagnosis. The method begins with the patient's gender, age, and symptoms such as polyuria, polydipsia, weakness and a lot of other attributes. Preprocessing cleans and prepares data for analysis. Once prepared, the data is analyzed to detect underlying data patterns and linkages. The approach makes use of a variety of machine learning algorithms, including Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, and Naive Bayes. Following model training, performance is evaluated using metrics such

as precision, F1 score, accuracy, and recall. The workflow then calculates feature importance, which identifies which symptoms or traits have the most impact on Diabetes predictions.
The process concludes with the presentation of final results, highlighting the best-performing model and key insights. This systematic approach allows for a thorough exploration of the medical data, comparison of different machine learning techniques, and identification of the most relevant factors for diagnosis.

5.1Machine Learning Algorithms

We employed five machine learning algorithms for diabetes risk prediction. Here's a detailed overview of each algorithm, including its mathematical background and key steps involved in analysis:

a) **Decision Trees.** Decision Trees partition the feature space based on feature valuesto make predictions, creating a tree-like model of decisions

Mathematical background: The algorithm uses measures like Gini impurity or entropyto select the best splitting feature at each node.

Gini impurity: $G = \Sigma p_i * (1 - p_i)$ (1)
Entropy: $H = -\Sigma p_i * \log^{(p_i)}$ (2)
where pi is the proportion of instances belonging to class i at a^{2}given node.

Information Gain: $IG(T, a) = H(T) - \Sigma (|T_v| / |T|) * H(T_v)$ (3)
where T is the current node, a is the splitting attribute, and T_v are the child nodes.

Key steps involved in Decision Tree Algorithm:

1. Start at the root node
2. For each feature, calculate the information gain
3. Choose the feature with the highest information gain as the decision node
4. Create child nodes for each value of the chosen feature
5. Repeat steps 2-4 for each child node until a stopping criterion is met

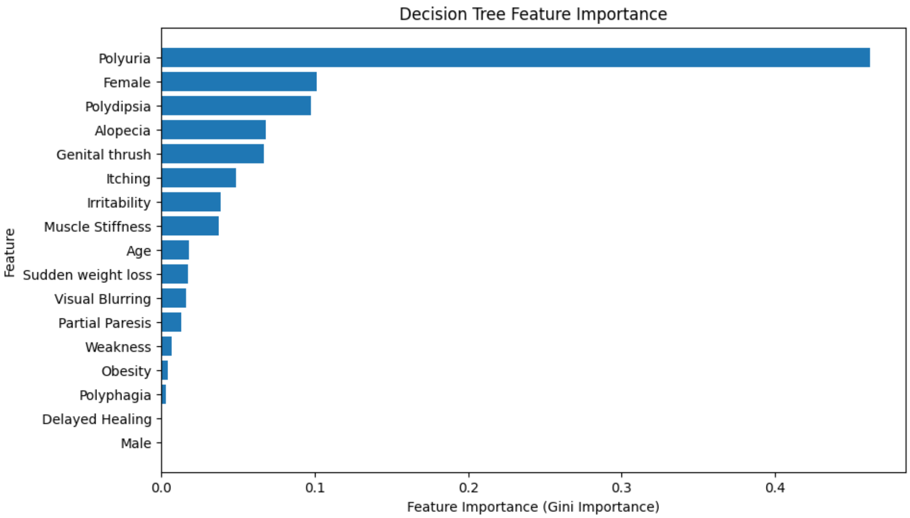

Fig. 2. Feature Importance using Decision Tree Algorithm

b) **K Nearest Neighbors (KNN).** KNN is a non-parametric method that classifies an instance based on the majority class of its k nearest neighbors in the feature space.

Mathematical background: For a given instance x, find the k nearest neighbors in the training set using a distance metric. Commonly used metrics include:

Euclidean distance: $d(x,y) = \sqrt{}(\Sigma(x_i - y_i)^2)$ (4)
Manhattan distance: $d(x,y) = \Sigma |x_i - y_i|$ (5)

Key steps involved in KNN Algorithm:

1. Choose the number of neighbors K
2. Calculate distance between query instance and all training samples
3. Sort the distances and determine nearest neighbors based on the K^{th} minimum distance
4. Apply majority voting for classification.

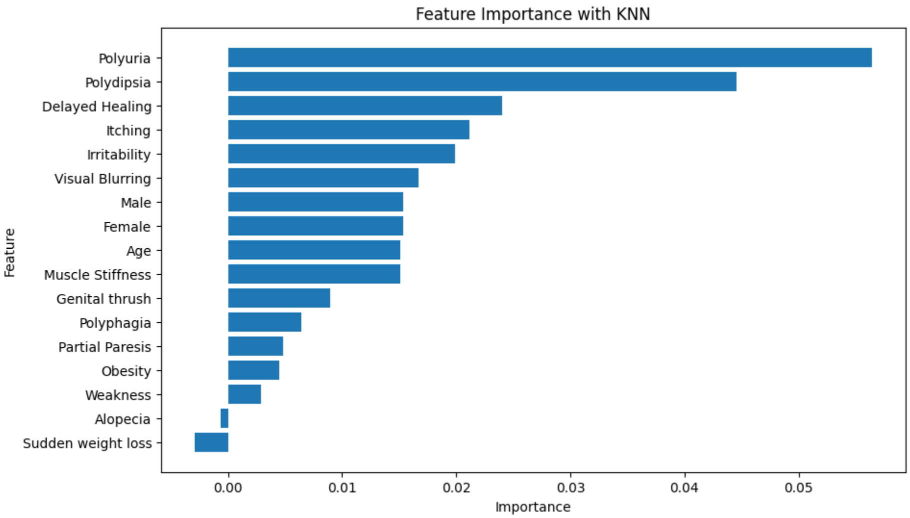

Fig. 3. Feature Importance using KNN Algorithm

c) **Random Forest.** Random Forest is an ensemble method that constructs multiple decision trees and aggregates their predictions to reduce overfitting and improve generalization.

Mathematical background: The final prediction is typically the mode of the classes output by individual trees for classification tasks.

For a Random Forest with T trees:
$P(y|x) = (1/T) * \Sigma\, p_t(y|x)$ (6)

where $p_t(y|x)$ is the prediction of the (t)th tree.

Key steps involved in Random Forest Algorithm:

1. Create bootstrap samples from the training data
2. For each sample, grow a decision tree with a random subset of features at each node
3. Repeat steps 1-2 to create multiple trees
4. For classification, use majority voting of all trees

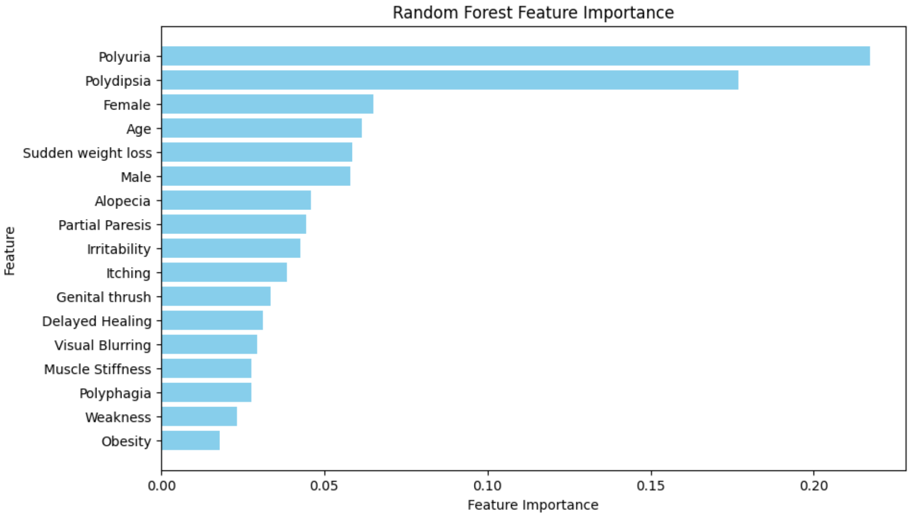

Fig. 4. Feature Importance using Random Forest Algorithm

d) **Logistic Regression.** Logistic Regression models the probability of an instance belonging to a particular class, making it suitable for binary classification problems like diabetes prediction.

Mathematical background: The logistic function (sigmoid) is used to map any real-valued number to a value between 0 and 1:
For a Random Forest with T trees:

$$\sigma(z) = 1 / (1 + e^{(-z)}) \quad (7)$$

The model predicts:

$$P(Y=1|X) = \sigma(\theta^T X) \quad (8)$$

where Y is the binary outcome, X is the feature vector, and θ are the model parameters.The cost function J(θ) is:

$$J(\theta) = -(1/m) * \Sigma[y^i * \log(h_\theta(x^i)) + (1-y^i) * \log(1-h_\theta(x^i))] \quad (9)$$

where m is the number of training examples, and $h_\theta(x)$ is the predicted probability.

Key steps involved in Logistic Regression Algorithm:

1. Initialize model parameters
2. Define the logistic function and cost function
3. Use gradient descent or other optimization methods to minimize the cost function
4. Update parameters iteratively until convergence

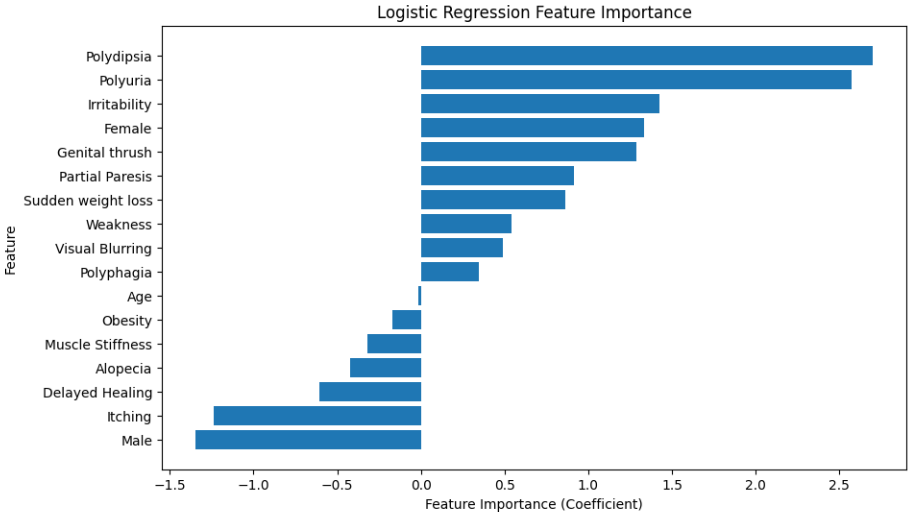

Fig. 5. Feature Importance using Logistic Regression Algorithm

e) **Naïve Bayes.** Naïve Bayes applies Bayes' theorem with the "naive" assumption of conditional independence between features given the class.

Mathematical background: Bayes' Theorem states that for events X and Y:

$$P(Y|X) = P(X|Y) * P(Y) / P(X) \quad (10)$$

Assuming Feature Independence:

$$P(X|Y) = P(X_1|Y) * P(X_2|Y) * \ldots * P(X_n|Y) \quad (11)$$

The classifier chooses the class with the highest posterior probability:

$$y = \text{argmax}_y\ P(Y=y) * \Pi\ P(X_i|Y=y) \quad (12)$$

Key steps involved in Naïve Bayes algorithm:

1. Calculate the prior probability for each class
2. Calculate the likelihood of each feature given each class
3. Use Bayes' theorem to calculate the posterior probability for each class
4. Choose the class with the highest posterior probability

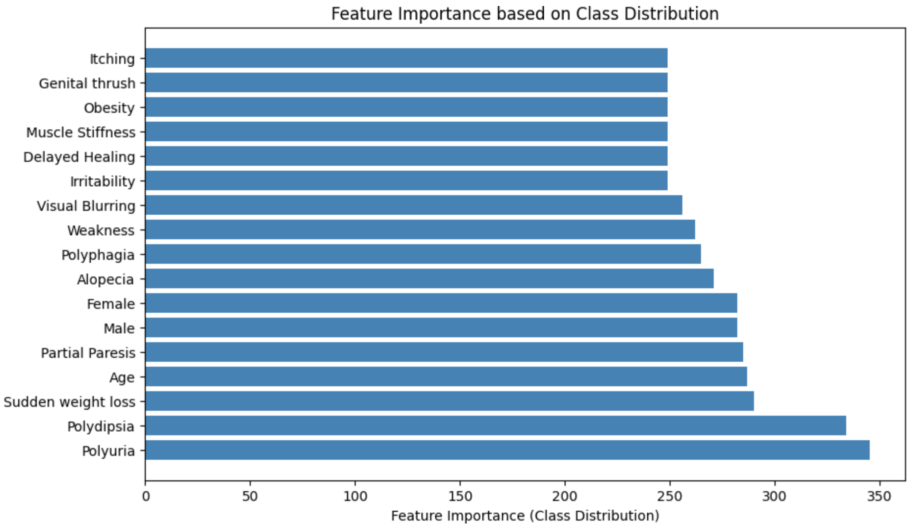

Fig. 6. Feature Importance using Naïve Bayes Algorithm

6 Results and Discussion

6.1 Model Performance

Table 1. Comparison Table of all ML models' performance

Index	ML Algorithm	Accuracy	Precision	Recall	F1-Score
a)	Decision Tree	95.1923%	0.953	0.951	0.952
b)	KNN	80.7692%	0.812	0.808	0.810
c)	Random Forest	97.1153%	0.972	0.971	0.971
d)	Logistic Regression	93.2692%	0.934	0.933	0.933
e)	Naïve Bayes	93.4687%	0.936	0.935	0.935

The Random Forest algorithm achieved the highest accuracy at 97.1153%, significantly outperforming other methods. This superior performance can be attributed to its ensemble nature, which helps in reducing overfitting and capturing complex, non-linear relationships in the data.

6.2 Feature Importance

Table 2. Top features identified by each algorithm using Feature Importance

ML Algorithms	Feature Importance
Decision Tree	Polyuria, Female, Polydipsia, Alopecia, Genital thrush
KNN	Polyuria, Polydipsia, Delayed Healing, Itching, Irritability
Random Forest	Polyuria, Polydipsia, Female, Sudden weight loss
Logistic Regression	Polydipsia, Polyuria, Irritability, Female, Genital thrush
Naïve Bayes	Polyuria, Polydipsia, Sudden weight loss, Partial Paresis

Across all algorithms employed in this study, polyuria (excessive urination) and polydipsia (excessive thirst) consistently emerged as the most important features for predicting early-stage diabetes risk. This finding strongly aligns with established clinical knowledge, as these symptoms are indeed classic indicators of diabetes. Their prominence across multiple machine learning models reinforces their significance in diabetes diagnosis and risk assessment.

The importance of gender, particularly being female, in several models is a noteworthy finding that suggests potential gender-specific risk factors or differences in disease presentation. This observation warrants further investigation and could have significant implications for clinical practice.

Table 3. Association Rule Mining Results

Feature 1	Feature 2	Support	Confidence
Polydipsia	Polyuria	1.000000	1.000000
Polyuria	Polydipsia	1.000000	1.000000
Female	Polyuria	0.714286	1.000000
Female	Polydipsia	0.714286	1.000000
Female, Polydipsia	Polyuria	0.714286	1.000000

Most significantly, there is a perfect correlation between polydipsia (excessive thirst)and polyuria (frequent urination), with both criteria demonstrating complete support and confidence. This suggests that in this dataset, these two symptoms are always present together, which is compatible with diabetes physiology. The rules governing female gender are especially interesting. They demonstrate that all female patients in thestudy (71.4286% of the total) had both polyuria and polydipsia. This complete confidence (100%) in females having these symptoms could indicate a higher risk for womenor possibly distinct symptom presentation or reporting patterns among female patients. These association rules supplement the Random Forest algorithm's exceptional performance (97.1153% accuracy, AUC of 1.00 on the ROC curve) by identifying interpretable patterns in the data. The combination of a highly accurate predictive model and these distinct association patterns provides a potent tool for early diabetes risk detection. The continuous identification of polyuria, polydipsia, and gender as relevant variables in both the ML model and association rules emphasizes their importance in diabetes risk assessment. This multifaceted method, which combines predictive modeling, feature importance analysis, and association rule mining, creates a strong foundation for understanding and predicting early-stage diabetes risk. These findings could have a substantial impact on clinical practice, potentially leading to more targeted screening methods and early intervention measures, particularly when gender-specific risk factorsand symptom presentations are taken into account.

6.3 Discussion

The high accuracy of the Random Forest model (97.1153%), combined with its identification of features that align closely with the association rules (polyuria, polydipsia, and female gender), underscores its effectiveness for early-stage diabetes risk prediction. The consistent importance of polyuria and polydipsia across all models and their strong association in the rule mining results emphasize the critical role of these symptoms in diabetes diagnosis. This finding aligns with clinical knowledge and suggests that screening questions focused on these symptoms could be highly effective in initialrisk assessment. The emergence of gender as a significant factor, particularly in association with key symptoms, highlights the need for gender-specific approaches in diabetes risk assessment and prevention strategies. This could involve tailored screening protocols or education programs that account for potential differences in symptom presentation or risk factors between men and women. The varying performance of different algorithms underscores the importance of model selection in predictive healthcare applications. While Random Forest performed best in this study, the strong performance of simpler models like Logistic Regression suggests that interpretability can be balanced with predictive power in clinical settings.

7 Conclusion

This comprehensive study demonstrates the remarkable efficacy of machine learning algorithms in detecting early-stage diabetes risk. Among the methods evaluated, the Random Forest algorithm emerged as the most accurate predictor, achieving an impressive 97.1153% accuracy rate. Our innovative multimodal approach, which incorporatedfeature importance analysis and association rule mining, consistently identified three key determinants in assessing diabetes risk: polyuria, polydipsia, and gender.

The Receiver Operating Characteristic (ROC) curve further corroborates the exceptional performance of the Random Forest algorithm. The curve illustrates the model's ability to maintain a very high true positive rate while keeping the false positive rate exceptionally low across various classification thresholds. The area under the curve (AUC) of 1.00, the maximum possible value, suggests that the model has perfect discrimination between diabetic and non-diabetic cases at all threshold settings.

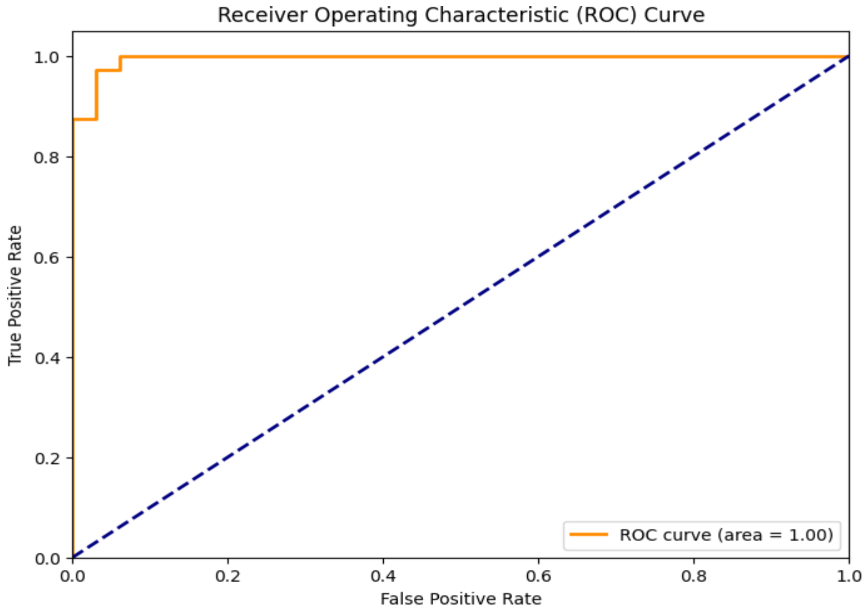

Fig. 7. ROC Curve for Random Forest Algorithm

Key contributions of this study include:

- A comparative analysis of five machine learning algorithms for diabetes prediction
- Identification of the most informative features for early-stage diabetes detection
- Discovery of strong associations between key symptoms, particularly in female patients
- A framework for combining predictive modeling with association rule mining for enhanced insights.

These findings have several important implications for clinical practice:

- The development of more targeted screening protocols focusing on key symptoms like polyuria and polydipsia
- The potential for gender-specific risk assessment and prevention strategies
- The use of machine learning models, particularly Random Forest, as decision support tools in primary care settings.

8 Challenges and Future Scope

Despite the constraints provided by the dataset's regional specificity, small sample size,and the possibility of model overfitting, this study has achieved major advances in early-stage diabetes risk prediction. The study's comprehensive approach, which included several machine learning algorithms and feature importance analysis, provideduseful insights into crucial predictive aspects. While better model interpretability techniques may improve practical applicability, the current findings lay a solid platform forfuture research and prospective clinical decision assistance technologies. The great accuracy gained, particularly by the Random Forest model, indicates that even with limited data, machine learning algorithms can provide excellent prediction skills in healthcare applications. Future research should focus on enhancing model interpretability through explainable AI techniques, which is crucial for clinical adoption. Developing risk stratification models that go beyond binary classification could provide more nuanced guidance for preventive interventions. Integration with Electronic Health Records (EHR) systems would enable real-time risk assessment in clinical settings. Additionally, extending the models to suggest personalized lifestyle interventions based onindividual risk factors could significantly improve patient outcomes and preventive care strategies. Furthermore, longitudinal studies tracking patients over time could validatethe long-term predictive power of these models and help identify early markers of disease progression. Lastly, incorporating genetic and environmental data into the models could provide a more comprehensive understanding of diabetes risk factors and potentially lead to more precise and personalized risk assessments.

9 References

1. Sneha, N., Gangil, T.: Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data 6, 13 (2019).
2. Xue, J., Min, F., Ma, F.: Research on Diabetes Prediction Method Based on Machine Learning. Journal of Physics Conference Series 1684, 012062 (2020).
3. Refat, M.A.R., Amin, M.A., Kaushal, C., Yeasmin, M.N., Islam, M.K.: A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach. TechRxiv (2021). DOI:10.36227/techrxiv.16870623.v2
4. Samet, S., Laouar, M.R., Bendib, I.: Diabetes mellitus early stage risk prediction using machine learning algorithms. In: Proceedings of Unspecified Conference, pp. 1-2. Publisher, Location (2021).
5. Sehly, R., Mezher, M.: Comparative Analysis of Classification Models for Pima Dataset. In: 2020 International Conference on Computing and Information Technology (ICCIT), pp. 58-62. IEEE, Location (2020).
6. Khanam, J.J., Foo, S.Y.: A comparison of machine learning algorithms for diabetes prediction. ICT Express (2021). DOI: 10.1016/j.icte.2021.02.004
7. Sisodia, D., Sisodia, D.S.: Prediction of Diabetes using Classification Algorithms. ProcediaComputer Science 132, 1578-1585 (2018).
8. Dutta I (n.d.) Data from: Early Stage Diabetes Risk Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset/data.
9. Sanghavi D (n.d.) Data from: Diabetes Prediction Sylhet Hospital Upgraded. Kaggle. https://www.kaggle.com/datasets/darsheelsanghavi/diabetes-prediction-sylhet-hospital-upgraded.