Efficient Diabetes Prediction Using Random Forests and Minimal Health Indicators on the BRFSS Dataset

Adnan Kutay Yüksel*[1] and Mehmet Serdar Güzel[2]

[1.] Department of Computer Engineering, Faculty of Engineering, Ankara University, Ankara 06830, Türkiye; (ORCID: 0000-0003-4057-3957)
[2.] Department of Computer Engineering, Faculty of Engineering, Ankara University, Ankara 06830, Türkiye; (ORCID: 0000-0002-3408-0083)
* Corresponding author: Adnan Kutay Yuksel (a.k.yuksel@hotmail.com)

Abstract: Early detection of diabetes is crucial for public health systems to implement timely interventions. In this study, we utilize the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset, particularly its balanced binary version, to build a Random Forest classifier for diabetes prediction. We begin with all 21 features and iteratively compare default, weighted, and hyperparameter-tuned models. Subsequently, we apply feature importance analysis to isolate the most significant predictors and retrain the model with a reduced feature set. Our tuned Random Forest model achieved an F1-score of 0.762 using all features. Notably, using only four features (GenHlth, HighBP, BMI, and Age), the model still achieved a robust F1-score of 0.751. These findings suggest that simpler models using fewer but high-impact features can be effectively deployed for diabetes prediction without sacrificing performance.

Keywords: Diabetes Prediction, Random Forest, Feature Selection, BRFSS Dataset, Healthcare Analytics

1. Introduction

Diabetes mellitus is a chronic metabolic disorder with increasing global prevalence and significant public health implications. According to the World Health Organization, the global prevalence of diabetes among adults has nearly quadrupled since 1980 (Zhou et al., 2016) and the latest official report on global diabetes prevalence comes from the International Diabetes Federation’s (IDF) Diabetes Atlas, which states that in 2021 approximately 537 million adults were living with diabetes worldwide and by 2045 an estimated 783 million people, equivalent of 12.2% of World population (Bergman et al., 2024). These statistics highlight both the growing clinical burden of diabetes and the urgent need for efficient, scalable screening and diagnostic methods to mitigate long-term health expenditures and improve patient outcomes (Ullah et al., 2022).

Recent advances in machine learning (ML) have enabled the development of predictive models that leverage health indicators derived from surveys or electronic medical records. Unlike traditional clinical assessments that may rely on expensive or invasive procedures, these models can offer scalable and cost-effective alternatives by utilizing easily collected data.

In this study, we leverage the 2015 binary-labeled and class-balanced subset of the BRFSS dataset to develop and optimize a Random Forest classifier. Random Forest (RF) is an ensemble technique that builds multiple decision trees using random subsets of features, thereby reducing overfitting while maintaining a high level of predictive accuracy. Prior research has supported the use of RF in diabetes prediction, with studies indicating that such tree-based ensemble models can achieve superior performance compared to other algorithms when applied to datasets like BRFSS (Mohamed et al., 2024; Liu et al., 2024)

Finally, by comparing the performance of models built on minimal versus full feature sets, we aim to provide practical insights into tailoring predictive models for diabetes. This analysis addresses not only the predictive performance in terms of standard classification metrics but also the trade-offs between model complexity and clinical interpretability. The potential to achieve high accuracy with reduced model complexity is particularly appealing for integration into cost-effective, scalable screening tools for diabetes, ultimately aiding public health policy and directing early intervention strategies (Horestani, 2024; Mohamed et al., 2024).

In summary, the increasing global prevalence of diabetes and its associated public health implications underscore the need for early prediction using efficient ML algorithms. The use of the BRFSS dataset combined with Random Forest classifiers, together with careful feature selection and data balancing techniques, offers a promising framework for developing predictive models that can outperform traditional clinical assessments while maintaining simplicity and interpretability. And we aim to explore two central research questions: (1) Can we achieve high prediction accuracy using a Random Forest model trained on a minimal subset of impactful features? and (2) How does the performance of simplified models compare with full-featured configurations in terms of standard classification metrics?

2. Background and Related Works

Recent studies in diabetes prediction predominantly employ the well-known Pima Indian Diabetes Dataset (PIDD) as a benchmark to test a variety of machine learning models. Conventional classifiers such as logistic regression, support vector machines (SVM), and decision trees have been extensively applied, while ensemble methods like Random Forests and XGBoost are frequently adopted for their ability to capture complex interactions within biomedical data (Asha et al., 2024).

Despite the robust performance reported with these models, a common limitation is that many investigations are restricted to relatively small or homogeneous datasets such as PIDD. This focus risks underrepresenting the variability and complexities of real-world clinical datasets (Kaliappan et al., 2024). In addition, although several studies incorporate feature selection techniques to filter out less informative attributes, few have systematically examined the impact of reducing the feature space on predictive power when moving to larger and more balanced datasets like the Behavioral Risk Factor Surveillance System (BRFSS) dataset (Asha et al., 2024).

The literature emphasizes that feature selection, whether implemented through filter-based methods (e.g., Pearson correlation, Chi-square) or wrapper methods (e.g., recursive feature elimination, Boruta), can significantly improve model interpretability while maintaining, or even enhancing, accuracy. For example, some studies have applied these techniques post hoc to refine classifiers trained on PIDD, demonstrating that reduced feature subsets can yield comparable performance to models leveraging the full set of features (Ashisha et al., 2024).

Our work seeks to address these identified gaps by providing a systematic benchmarking of default, weighted, and optimized Random Forest models. The approach first entails an extensive evaluation of these models using standard settings and weighted adjustments to account for potential class imbalances present in the dataset (Abousaber et al., 2024). Subsequently, we apply rigorous feature selection methodologies to identify a subset of high-importance features, after which the models are retrained on this reduced feature space. By doing so, we aim to ascertain whether similar, or even superior, predictive power can be achieved with fewer features on large and balanced datasets like those drawn from BRFSS (Ayoade et al., 2024).

3. Dataset and Preprocessing

The 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset provides comprehensive survey data collected across the United States on health-related risk behaviors, chronic health conditions, and use of preventive services. This dataset is particularly valuable for analyzing patterns related to diabetes prevalence, as it includes a wide range of demographic, lifestyle, and health status variables. Its large sample size and standardized methodology make it a reliable source for machine learning-based predictive modeling and public health research. The dataset is available on Kaggle: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/

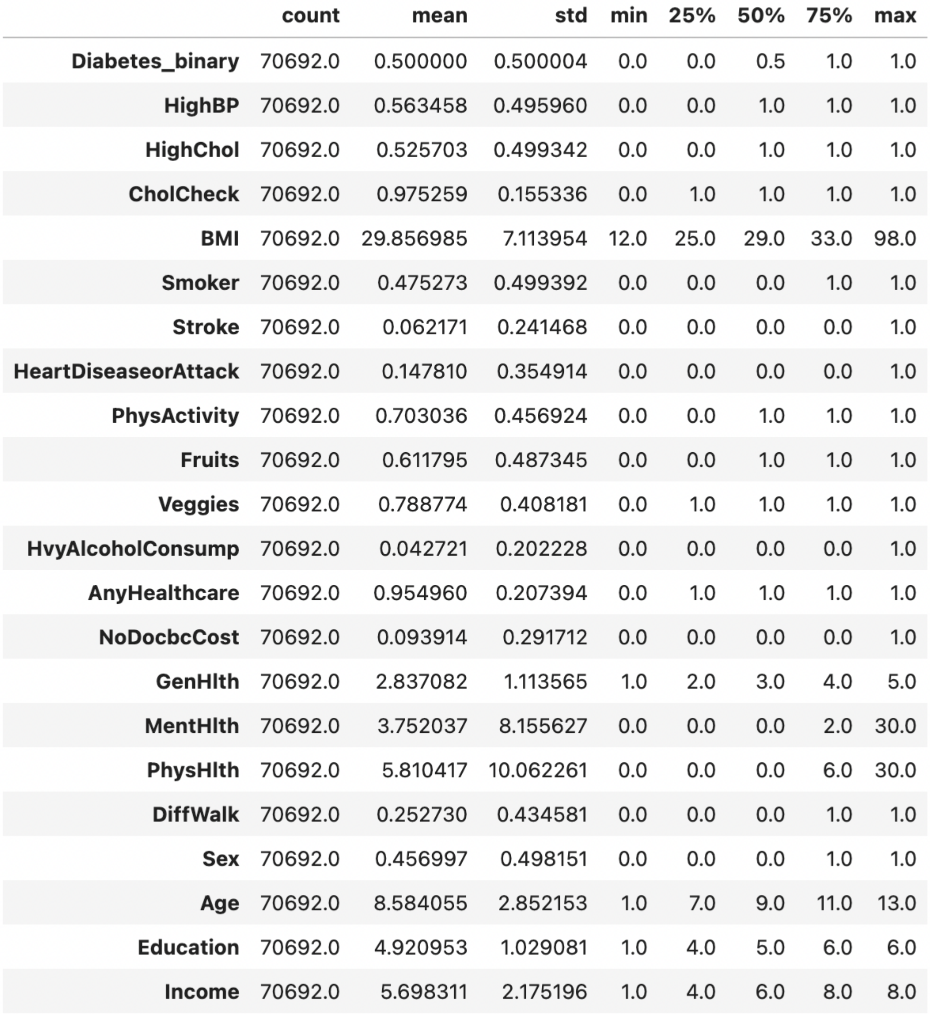
Within this dataset there are three CSV dataset files, each containing 21 health-related feature variables. The first dataset includes 253,680 responses with a three-class target variable (0: no diabetes or only during pregnancy, 1: prediabetes, 2: diabetes), but suffers from class imbalance. The second dataset contains 70,692 responses with a binary target variable (0: no diabetes, 1: prediabetes or diabetes) and features a balanced 50-50 class distribution. The third dataset, like the first, includes 253,680 responses with a binary target variable but remains imbalanced. Among these, the second dataset was selected for

model development due to its balanced nature, which helps reduce bias and supports more reliable model evaluation.
Key features include general health ratings, BMI, blood pressure status, physical activity, fruit and vegetable consumption, alcohol use, healthcare access, and demographic variables such as age, sex, education, and income.

We performed the following preprocessing steps:

1. Removed no columns due to missing values (dataset was clean)
2. Used Scikit-learn's train_test_split with an 80/20 train-test ratio
3. No normalization was required due to the tree-based nature of Random Forests

Dataset structure is provided below:



	count	mean	std	min	25%	50%	75%	max
Diabetes_binary	70692.0	0.500000	0.500004	0.0	0.0	0.5	1.0	1.0
HighBP	70692.0	0.563458	0.495960	0.0	0.0	1.0	1.0	1.0
HighChol	70692.0	0.525703	0.499342	0.0	0.0	1.0	1.0	1.0
CholCheck	70692.0	0.975259	0.155336	0.0	1.0	1.0	1.0	1.0
BMI	70692.0	29.856985	7.113954	12.0	25.0	29.0	33.0	98.0
Smoker	70692.0	0.475273	0.499392	0.0	0.0	0.0	1.0	1.0
Stroke	70692.0	0.062171	0.241468	0.0	0.0	0.0	0.0	1.0
HeartDiseaseorAttack	70692.0	0.147810	0.354914	0.0	0.0	0.0	0.0	1.0
PhysActivity	70692.0	0.703036	0.456924	0.0	0.0	1.0	1.0	1.0
Fruits	70692.0	0.611795	0.487345	0.0	0.0	1.0	1.0	1.0
Veggies	70692.0	0.788774	0.408181	0.0	1.0	1.0	1.0	1.0
HvyAlcoholConsump	70692.0	0.042721	0.202228	0.0	0.0	0.0	0.0	1.0
AnyHealthcare	70692.0	0.954960	0.207394	0.0	1.0	1.0	1.0	1.0
NoDocbcCost	70692.0	0.093914	0.291712	0.0	0.0	0.0	0.0	1.0
GenHlth	70692.0	2.837082	1.113565	1.0	2.0	3.0	4.0	5.0
MentHlth	70692.0	3.752037	8.155627	0.0	0.0	0.0	2.0	30.0
PhysHlth	70692.0	5.810417	10.062261	0.0	0.0	0.0	6.0	30.0
DiffWalk	70692.0	0.252730	0.434581	0.0	0.0	0.0	1.0	1.0
Sex	70692.0	0.456997	0.498151	0.0	0.0	0.0	1.0	1.0
Age	70692.0	8.584055	2.852153	1.0	7.0	9.0	11.0	13.0
Education	70692.0	4.920953	1.029081	1.0	4.0	5.0	6.0	6.0
Income	70692.0	5.698311	2.175196	1.0	4.0	6.0	8.0	8.0

Figure 1. Dataset Feature Overview Table

4. Methodology

We employed Scikit-learn's implementation of the Random Forest classifier with the following experimental configurations:

- **Baseline model** with default parameters
- **Balanced model** using class_weight='balanced'
- **Tuned model** with n_estimators=200, max_depth=10, min_samples_split=10

For feature importance evaluation, we extracted feature_importances_ from the tuned model and sorted them in descending order. We then retrained models using only the top 4, 8, 12, and 16 features. Performance metrics included accuracy, F1-score, precision, recall, and confusion matrix evaluation. The steps involved are:

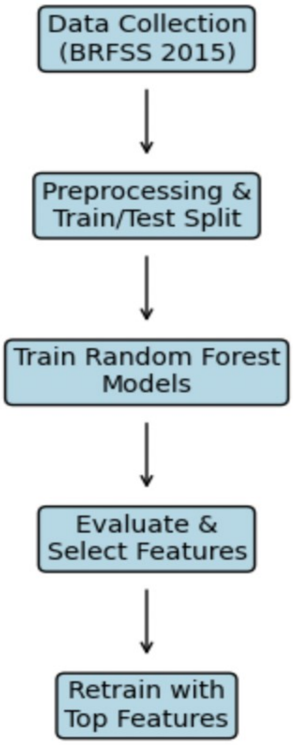

Data Collection (BRFSS 2015)
Preprocessing & Train/Test Split
Train Random Forest Models
Evaluate & Select Features
Retrain with Top Features

Figure 2. Methodology Flowchart

5. Results

5.1 Model Comparisons

While the fully tuned RF model using all 21 features achieved the highest performance (Accuracy = 0.752, F1 Score = 0.762), reducing the feature set to the top 16 resulted in only a slight decrease in performance (Accuracy = 0.749, F1 Score = 0.756). Further reductions to 12, 8, and 4 features caused gradual declines in both metrics, with the model still maintaining competitive results (e.g., F1 Score = 0.751 for 4 features). These findings, as shown in Table 1, indicate that a smaller feature set can preserve most of the full model's predictive power, highlighting a potential trade-off between performance and simplicity.

Table 1. Model Accuracy and F1 Score Comparison

Model	Feature Count	Accuracy	F1 Score
Default RF	21	0.737	0.747
RF + Balanced	21	0.738	0.748
RF Tuned	21	0.752	0.762
Top 16 Features (after having tuned the model)	16	0.749	0.756
Top 12 Features (after having tuned the model)	12	0.748	0.755
Top 8 Features (after having tuned the model)	8	0.745	0.751
Top 4 Features (after having tuned the model)	4	0.741	0.751

5.2 Feature Importance

Feature importances before having tuned the model are illustrated in Figure 3 below:

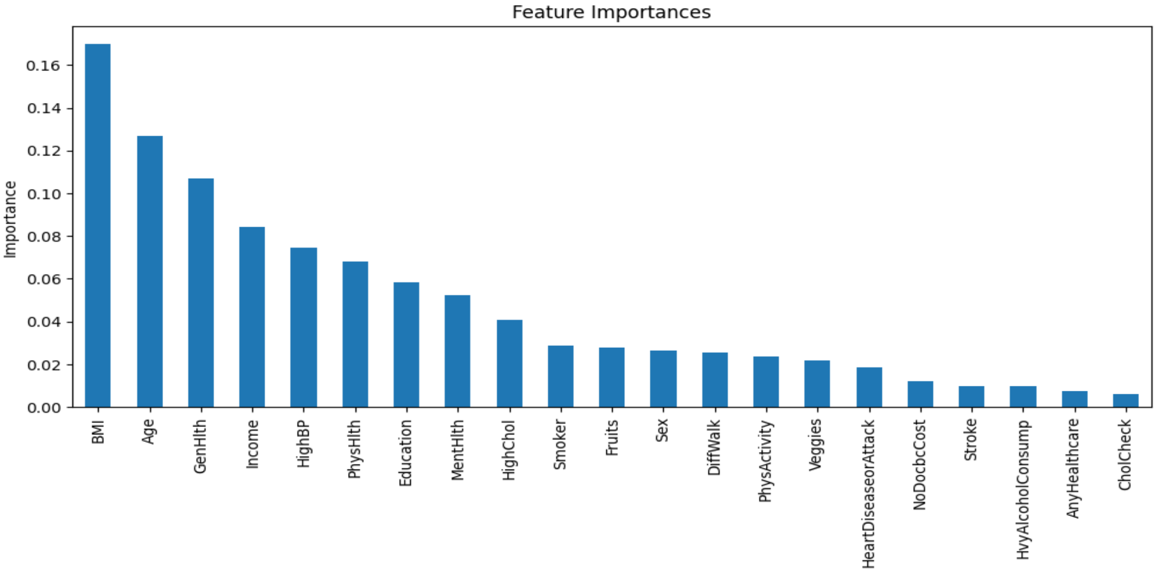


Figure 3. Feature importances before having tuned the model

This visualization shows the baseline importance values of all features before any hyperparameter tuning was applied. It highlights which variables initially had the greatest influence on the model's decisions.

After having tuned the model, our most important feature findings have changed as illustrated in Figure 4**:**

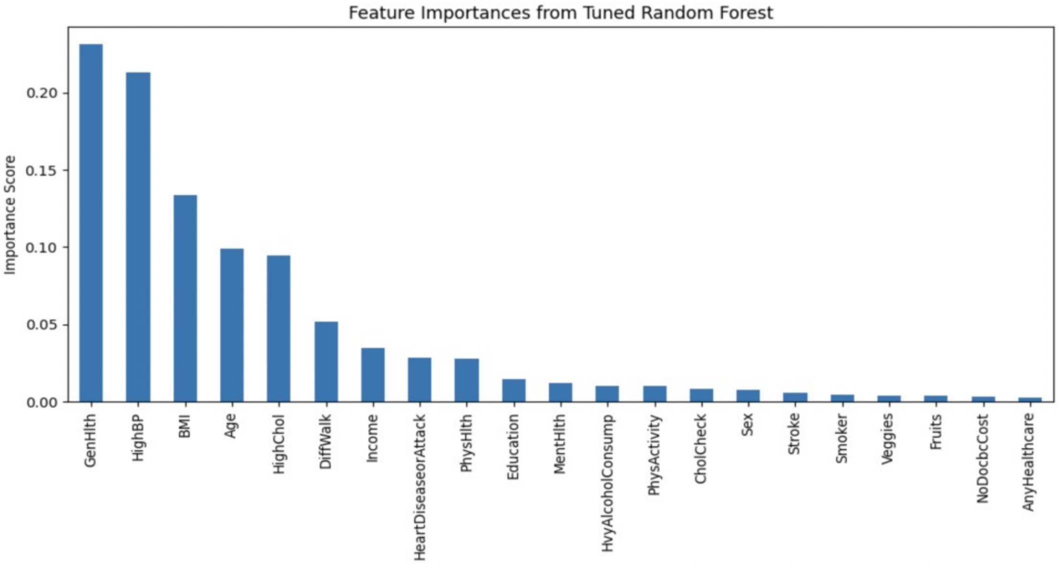


Figure 4. Feature importances after having tuned the model

This figure demonstrates how feature importance rankings shifted following hyperparameter optimization. It reflects a refined model understanding, emphasizing the predictive power of a different set of variables compared to the baseline.

Eventually, the top four most important features are:

1. General Health (GenHlth)
2. High Blood Pressure (HighBP)
3. Body Mass Index (BMI)
4. Age

And the final optimized tree using the top 4 features revealed the following forest, as shown in Figure 5**:**

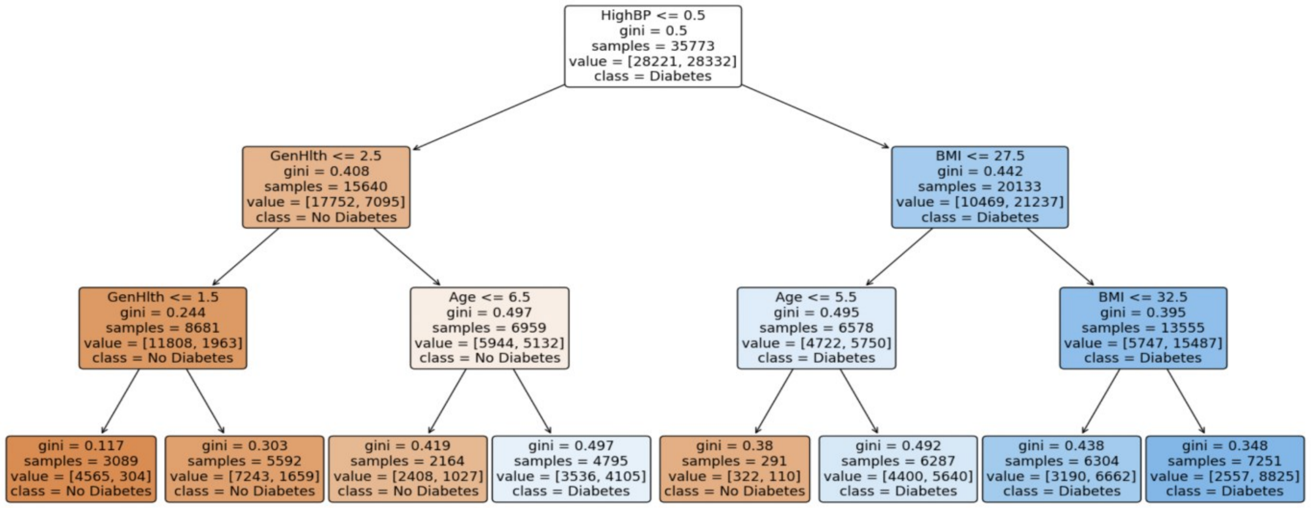


Figure 5. Random Forest Decision Tree Using Top 4 Features (Optimized – Our Approach)

While Figure 5 displays a clear and concise tree, Figure 6 displays a full-depth decision tree trained using all 21 features. This tree is grown without depth limitation and illustrates the inherent complexity of the feature-rich model. Compared to the simplified tree in Figure 5, it demonstrates how much decision paths can be compressed with proper feature selection, without compromising accuracy.

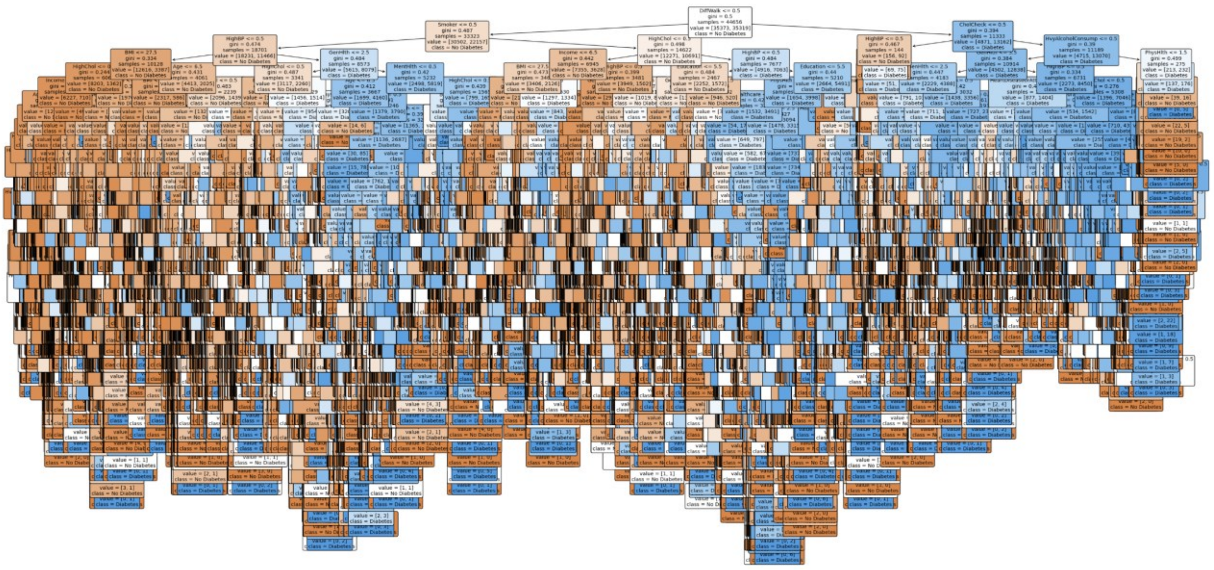


Figure 6. Random Forest Decision Tree Using All 21 Features (Unoptimized –Baseline)

This is exactly where our work comes in: demonstrating that interpretability and performance can coexist, and that smart feature selection can make complex models practically deployable.

5.3 Confusion Matrix and Classification Metrics

To evaluate the performance of our final model, we utilized a confusion matrix and a detailed classification report.

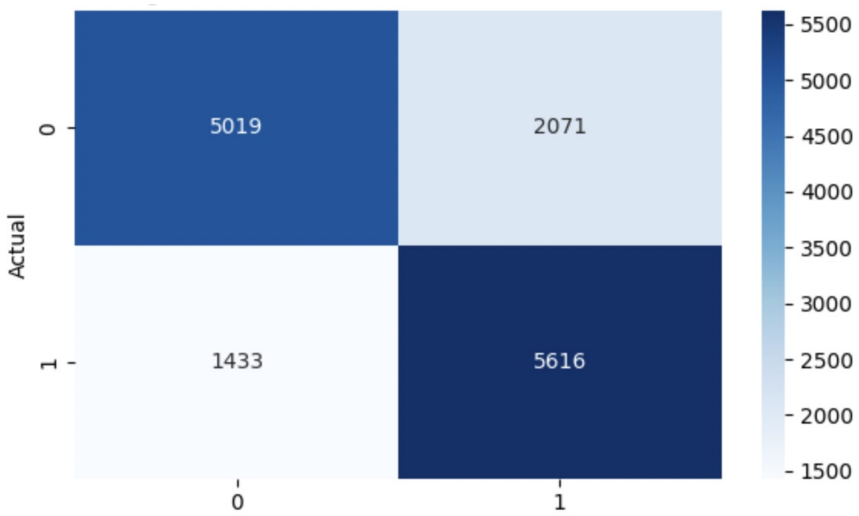
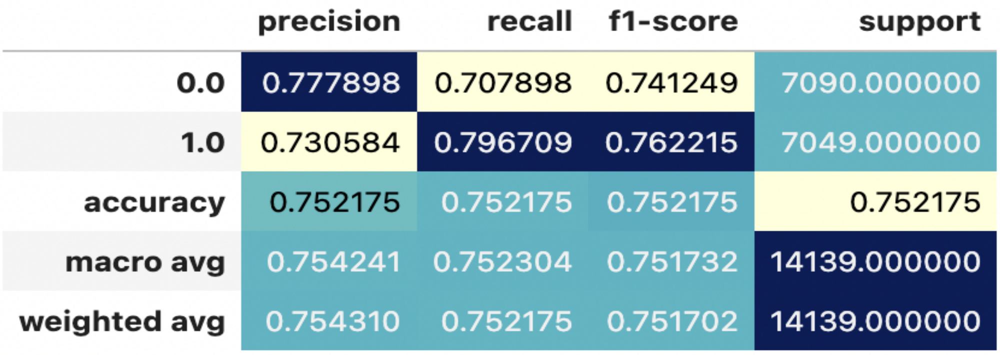


Figure 7. Confusion Matrix for Tuned Model

As shown in Figure 7, the confusion matrix illustrates the distribution of correct and incorrect predictions. The model correctly predicted 5,019 true negatives and 5,616 true positives, while there were 2,071 false positives and 1,433 false negatives. This matrix highlights the model's relative balance in classifying both classes, although some false positives and false negatives still exist.



	precision	recall	f1-score	support
0.0	0.777898	0.707898	0.741249	7090.000000
1.0	0.730584	0.796709	0.762215	7049.000000
accuracy	0.752175	0.752175	0.752175	0.752175
macro avg	0.754241	0.752304	0.751732	14139.000000
weighted avg	0.754310	0.752175	0.751702	14139.000000

Figure 8. Classification Report Table

Figure 8 provides precision, recall, and F1-score metrics for each class:

1. **Class 0.0 (Negative Class)**: Precision = 0.778, Recall = 0.708, F1-score = 0.741
2. **Class 1.0 (Positive Class)**: Precision = 0.731, Recall = 0.797, F1-score = 0.762
3. **Overall Accuracy**: 75.2%

The model performs slightly better in detecting the positive class (1.0) in terms of recall, but slightly better in precision for the negative class (0.0). The balance between precision and recall, especially in F1-score, shows that the model generalizes well across both classes.

5.4 Accuracy Comparison

To evaluate the effectiveness of our model, we compared its performance with two recent studies, focusing on the number of features used, overall accuracy, and AUC (Area Under the Curve) values. The comparison is summarized in Table 2:

Table 2. Comparison of Accuracy and AUC with Recent Studies

	No. of Features	Accuracy	AUC
Our Study	4	0.75	0.83
Horestani (2024)	6	0.71	0.7743
Koushik et al (2023)	21	0.74	0.81

As shown in Table 2, our model achieves an accuracy of 0.75 and an AUC of 0.83, using only 4 features**.** In contrast, Horestani (2024) used 6 features and reached an accuracy of 0.71 with an AUC of 0.7743. Koushik et al. (2023), despite using 21 features, achieved slightly lower accuracy (0.74) and AUC (0.81) than our model.

This comparison demonstrates the strength of our feature selection and optimization process, showing that a smaller set of well-selected features can yield not only a more interpretable model but also superior performance.

5.5 Accuracy Comparison

To further evaluate the classification performance of our model, we analyzed the Receiver Operating Characteristic (ROC) curve, which illustrates the model's ability to differentiate between the two classes at various threshold settings.

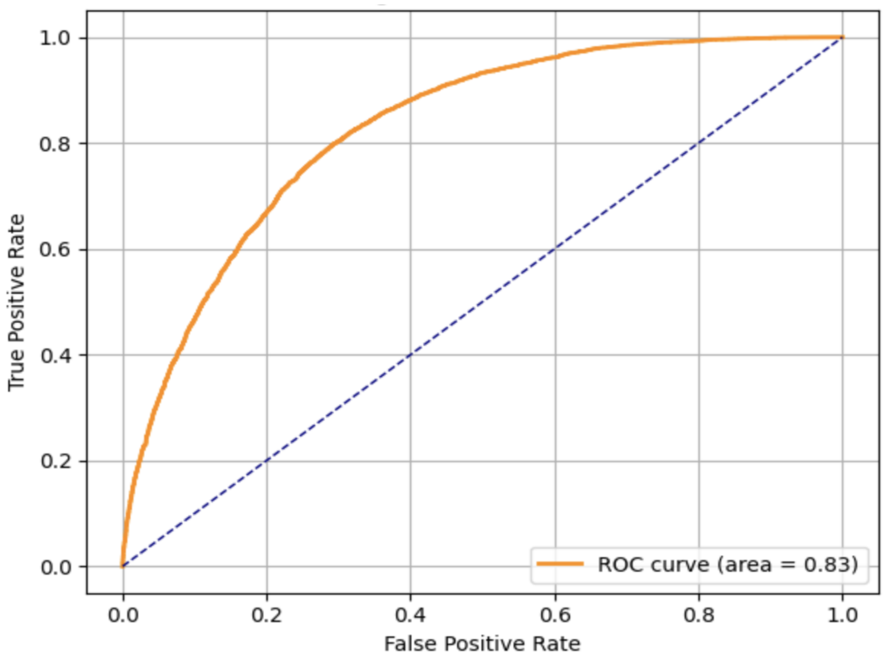


Figure 9. ROC Curve

As depicted in Figure 9, the ROC curve demonstrates a high true positive rate (TPR) across a wide range of false positive rates (FPR). The Area Under the Curve (AUC) is calculated as 0.83**,** which indicates strong classification performance.

The smooth upward curve and the large AUC value reflect the model's high discriminative capacity, even with a reduced number of features (only 4). This reinforces the effectiveness of our feature selection and tuning strategy, emphasizing that optimal performance can be achieved without relying on complex or high-dimensional models.

6. Discussion

The results indicate that the Random Forest classifier performs well on the balanced BRFSS dataset even with a significant reduction in feature dimensionality. The minimal performance drop from using 21 features to using just four suggests that most predictive value is concentrated in a few variables. This insight supports the design of lightweight diagnostic tools for real-time screening. Importantly, such simplification improves model interpretability and reduces data collection burden, making it more feasible for large-scale deployment, particularly in resource-limited healthcare environments.

While advanced models such as XGBoost may offer marginal improvements, Random Forests remain highly interpretable and easy to implement. Further studies can validate these findings across different datasets and include temporal, regional, or demographic segmentation. It should also be noted that the BRFSS dataset is based on self-reported responses, which may introduce recall or reporting biases. Therefore, the generalizability of the results beyond the BRFSS population may be limited and should be further assessed through external validation.

7. Conclusion

This study demonstrates that a carefully tuned Random Forest model can achieve strong performance in predicting diabetes using only four features from a balanced health survey dataset. The proposed model strikes a favorable balance between simplicity, interpretability, and predictive power.

Our approach can inform the design of decision support tools in primary healthcare, telemedicine platforms, and public health surveillance systems. Future work should explore cross-dataset generalization, longitudinal prediction, and incorporation of domain knowledge through hybrid model structures.

Author Contributions: Conceptualization, A.K.Y. and M.S.G.; Methodology, A.K.Y.; Software, A.K.Y.; Validation, A.K.Y.; Formal Analysis, M.S.G.; Investigation, A.K.Y.; Resources, A.K.Y.; Writing – Original Draft Preparation, A.K.Y.; Writing – Review & Editing, M.S.G.; Visualization, A.K.Y. and M.S.G.; Supervision, M.S.G.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

Abousaber, I., Abdallah, H. F., & El-Ghaish, H. (2025). Robust predictive framework for diabetes classification using optimized machine learning on imbalanced datasets. *Frontiers in Artificial Intelligence*, *7*, 1499530.

Aliverdi, F. (2024). Reproducibility in Diabetes Research Articles Using Machine Learning Classifiers.

Asha, V., Lamani, M. R., Padmaja, K., & Gondhale, T. A. (2024). Enhancing diabetes prediction accuracy with AdvanSVM: A machine learning approach using the PIMA dataset. *Seybold Report Journal, 19*(05), 51-72.

Ashisha, G. R., Mary, X. A., Kanaga, E. G. M., Andrew, J., & Eunice, R. J. (2024). Random Oversampling-Based Diabetes Classification via Machine Learning Algorithms. *International Journal of Computational Intelligence Systems*, *17*(1), 270.

Ayoade, O. B., Shahrestani, S., & Ruan, C. (2025). Machine Learning and Deep Learning Approaches for Predicting Diabetes Progression: A Comparative Analysis.

Bergman, M., Manco, M., Satman, I., Chan, J., Schmidt, M. I., Sesti, G., ... & Tuomilehto, J. (2024). International Diabetes Federation Position Statement on the 1-hour post-load plasma glucose for the diagnosis of intermediate hyperglycaemia and type 2 diabetes. *Diabetes research and clinical practice*, *209*, 111589.

Horestani, F. J. (2024). Predicting Diabetes with Machine Learning Analysis of Income and Health Factors. *arXiv preprint arXiv:2404.13260*.

Kaliappan, J., Saravana Kumar, I. J., Sundaravelan, S., Anesh, T., Rithik, R. R., Singh, Y., ... & Srinivasan, K. (2024). Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets. *Frontiers in Artificial Intelligence*, *7*, 1421751.

Koushik, P. S., Indira, B., & Ponnala, I. R. Dia-Analyze: A Comprehensive Data Analytics Suite for Type 2 Diabetes.

Liu, Z., Zhang, Q., Zheng, H., Chen, S., & Gong, Y. (2024, October). A Comparative Study of Machine Learning Approaches for Diabetes Risk Prediction: Insights from SHAP and Feature Importance. In *2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)* (pp. 35-38). IEEE.

Mohamed, M. H., Khafagy, M. H., Mohamed, N., Kamel, M., & Said, W. (2024). Diabetic mellitus prediction with brfss data sets. *Journal of Theoretical and Applied Information Technology*, *102*(3).

Ullah, Z., Saleem, F., Jamjoom, M., Fakieh, B., Kateb, F., Ali, A. M., & Shah, B. (2022). Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods. *Computational Intelligence and Neuroscience*, *2022*(1), 2557795.

Zhou, B., Lu, Y., Hajifathalian, K., Bentham, J., Di Cesare, M., Danaei, G., ... & Gaciong, Z. (2016). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants. *The lancet*, *387*(10027), 1513-1530.