# Novel Ensembled Approach To Healthcare Data Analysis

# Introduction

- Diabetes mellitus (DM) is a prevalent chronic disease worldwide, affecting millions of individuals.
- Our project aims to develop a predictive model that can assess the risk of diabetes based on various health parameters.
- Leveraging machine learning algorithms, we seek to create a tool that aids in early detection and management of diabetes, ultimately improving healthcare outcomes.
- Improvement: Regular updates and detailed documentation to maintain effectiveness and support future research.

# Problem Statement

"DEVELOPING A NOVEL ENSEMBLE APPROACH TO ENHANCE THE ACCURACY AND RELIABILITY OF HEALTHCARE DATA ANALYSIS TO PREDICT DIABETES DISEASE."

# Motivation

- Enhancing Accessibility to Healthcare
- Enhancing Patient Outcomes
- Reducing Healthcare Costs
- Empowering Personalized Healthcare
- Research and Development

# *Objectives*

- To increase the accuracy of the model using ensembled technique
- To provide more stable and reliable predictions by mitigating the weaknesses of individual models
- To performs well on unseen data, improving its applicability in real-world scenarios.
- To Validats model performance through techniques like k-fold cross-validation, ensuring robust evaluation.

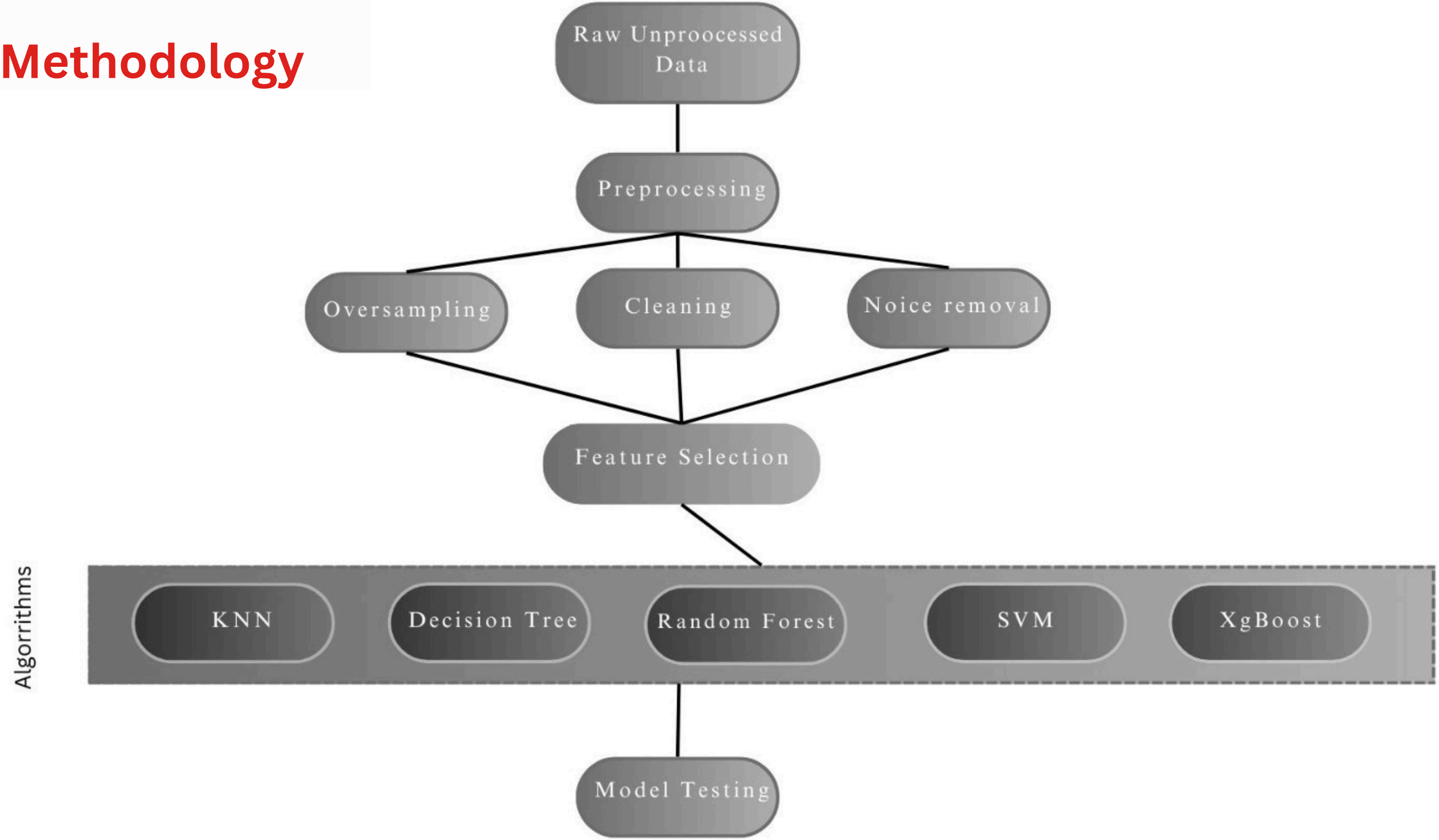
# Liturature Survey:

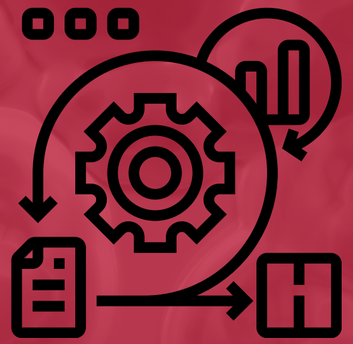
| Research Papers | Description | Author | Publucation &Year of Publication |
|---|---|---|---|
| CatBoost Ensemble Approach for Diabetes Risk Prediction at Early Stages | In this paper, we propose an ensemble technique CatBoost which is a Gradient Boosting Decision Tree (GBDT) for diabetes prediction at early stages. | P. Suresh Kumar1 Anisha Kumari K Subhashree Mohapatra | IEEE 2021 |
| Genome-wide association analysis of type 2 diabetes in the EPICInterAct study | The EPIC-InterAct project, centred in 8 countries in the European Prospective Investigations into Cancer and Nutrition study, is one of the largest prospective studies of T2D. | Lina Cai | 2020 |
| Genomic Prediction of 16 Complex Disease Risks Including Heart Attack, Diabetes, Breast and Prostate Cancer | Our results indicate that substantial improvements in predictive power are attainable using training sets with larger case populations. We anticipate rapid improvement in genomic prediction as more casecontrol data become available for analysis. | Louis Lello TimothyG. Raben SokeYuenYong | 2019 |

| | | | |
|---|---|---|---|
| Random Forest Algorithm for the Prediction of Diabetes | The literature introduces a weighted ensembling approach, where weights are determined based on the Area Under ROC Curve (AUC) of each ML model. AUC is chosen as the performance metric, optimized through hyperparameter tuning using the grid search technique. | Md. Kamrul Hasan | IEEE 2019 |
| HYBRID PREDICTION MODEL FOR TYPE-2 DIABETES WITH CLASS IMBALANCE | This paper presents an efficient model for classifying type-2 diabetes using a hybrid approach. The dataset, obtained from Vanderbilt University, USA, focuses on 390 African-American individuals and includes parameters such as Glucose, Age, Cholesterol, Weight, and Waist/Hip ratio. To address imbalance, oversampling and undersampling methods are applied. | Rishi Kashyap Surya Teja CVN | IEEE 2020 |
| Predictive Diabetes Mellictus by DNA sequences using Deep Learning | Deep learning algorithms and artificial intelligence (AI) are introduced. Based on DNA sequencing, a recent study used Long Short-Term Memory (LSTM) algorithms and Convolutional Neural Networks (CNNs) to identify diabetes types. The results were remarkable: on a labeled dataset, the suggested CNN-LSTM model reached 100% accuracy, highlighting AI's promise in healthcare | Lena abed ALraheim Hamza, Hussein Attya Lafta. Sura Zaki Al-Rashid, | IEEE 2023 |
| Prediction of type 2 diabetes mellitus onset using logistic regressionbased scorecards | This study focused on analyzing data from 44,709 nondiabetic participants aged 40–69 in the UK Biobank to predict the risk of Type 2 Diabetes (T2D) onset within a mean timeframe of 7.3 years. Initially considering 798 potential predictors for T2D onset, the study employed gradient boosting decision trees, survival analysis, and logistic regression methods | Yochai Edlitz1 Eran Segal | 2020 |

| | | | |
|---|---|---|---|
| Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction | A novel ensemble WeightedVotingLRRFs ML model is introduced to enhance diabetes prediction, achieving an Area Under the ROC Curve (AUC) of 0.884. The model incorporates optimal weights determined through a bi-objective genetic algorithm based on Sensitivity and AUC. | NIKOS FAZAKIS<br>OTILIA KOCSIS | IEEE<br>2021 |
| Prediction Model for Type 2 Diabetes using Stacked Ensemble Classifiers | The study employs 10-fold cross-validation and assesses performance using precision, recall, F-measure, and accuracy metrics. Experimental results demonstrate the superiority of the proposed model, achieving high accuracy rates of up to 93.18%, 98.87%, and 96.09% for datasets I, II, and III, respectively. | Norma Latif Fitriyani<br>Muhammad Syafrudin | 2020 |
| Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers | In this literature, we are proposing a robust framework for diabetes prediction where the outlier rejection, filling the missing values, data standardization, feature selection, K-fold cross-validation, and different Machine Learning classifiers ere employed. T | MD. KAMRUL HASAN1<br>MD. ASHRAFUL ALAM1 | 2020 |

# Methodology





# Proposed Methodology

- Data Collection and Preprocessing
- Model Selection and Ensemble Techniques
- Model Evaluation
- Cross-Validation and Generalization
- Implementation and Deployment

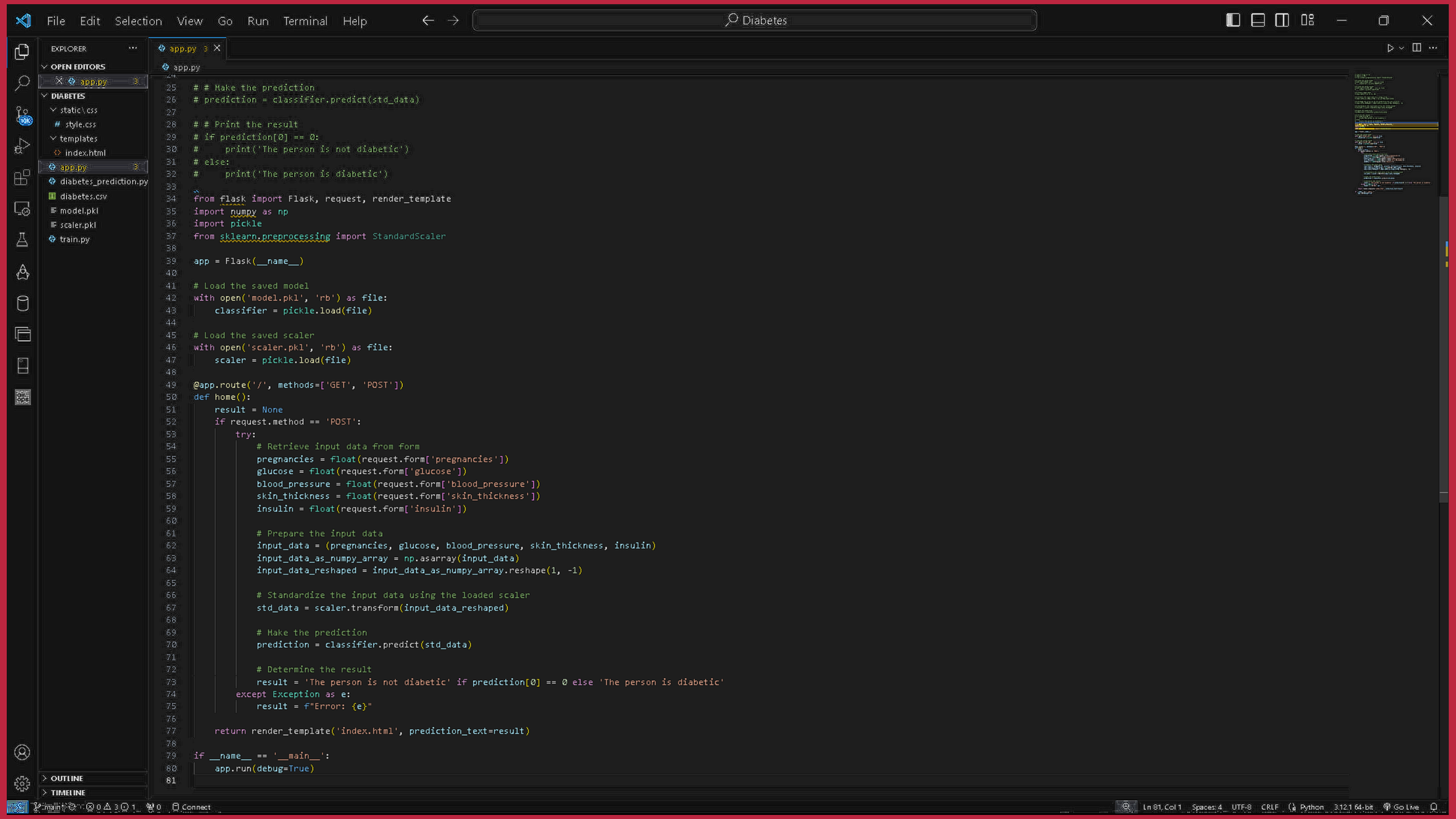
# Algorithms Used

1. K-Nearest Neighbors (KNN)
2. Support Vector Machines (SVM)
3. Decision Trees
4. Random Forest
5. XGBoost (Extreme Gradient Boosting)

# Experimental Results

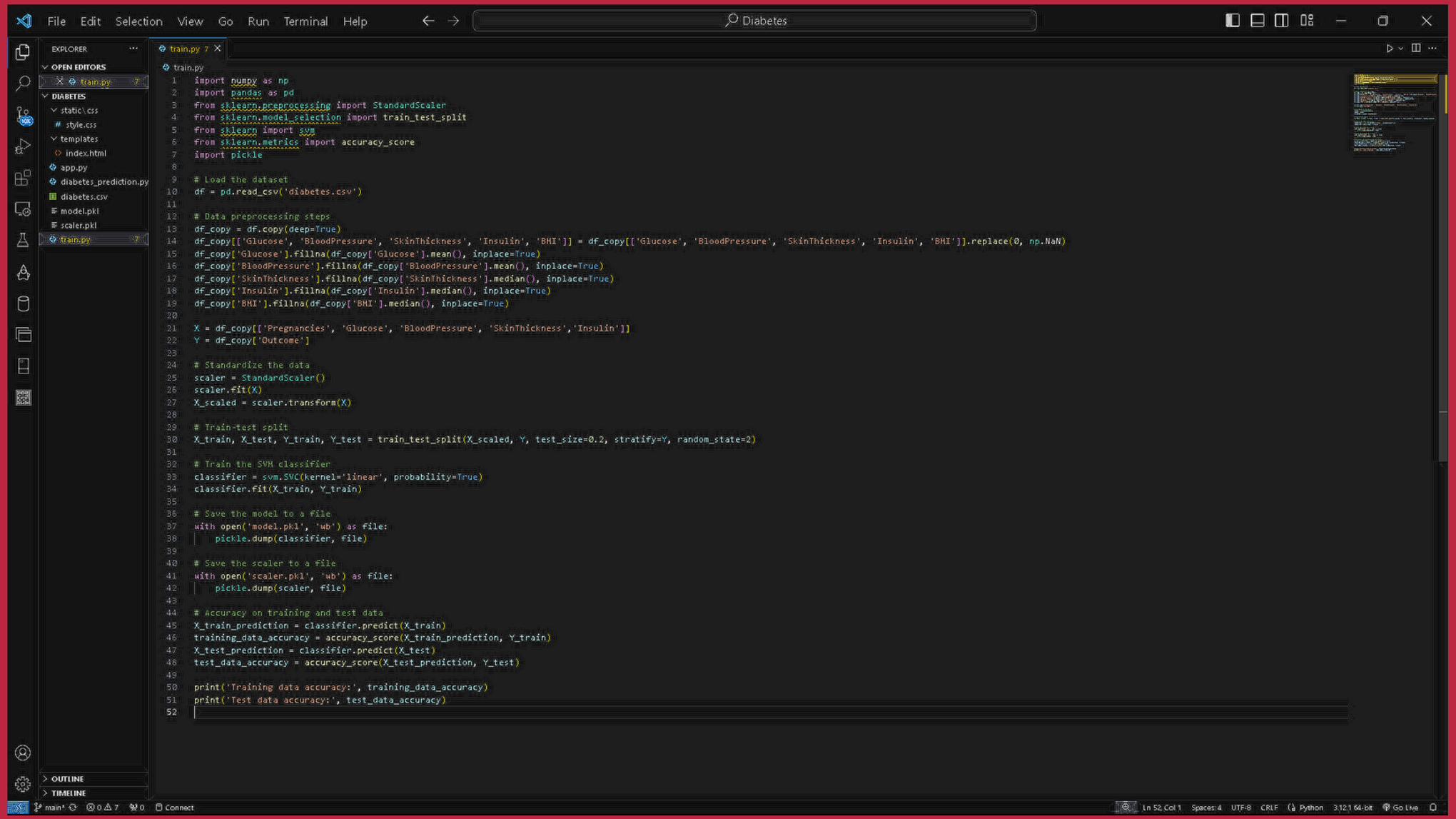
| Algorithms | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| KNeighboursClassifier() | 69% | 60% | 39% | 47% |
| SVM | 76% | 75% | 50% | 60% |
| Decision Tree | 68% | 57% | 46% | 51% |
| Random Forest | 72% | 63% | 50% | 53% |
| XgBoost | 75% | 65% | 65% | 65% |

# Implementation



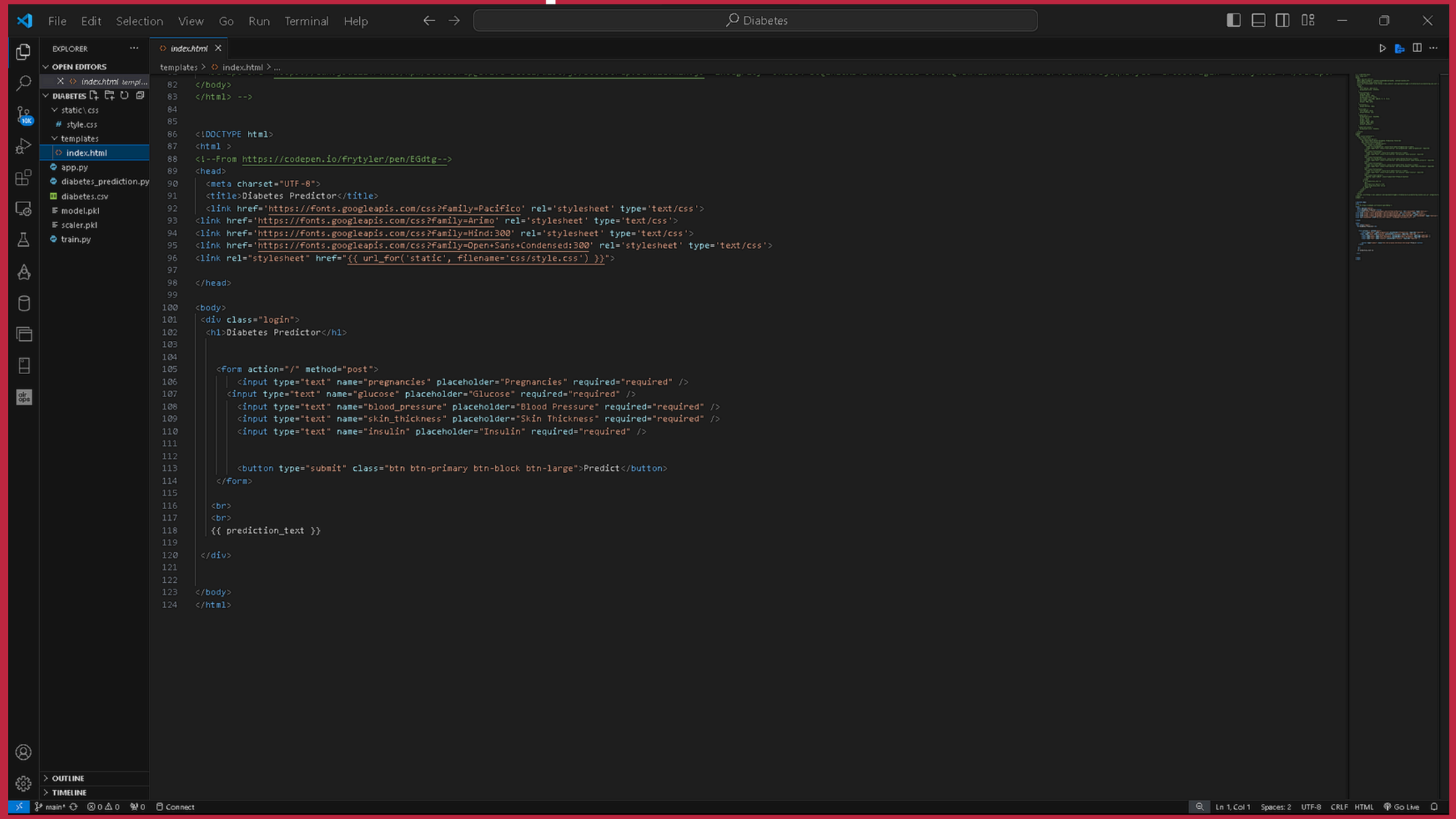

Flask App implementation

# Implementation



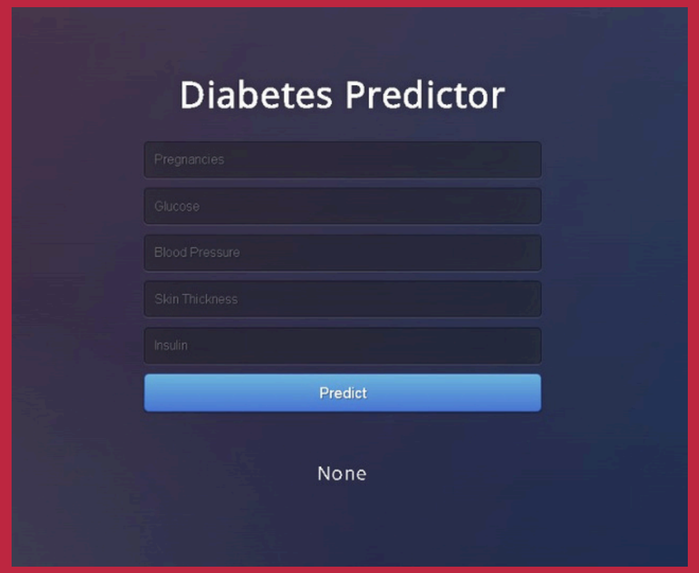

**Model Training**

# Implementation

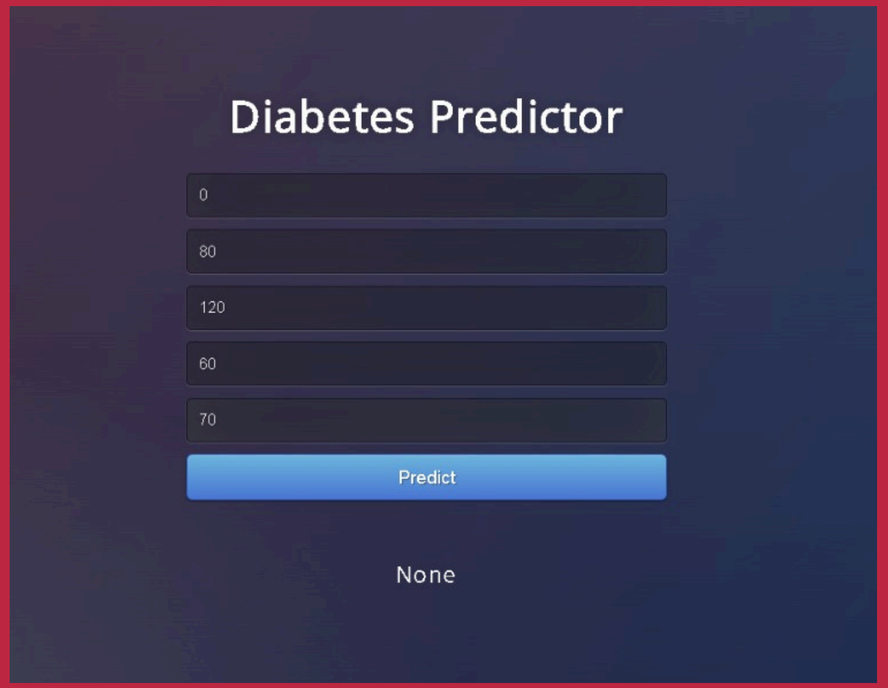



User interface implementaion

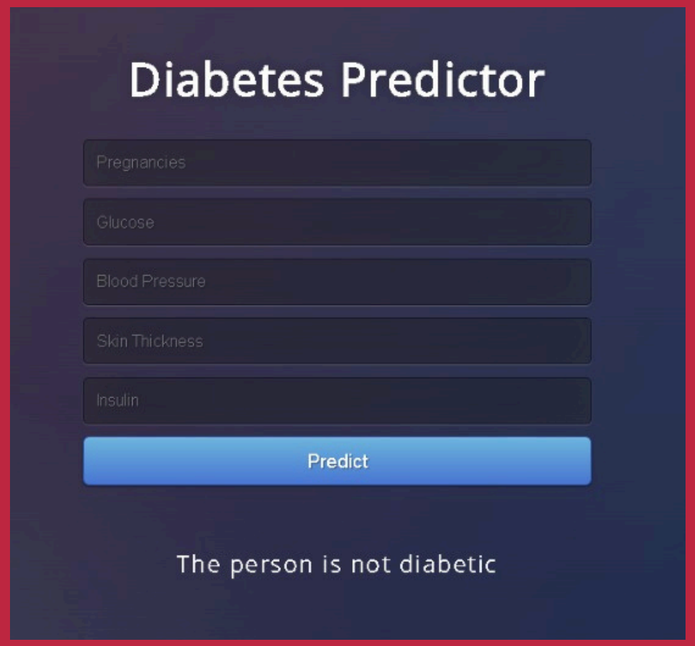
# Our Model




**Home Page**




**User Input**




**Output**

# Future Scope

1.Integration of Genetic Data

2. Real-time Monitoring Systems

3. Expansion to Other Chronic Diseases

4. Explore Additional Algorithms

# Conclusion

- Ensemble methods significantly improve the accuracy of diabetes prediction by combining the strengths of multiple models.
- Ensemble techniques effectively reduce both bias and variance, leading to better generalization on unseen data.
- The flexibility to combine different algorithms allows for the creation of highly customized models tailored to specific prediction tasks.

# References


1. Hamza, L.A. et al. (2023). Predictive Diabetes Mellitus From DNA Sequences Using Deep Learning. Al-Bahir Journal for Engineering and Pure Sciences, 3(2). https://doi.org/10.55810/2313-0083.1042

2. Zou, Q. et al. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. Front. Genet. 9:515. https://doi.org/10.3389/fgene.2018.00515

3. Das, B. (2022). A deep learning model for identification of diabetes type 2 based on nucleotide signals. Neural Comput & Applic, 34, 12587–12599. https://doi.org/10.1007/s00521-022-07121-8

4. Kim, J. et al. (2018). Genetic prediction of type 2 diabetes using deep neural network. Clin Genet, 93(4), 822-829. doi: 10.1111/cge.13175.

5. Alshamlan, H. et al. (2020). A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression. doi: 10.1109/ICICS49469.2020.239549.

6. Kumar, A. et al. (2017). SVMRFE based approach for prediction of most discriminatory gene target for type II diabetes. Health Cares Data, 12, 28-37. https://doi.org/10.1016/j.gdata.2017.02.008.

7. Lello, L. et al. (2019). Genomic Prediction of 16 Complex Disease Risks Including Heart Attack, Diabetes, Breast and Prostate Cancer. Sci Rep, 9(1):15286. https://doi.org/10.1038/s41598-019-51258-x.

8. Loh, M. et al. (2022). Identification of genetic effects underlying type 2 diabetes in South Asian and European populations. Commun Biol, 5(1):329. https://doi.org/10.1038/s42003-022-03248-5.

9. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, & others. (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science, 316(5829), 1331-6. https://doi.org/10.1126/science.1142358.

10. Cai, L. et al. (2020). Genome-wide association analysis of type 2 diabetes in the EPIC-InterAct study. Sci Data, 7(1):393. https://doi.org/10.1038/s41597-020-00716-7.

11. Cole, J.B. & Florez, J.C. (2020). Genetics of diabetes mellitus and diabetes complications. Nat Rev Nephrol, 16(7), 377-390. https://doi.org/10.1038/s41581-020-0278-5.

12. Cho, N.H. & others. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res Clin Pract, 138, 271–281. https://doi.org/10.1016/j.diabres.2018.02.023.

13. Butt, U.M. et al. (2021). Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. J Healthc Eng, 2021, 9930985. https://doi.org/10.1155/2021/9930985.

14. Oliullah, K. et al. (2023). A stacked ensemble machine learning approach for the prediction of diabetes. J Diabetes Metab Disord. https://doi.org/10.1007/s40200-023-01321-2.

15. Tripathi, D. et al. (2022). Diabetes Prediction Using Machine Learning Analytics: Ensemble Learning Techniques.

16. Nagpal, A. et al. (2023). A novel ensemble machine learning framework for early stage diabetes mellitus prediction. Multidisciplinary Science Journal.

17. National Diabetes Statistics Report | Diabetes | Centers for Disease Control and Prevention. 2022. https://www.cdc.gov/diabetes/data/statistics-report/index.html.

18. Hosseini Sarkhosh, S.M. et al. (2022). Predicting diabetic nephropathy in type 2 diabetic patients using machine learning algorithms. J Diabetes Metab Disord, 21(2), 1433–41.

19. Hemanth, S. & Alagarsamy, S. (2023). Hybrid adaptive deep learning classifier for early detection of diabetic retinopathy using optimal feature extraction and classification. J Diabetes Metab Disord, 1–15.