Using Machine Learning Algorithms to Perform Diabetes Probability Prediction from Multiple Risk Factors in Diabetes Health Indicators Dataset

Wenye Song

Emory University

Introduction

Diabetes stands as a significant chronic disease in the United States, affecting over 34 million individuals and placing a substantial economic and health burden on the country. In addition to those diagnosed, approximately 88 million Americans are living with prediabetes or diabetes, and many of them are unaware of their medical condition. Diabetes can lead to severe health issues such as heart disease, kidney failure, and other critical conditions. However, the impact of diabetes can be mitigated through timely lifestyle changes and appropriate medical interventions in advance. Therefore, it is important to predict potential diabetes patients and help them prepare in advance. In this case, machine learning algorithms can help healthcare workers with precise predictions. Diabetes incidence is closely linked to various lifestyle and socioeconomic factors, and my project aims to leverage advanced predictive models to assess the risk of diabetes through these factors.

By analyzing data from the CDC's survey [1], my goal is to reflect the probability of individuals developing diabetes based on a comprehensive array of risk factors. Instead of solely predicting diabetic or not, the probability can better inform the potential of developing or severity of diabetes. This project also aims to identify the most significant indicators that contribute to diabetes risk. Through this predictive approach, I intend to enhance preventive measures and ultimately reduce the prevalence and adverse effects of diabetes, focusing on tailored interventions for individuals predisposed to this debilitating disease.

Previously, Xie et al. (2019) explored the prediction of type 2 diabetes risk factors and their classification through the use of various machine learning models, including support vector machine, decision tree, Gaussian Naive Bayes classifiers, and so on [2]. This foundational work discovered known risk factors, including BMI and age, and unveiled new potential ones, including hours of sleep and last doctor's visit. Tasin et al. (2023), building models with a private Bangladeshi dataset, utilized mutual information for feature selection and employed a semi-supervised model to predict insulin levels as an indication of diabetic state [3].

Building upon these insights, my research introduces several novel approaches to further enhance the predictive accuracy and interpretability of risk factor identification for diabetes. First, other than the regular feature selection process, I also used a random forest algorithm to rank the importance of features and then run algorithms against these selected features. Second, I predict the probability of diabetes in patients. Third, I compared socio and physical feature importance, to get a better understanding of how each groups of them contribute to diabetes prediction.

Methods

Below are the steps I followed in the project.

1) Exploratory Data Analysis and Data Preprocessing
2) Feature Extraction and Selection (Regular Approach and Random Forest)
3) Model Selection
4) Hyperparameter Tuning via Random Search
5) Model Implementation and Evaluation

I selected robust classifiers focused on predicting the probability that a sample belongs to one of two binary classes. Here's a breakdown of the chosen models and why they are suitable for this task:

XGBoost Classifier

XGBoost leverages a gradient boosting framework to build an ensemble of weak prediction models, primarily decision trees, in sequence. Each model aims to correct errors from its predecessors, making it highly effective for complex datasets like my diabetes dataset. The model's computational efficiency makes it ideal for predicting probabilities in binary classification. The formula for updating in XGBoost is given by:

$$F_t(x) = F_{t-1}(x) + \eta \cdot \sum_{j=1}^{J} \gamma_j \cdot I(x \in R_j)$$

CatBoost

This gradient boosting algorithm excels with categorical input variables and also performs well with numerical data. CatBoost uses ordered boosting to minimize overfitting and optimally processes categorical variables, making it particularly effective for datasets that combine categorical and numerical features, such as BMI and high blood pressure presence in my case.

DecisionTree

Decision Trees are transparent in their decision-making process, making them intuitive for binary classification tasks. They provide clear probabilities at each leaf node, representing the confidence in predictions based on the proportion of class members within each terminal node, which is crucial for assessing risk levels in medical diagnostics.

LogisticRegression

Regression directly estimates the probability of belonging to a specific class via the logistic function. This model is inherently calibrated to provide confidence scores that are interpretable as the likelihood of an event (e.g., developing diabetes), making it invaluable for precise risk assessment in patient care. The model is computed as:

$$p(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)}}$$

K-Nearest Neighbors

KNN classifies samples based on the majority vote of their nearest neighbors, offering straightforward probability calculations by considering the proportion of neighbors in each class. This method is non-parametric and can be very effective if the dataset has informative, noise-free features. This method not only assigns a class but also quantifies certainty based on local data density, which is essential for robust and reliable medical predictions.

Neural Network (MLPClassifier)

Multilayer Perceptrons (MLPs) are a type of neural network. MLPs can learn complex patterns using layers of nodes in a directed graph, which means they can handle complex interactions between features. They are especially good at capturing interactions in large amounts of data, which can be crucial for identifying subtle patterns in the risk factors of diabetes.

GaussianNB

Gaussian Naive Bayes implements Bayes' theorem under the assumption that all features are independent and normally distributed within each class. This model excels in handling numerical data and can be effective in probabilistic classification by estimating the likelihood of outcomes based on feature evidence. The probabilities it computes reflect the confidence in predictions, aiding in decision-making where data relationships are assumed to be simple.

RandomForestClassifier

Random Forest constructs multiple decision trees during training and determines the output class based on the mode of the classes predicted by individual trees. This method's ensemble approach not only provides a robust mechanism against overfitting but also improves prediction accuracy. Additionally, it offers insights into feature importance and provides probability estimates from the ensemble, reflecting confidence levels in its classifications.

Stacking

In my stacking approach, I use the predictions from base models described above to train a final logistic regression model, aiming to boost overall predictive performance for diabetes risk. The logistic regression meta-learner is chosen for its simplicity and effectiveness in integrating the diverse outputs of the base models. I aim to enhance the accuracy of my predictions and improve the reliability by providing clear probabilities.

Experiments/Results

Data Description

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual survey initiated by the CDC that collects health-related information from over 400,000 Americans each year. It focuses on health behaviors, chronic health conditions, and the usage of preventive services, etc. The specific dataset used for this study comes from the 2015 telephone survey results, including data from 441,455 participants across 330 different features, comprising both direct questions and derived variables. The Diabetes Health Indicators Dataset is derived from this original data and is available on Kaggle. It is segmented into three distinct files with different preprocessing and sampling methods. I chose to focus on the first file, which contains 253,680 responses classified into no diabetes, prediabetes, and diabetes categories and unbalanced among classes. It includes 21 critical indicators for diabetes risk such as high blood pressure, BMI, smoking habits, and heart disease. All features have been cleaned and converted to numerical types, and binary variables have been one-hot encoded for simplified analysis.

Exploratory Data Analysis and Preprocessing

The analysis started off with data preprocessing and exploratory data analysis. I examined the feature distributions, number of samples in each class, and missing values, and the following preprocessing steps were utilized.

1) balancing the dataset and retaining the binary class

The dataset is very imbalanced, with nondiabetes (class 0) having 213703 samples, diabetes (class 2) having 35346 samples and prediabetes (class 1) with 4631 samples. Prediabetic samples were dropped to reduce variability in the feature space and ensure greater distinction between the classes. Class 2 labels were then changed to class 1 for better interpretability.

2) Feature scaling and selection

With a ratio of 2:8, I split the training and testing dataset. Then, I applied standard scaler and feature selection based on Pearson correlation and mutual information.

Pearson correlation was generated based on a training dataset. The feature which has a greater than 0.8 correlation coefficient with another feature and the one with a lower correlation to the target is dropped. Features with lower than 0.05 correlation with the target were also dropped.

In addition to Pearson, mutual information was employed to account for nonlinear relationships between variables and dependencies between variables. Features with <0.01 mutual information with the target were dropped. To safeguard against potential information leakage and ensure that my model remains applicable to real-world scenarios, I inspected mutual information exceeding the threshold of 0.5.

After the preprocessing steps, columns were retained: *'HighBP', 'HighChol', 'BMI', 'HeartDiseaseorAttack', 'PhysActivity', 'GenHlth', 'PhysHlth', 'DiffWalk', 'Age', 'Education', 'Income'*, while variables with greater noise were dropped, including features like smoking habit, diet, alcohol consump, mental health, sex. The same feature selection were performed for xTest.

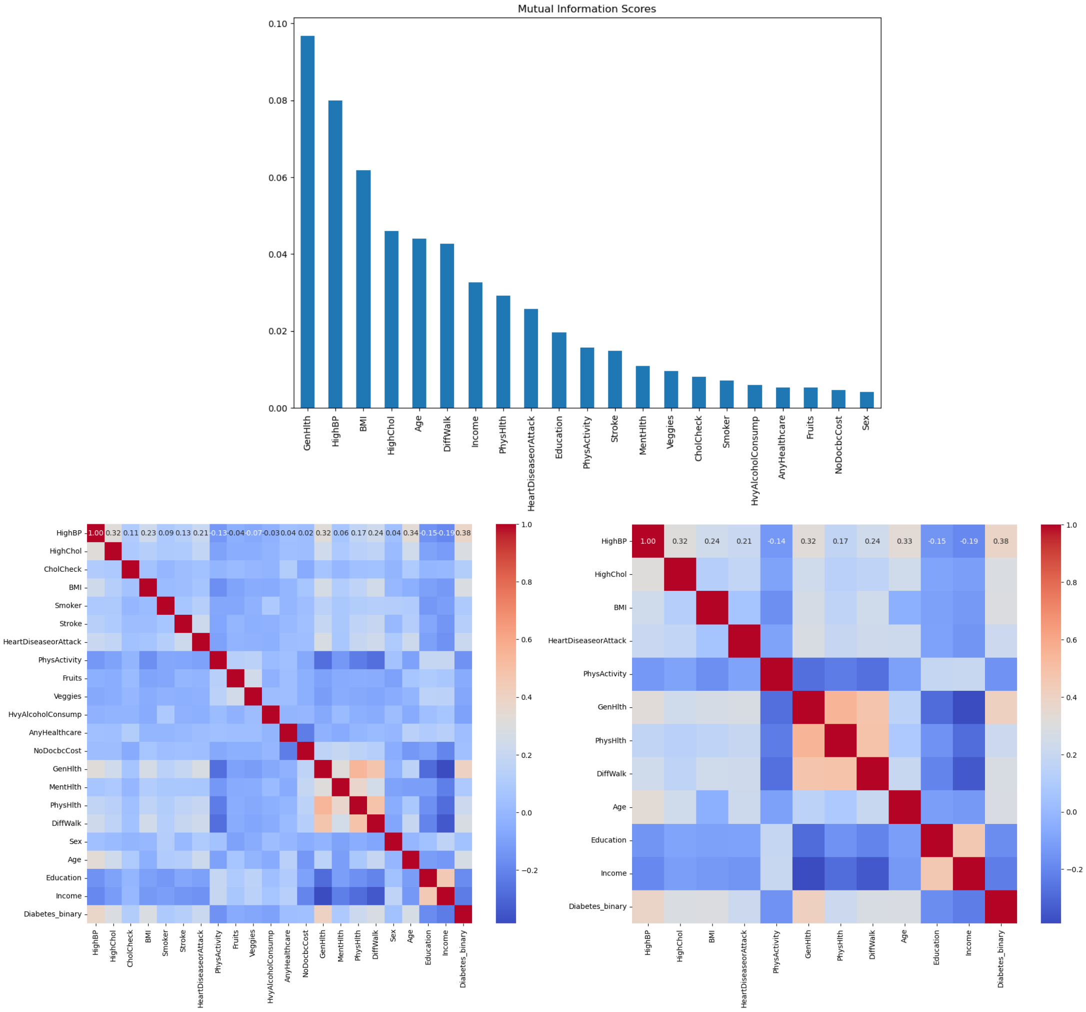

Figure 1. Analysis of Feature Relationships and Feature Selection.

Top plot: Mutual information scores between features and target; Bottom left plot: Pearson correlation of features before feature extraction; Bottom right plot: Feature correlations after feature extraction using pearson and mutual information methods

3) Hyperparameter tuning with random search and Model Choices

I employed Random search cross-validation to tune the hyperparameters of my predictive models. Subsets of the parameter space were defined for each model type, with the number of iterations set to one-third of the total possible combinations. This approach enabled us to efficiently identify optimal configurations across diverse models, including decision trees, logistic regression, K-nearest neighbors, neural networks, Naive Bayes, random forests, XGBoost, and CatBoost. The effectiveness of each configuration was assessed based on the receiver operating characteristic area under the curve (ROC AUC), ensuring that the selected parameters maximized the predictive accuracy of the models.

Table 1. Hyperparameters Tested and Final Selection for Models. The hyperparameters highlighted in red indicate the final selected best-performing options.

Model	Parameter	Values
Decision Tree (DT)	max_depth	None, 10, 20, 30, 40, 50
	min_samples_split	2, 10, 20, 40
	min_samples_leaf	1, 2, 4, 6
	criterion	gini, entropy
Logistic Regression (L1)	solver	liblinear, saga
	C	0.01, 0.1, 1, 10, 100
	penalty	l1
Logistic Regression (L2)	solver	newton-cg, lbfgs, saga
	C	0.1, 1
	penalty	l2
K-Nearest Neighbors (KNN)	n_neighbors	3, 5, 7, 9, 11
	metric	euclidean, manhattan
Neural Network (NN)	hidden_layer_sizes	(50,), (100,), (50, 50), (100, 100)
	activation	relu, tanh, logistic
	solver	adam, sgd
	alpha	0.0001, 0.001, 0.01
	learning_rate	constant, invscaling, adaptive
Naive Bayes (NB)	var_smoothing	1e-8, 1e-7, 1e-6, 1e-5
Random Forest (RF)	n_estimators	100, 200, 300
	max_depth	None, 10, 20, 30
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 4
	bootstrap	True, False
XGBoost (XGB)	max_depth	6, 10, 15, 20
	min_child_weight	5, 10

	gamma	0, 0.1, 0.5, 1, 1.5
	subsample	0.5, 0.7, 0.9
	colsample_bytree	0.5, 0.7, 0.8, 0.9
	learning_rate	0.01, 0.05, 0.1, 0.2
CatBoost (CatBoost)	depth	6, 8, 10
	learning_rate	0.01, 0.05, 0.1
	iterations	100, 500, 1000

4) Empirical results and comparisons

4.1 Model Performance on Feature Extraction with Filter Method

With the chosen models and hyperparameters via random search, I evaluated the models performance on features selected using pearson and mutual information. The assessment focused on the following metrics: Accuracy, Precision, Recall, F1 Score, AUC, and AUPRC.

Table 2. Comparison of Performance of Different Models

Method	Accuracy	Precision	Recall	F1	AUC	AUPRC
KNN	0.718	0.7043	0.7482	0.7256	0.7837	0.7321
NB	0.7031	0.7428	0.6184	0.6749	0.7972	0.7569
DT	0.741	0.7245	0.7749	0.7489	0.8157	0.7816
LR (L2)	0.7501	0.7435	0.7613	0.7523	0.8244	0.7958
LR (L1)	0.7501	0.7433	0.7613	0.7522	0.8244	0.7958
NN	0.7533	0.7401	0.7783	0.7587	0.8306	0.8067
RF	0.7542	0.7352	0.7921	0.7626	0.8289	0.8042
XGB	0.7549	0.7362	0.7919	0.7631	0.8307	0.8076
CatBoost	0.7563	0.7378	0.7926	0.7642	0.8315	0.8075

According to the table above, most of the models have accuracy scores between 0.71 to 0.76. Among these, CatBoost shows the highest Accuracy of 0.7563, F1 Score of 0.7642, and AUC of 0.8315, suggesting strong overall performance.

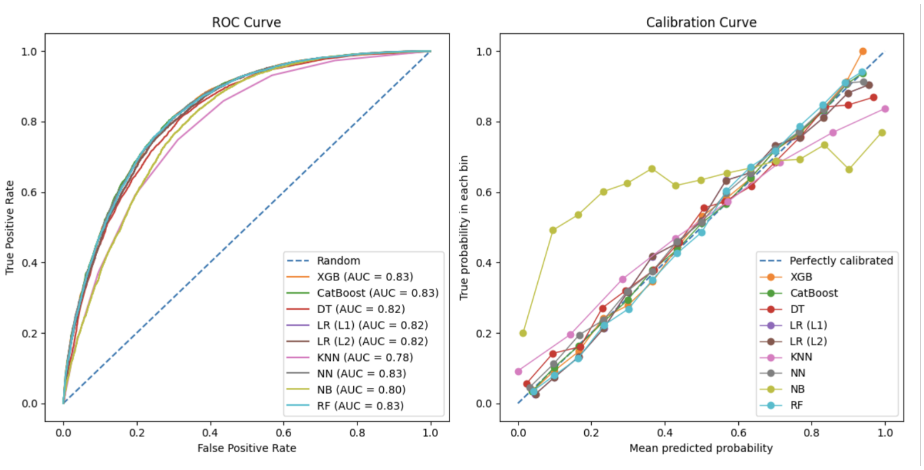

Figure 2: Model Performance Evaluation Curves. Left plot: ROC Curves comparing true positive rate and false positive rate across models; Right plot: Calibration curves showing the accuracy of predicted probabilities for each model.

According to the ROC plot, CatBoost, XGB, and RF have a relatively larger area under the curve compared with other models (Figure 2 Left). Comparing the calibration curves of each model, which show how well the predicted probabilities from each model compared with the actual outcomes, the result indicates that most models demonstrate good probability prediction except for Naive Bayes (Figure 2 Right). Among these, the calibrated curves of CatBoost, NN, and RF are relatively closer to the perfectly calibrated standard line than others, meaning that the probability predictions made by these three models are more reliable.

4.2 Feature Importance

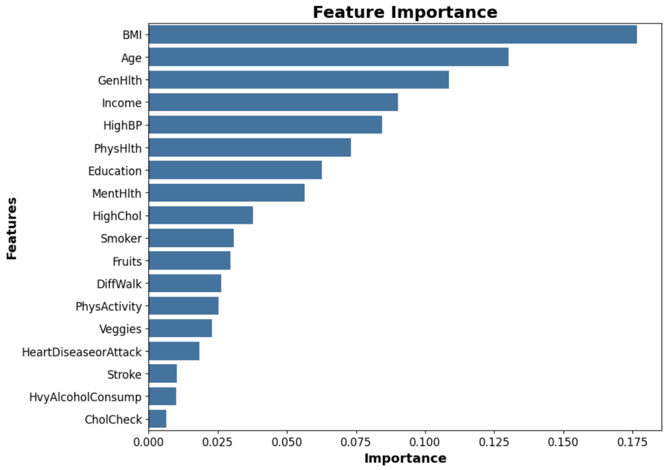

Figure 3: Feature Importance Ranking via Decrease in Impurity within Random Forest

Random forest was used to determine feature importance by measuring how much each feature decreases the impurity in the splits of the decision trees across the forest (Figure 3).

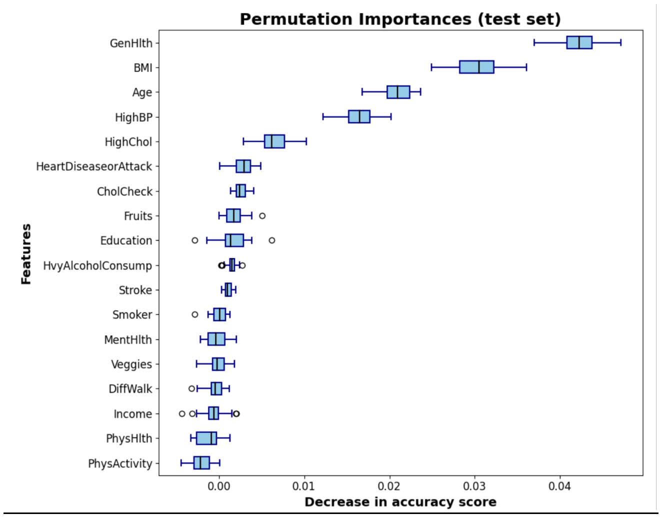

Figure 4: Permutation Importances on Test Set by Measuring Decrease in Model Accuracy

In addition, I determined feature importance by randomly shuffling each feature in the validation set, and measuring the change in the model's accuracy. Each feature is shuffled individually and the process is repeated 30 times to measure the average impact on model accuracy (Figure 4).

Both figures show that features like *General Health, BMI, and Age* have a relatively significant influence on outcomes. However, there are still differences between the two importance assessment methods. In the feature importance assessment, features including High blood pressure and High Cholesterol may have more potential split points inside the decision tree, which is for reducing model impurity during the training process. This may overemphasize the more frequently appeared features, which may not be the most critical features in real-world prediction. In contrast, permutation importance assessment can be used with any model and does not depend on the internal structure of the model, thus providing a model-agnostic way to measure feature importance.

Therefore, by comparing these two methods, I decided to use permutation importance to determine which features are more important. The top 5 important features are *GenHlth, BMI, Age, HighBP , HighChol*, and the 5 least important features are *physical activity, physical health, income, DiffWalk, Veggies*.

4.3 Performance for each model based on the filtered important features

Based on random forest feature importance, 9 features were retained to measure the effect of feature selection with random forest on model performance, with *GenHlth, BMI, Age, HighBP, HighChol, HeartDiseaseorAttack, CholCheck, Fruits, Education.*

The preprocessing steps including scaling were then repeated and hyperparameters were selected again via random search.

Table 3. Comparative Performance of Different Models after Manual Feature Selection

Method	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**	**AUPRC**
KNN	0.7353	0.7219	0.7645	0.7426	0.805	0.7592
NB	0.7449	0.715	0.8137	0.7611	0.8109	0.7733
DT	0.7481	0.7398	0.7648	0.7521	0.8217	0.795
LR (L1)	0.7505	0.7451	0.7607	0.7528	0.8289	0.8012
LR (L2)	0.7505	0.7454	0.7604	0.7528	0.8289	0.8012
NN	0.7556	0.744	0.7786	0.7609	0.8353	0.8152
RF	0.7537	0.74	0.7814	0.7601	0.8331	0.8132
XGB	0.7571	0.7415	0.7888	0.7644	0.8354	0.8153
CatBoost	0.7576	0.7406	0.7920	0.7655	0.8358	0.8164

CatBoost demonstrates the highest overall performance across most metrics (Table 3). CatBoost shows the highest accuracy of 0.7576, and a strong balance between Precision of 0.7406 and Recall of 0.792, contributing to the highest F1 Score of 0.7655 among the models. It also has a relatively high AUC value of 0.8358.

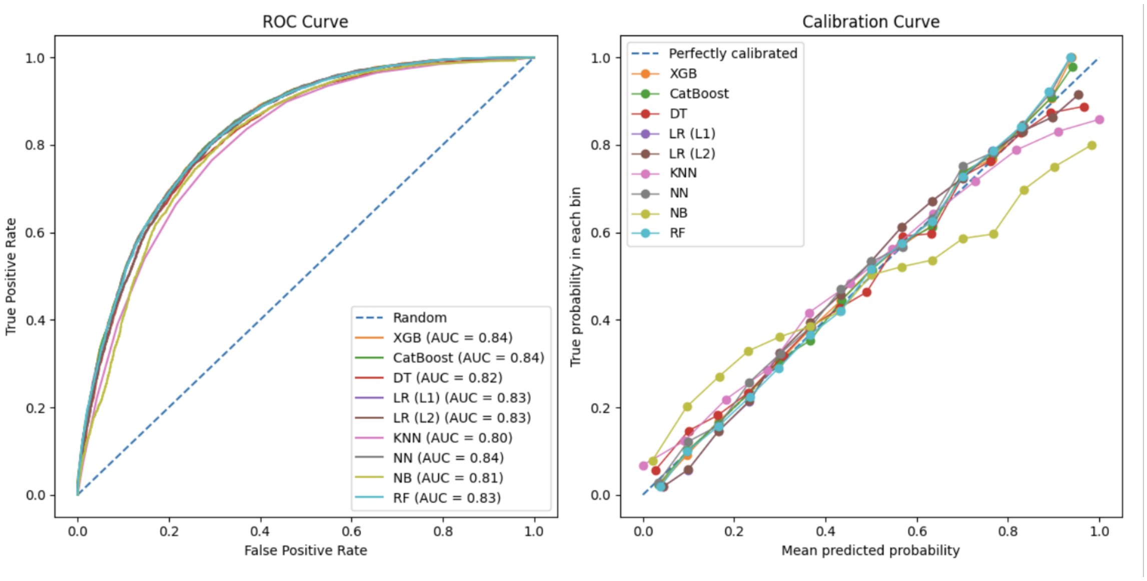

Figure 5: Model Performance Evaluation Curves After Manual Feature Selection. Left plot: ROC Curves comparing true positive rate and false positive rate across models; Right plot: Calibration curves showing the accuracy of predicted probabilities for each model.

XGB, CatBoost, and NN exhibit a larger area under the ROC curve, indicating higher true positive rates and lower false positive rates (Figure 5 Left). The calibration plot demonstrates that most models are well-calibrated, with the mean predicted probability closely aligning with the actual outcomes across different probability thresholds. According to the plot, XGB and CatBoost show relatively better alignment, suggesting that their probability estimates are more reliable (Figure 5 Right). Overall, CatBoost, XGB, and NN stand out based on their overall performance and reliability of probability estimates.

Comparing the model performances before and after manual feature selection (Table 4), the CatBoost model showed slight improvements in most metrics, particularly in precision and AUPRC.

Table 4. Performance Comparison of CatBoost Model With and Without Manual Feature Selection based on Random Forest

Method	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**	**AUPRC**
CatBoost	0.7563	0.7378	0.7926	0.7642	0.8315	0.8075
CatBoost with manual feature selection	0.7576	0.7406	0.7920	0.7655	0.8358	0.8164

4.4 Stacking

My analysis of the stacking model demonstrated some improvements in predictive performance across various metrics compared to individual base models (Table 5), suggesting that the stacking method effectively combines the strengths of individual models, enhancing both the reliability and the accuracy of predictions.

Table 5. Performance Metrics of Stacking Model for Diabetes Prediction

Metric	Accuracy	F1 Score	ROC AUC	AUPRC	Precision	Recall
Stacking result	0.76	0.7666	0.8381	0.8135	0.7385	0.797

Despite these gains, the accuracy did not surpass 0.8, potentially due to the inherent complexity and variability of the diabetes dataset. If the dataset's features do not fully capture all the nuances necessary to distinguish classes effectively or if there is significant noise, these factors could inherently limit the performance ceiling of any modeling approach used.

4.5 Comparison between Social and Clinical Features

To compare the predictive power of social features versus clinical features, the entire dataset was divided after feature selection and preprocessing into two categories: social features, including *Education, Income, Age, and Sex*; and clinical features, including *HighBP, HighChol, BMI, HeartDiseaseorAttack, PhysActivity, GenHlth, PhysHlth, and DiffWalk*. The predictive power of these features was evaluated using CatBoost, Neural Network (NN), and Random Forest (RF) due to their high performance in previous tasks and their computational efficiency.

Table 6. Performances of Models Using Social and Clinical Features for Diabetes Prediction.

Feature	Method	Accuracy	Precision	Recall	F1 Score	ROC AUC
Social	CatBoost	0.6607	0.6318	0.7524	0.6869	0.7214
Social	NN	0.6586	0.6261	0.7691	0.6903	0.7117
Social	RF	0.6572	0.6292	0.7475	0.6833	0.718
Clinical	CatBoost	0.7388	0.7188	0.7752	0.746	0.8181
Clinical	NN	0.7407	0.7236	0.7699	0.7461	0.8196
Clinical	RF	0.7407	0.7198	0.7792	0.7483	0.8178

The results demonstrate a clear distinction in predictive power between the two feature sets across all metrics. Models trained on clinical features consistently outperformed those trained on social factors, with accuracy scores ranging from 0.7388 to 0.7407 for clinical features compared to 0.6572 to 0.6607 for social features (Table 6). Similarly, clinical features also yielded higher precision, recall, F1 scores, and ROC AUC values. However, the accuracy of clinical feature predictions was less than the overall performance when both feature sets were combined, which achieved an accuracy of 0.7576 with catboost. This indicates that while clinical features are highly predictive, social factors also play a role and can enhance the model's performance when integrated.

Discussion

In this comprehensive analysis of various predictive models for diabetes using the Behavioral Risk Factor Surveillance System dataset, I investigated the performance of various machine learning models and assessed the ranking of various diabetic risk factors. The study utilized a range of machine learning techniques, including CatBoost, Neural Networks, and Random Forests, evaluated on a preprocessed dataset emphasizing diabetes risk indicators.

Firstly, my evaluation revealed that CatBoost outperformed other models, likely due to its ability to efficiently handle categorical features and its robust mechanism to combat overfitting (Table 2). Moreover, my feature selection analysis further enhanced model performance (Table 5). By manually excluding less important features on random forest permutation ranking, I were able to enhance model accuracy by diminishing noise and irrelevant data inputs. This step not only streamlined the modeling process but also improved the precision and interpretability of the models.

When comparing social and clinical features, the results clearly show that clinical features possess stronger predictive power for diabetes prediction than social factors. However, the slight improvement in overall model performance when both feature sets are combined suggests that including socio-economic factors offers a more comprehensive approach. Integrating these social determinants of health with clinical data not only refines accuracy but also boosts the reliability of the predictive models. This holistic approach recognizes the complex nature of health and disease, potentially leading to more effective interventions and personalized care strategies in clinical settings.

Reference

[1] A. Teboul, "Diabetes Health Indicators Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data?select=diabetes_012_health_indicators_BRFSS2015.csv.[Accessed: 20-Mar-2024].

[2] Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques," *Prev Chronic Dis*, vol. 16, E130, Sep. 2019, doi: http://dx.doi.org/10.5888/pcd16.190109

[3] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable AI techniques," *Healthc Technol Lett*, vol. 10, no. 1-2, pp. 1–10, Feb.-Apr. 2023, doi: https://doi.org/10.1049/htl2.12039