Using Lifestyle Markers and Machine Learning to Detect Diabetes

Ayush Srivastava[1]

Received November 28, 2025
Accepted April 5, 2026
Electronic access April 30, 2026

Diabetes is a prevalent chronic disease throughout the world and affects more than 100 million people just in the United States itself. Type 2 diabetes (T2D) can be prevented and controlled by managing lifestyle markers, such as diet and amount of exercise. In this project, a mathematical analysis of different lifestyle markers is presented along with a study using machine learning (ML) classifiers to detect prediabetes/diabetes cases. For this project, a dataset from the Centers for Disease Control and Prevention (CDC) containing 21 lifestyle markers (age, BMI, physical health, high blood pressure, etc.) and diagnosis of prediabetes/diabetes (binary label) from more than 250,000 adults was used. After conducting a comprehensive evaluation of several ML models (using 10-fold stratified cross-validation), it has been found that prediabetes/diabetes cases can be classified or detected with a high recall (>80%) using a Linear SVC model. In addition, this project presents an empirical relationship between lifestyle markers and their importance in the determination of diabetes by the model. A web application Diabetes Risk Analyzer has also been developed using the model trained in this project to detect prediabetes/diabetes based on lifestyle markers for use as a quick, robust, and widely accessible tool.

Keywords: Diabetes, machine learning, supervised learning, classification

Introduction

Diabetes is a chronic disease that affects millions of people in the United States and the rest of the world. According to WHO, this chronic disease occurs when the pancreas does not produce enough insulin or when the body cannot use the insulin it produces effectively. By 2022, nearly 830 million people in the world were living with diabetes. In 2022, 14% of adults 18 years and older lived with diabetes. In 2021, diabetes was the direct cause of 1.6 million deaths and 47% of all deaths due to diabetes occurred before the age of 70 years.

Diabetes can be divided into two categories, Type 1 diabetes (T1D) and Type 2 diabetes (T2D). More than 95% of people with diabetes have Type 2 diabetes (T2D). Type 2 diabetes was previously called non-insulin dependent or adult-onset. It prevents the body from using insulin properly, which can lead to high levels of blood sugar if not treated, and, over time, Type 2 diabetes can cause serious damage to the body, especially to nerves and blood vessels. Type 2 diabetes is often preventable. Factors that contribute to the development of type 2 diabetes include being overweight, not getting enough exercise, and genetics. Lifestyle changes are the best way to prevent or delay the onset of type 2 diabetes. Early diagnosis is important in preventing the worst effects of type 2 diabetes. In this project, the main contributions are the following:

- Presenting the empirical relationship between different lifestyle markers, including their correlations and distributions between prediabetic/diabetic and nondiabetic populations.
- Conducting a comprehensive evaluation of machine learning classifiers to detect prediabetes/diabetes using these lifestyle markers. This also includes identifying the top most influential markers for the model's prediction.
- Presenting a web application which uses the trained model in this project to detect prediabetes/diabetes based on the user's current lifestyle markers.

While genetics and blood results can provide more comprehensive insights into the risk of diabetes for any individual, the scope of this project shows the screening potential using only lifestyle markers. Lifestyle markers are more widely available and easily accessible without incurring any medical tests or costs. An ML model is trained on the dataset used for this project in order to accurately classify prediabetes/diabetes cases based on lifestyle markers.

Related Work

In this paper[1], a machine learning model was trained to predict blood glucose levels using real-time data from monitored patients. Although this research and tool are useful for Type 1 diabetics, this applies only to patients already diagnosed with diabetes and insulin dependent.

[1] *North Allegheny Senior High School, USA*

Similarly, in this paper[2], machine learning-based glucose prediction with the use of continuous glucose and physical activity monitoring data: The Maastricht study from the National Library of Medicine and machine learning is used to predict glucose levels accurately and safely based on continuous glucose monitoring (CGM).

In this paper[3], the authors have shown that physical activity is beneficial and can be used both for the management of Type 1 diabetes and for the prevention and management of Type 2 diabetes. Here too, it is emphasized that the presence of physical activity correlates with better physical and mental health. In addition, general health is one of the key attributes of the ML model in diagnosing diabetes.

In this paper[4], machine learning algorithms have also been used to diagnose diabetes with high precision. Here, the authors use a combination of lifestyle markers and medical information (insulin, glucose level, etc.) as input features for the ML models. However, medical measurements are not widely accessible to the general population and the determination of diabetes risk using such methods becomes limited.

In this paper[5], the authors have performed a similar analysis using the Receiver Operating Characteristic (ROC) on multiple data sets on diabetes. Their analysis spans only four machine learning classifiers, though, whereas this project covers the evaluation of eight machine learning classifiers.

In this paper[6], six ML classifiers are evaluated using a dataset with only 768 data points consisting of eight markers (including BMI, age and clinical markers) and testset accuracies.

In this paper[7], three ML classifiers (logistic regression, naive bayes, adaboost) are evaluated on a dataset with lifestyle and clinical markers.

A key difference between this research paper and most related work on diabetes is that very few lifestyle markers have been explored for causative influence and correlation compared to this research, where 21 lifestyle markers from more than 250,000 adults were used.

Methods

Dataset

For this project, the CDC Diabetes Dataset from the UCI Machine Learning Repository was used. This dataset contains 253,680 data points, with 21 lifestyle markers (14 binary and 7 integer markers) pertaining to both men and women. 35,346 (=13.93%) cases are labeled as Yes (prediabetes/diabetes diagnosed) whereas the remaining 86.1% cases are labeled as No (prediabetes/diabetes not diagnosed). Table 1 lists the different lifestyle markers to be used as input features for our model.

Age, Education, and Income are multi-category variables. Since their respective integer values correspond to intrinsic ordered ranges as shown in Table 1 above, we use these numeric values directly and do not use one hot encoding. This keeps our feature dimensionality low and avoids addition of sparse features.

These lifestyle markers play a vital role in the classification of prediabetes/diabetes cases. Using these lifestyle markers and the data from the dataset, the machine learning model chosen will be able to detect prediabetic/diabetic cases.

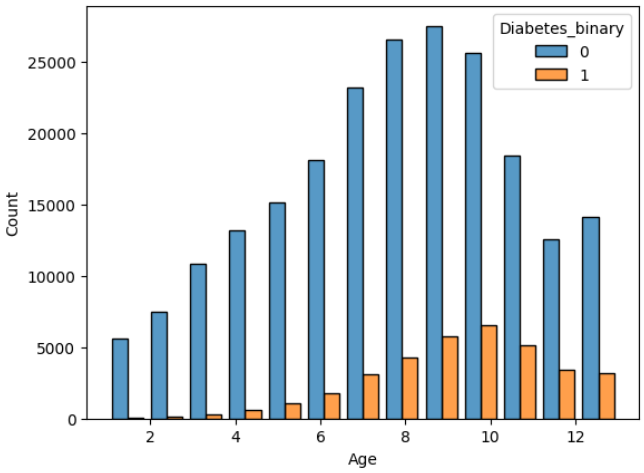

Fig. 1 Age distribution (grouped age bands)

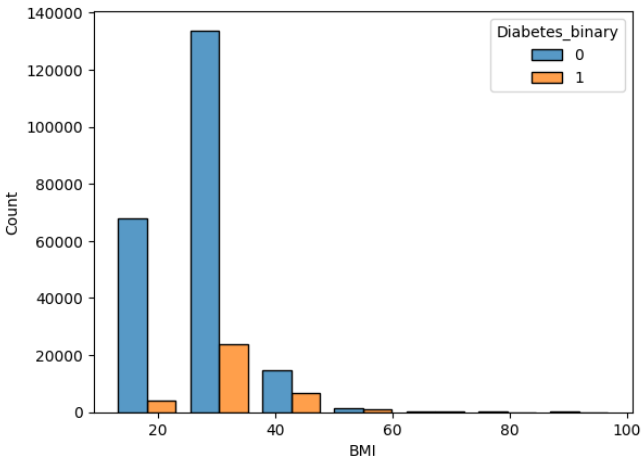

Fig. 2 BMI distribution

Figure 1 shows the distributions of the different age groups for the prediabetic/diabetic and non-diabetic population. For group 10 (ages[8] 65-70), the count for diabetes peaks. Similarly, Figure 2 shows the distribution of BMI. Some lifestyle markers cannot be changed, such as gender and age, while others can be altered, such as intake of fruits, vegetables, and

Table 1 Input features in dataset

Feature	Type	Range	Description
Blood Pressure	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Cholesterol	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Cholesterol Check	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Smoker	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Stroke	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Heart Disease or Attack	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Physical Activity	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Fruits	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Veggies	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Heavy Alcohol Consumption	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Health Care	Binary	0 - 1	0 indicates absence, 1 indicates presence.
No Doctor Due to Cost	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Walking Difficulty	Binary	0 - 1	0 indicates absence, 1 indicates presence.
Gender	Binary	0 - 1	0 = female 1 = male
BMI (kg/m^2)	Integer	12 - 98	Body Mass Index in kg/m2
General Health	Integer	1 - 5	1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
Mental Health	Integer	1 - 30	For how many days during the past 30 days was your mental health not good?
Physical Health	Integer	1 - 30	For how many days during the past 30 days was your physical health not good?
Age	Integer	1-13	13-level age category 1 (18-24 years) 2 (25-29 years) 3 (30-34 years) 4 (35-39 years) 5 (40-44 years) 6 (45-49 years) 7 (50-54 years) 8 (55-59 years) 9 (60-64 years) 10 (65-69 years) 11 (70-74 years) 12 (75-79 years) 13 (80+ years)
Education	Integer	1-6	Education level 1 (No school/only Kindergarten) 2 (Grades 1-8) 3 (Grades 9-11) 4 (High School Graduate) 5 (1-3 years of college) 6 (College Graduate)
Income	Integer	1-8	Income scale 1 (< $10,000) 2 ($10,000 - $15,000) 3 ($15,000-$20,000) 4 ($20,000-$25,000) 5 ($25,000-$35,000) 6 ($35,000-$50,000) 7 ($50,000-$75,000) 8 ($75,000+)

increase in physical activity. A Pearson correlation matrix, shown in Figure 3, was computed to explain and analyze the correlations between the different lifestyle markers.

Cells in lighter colors show positive correlation, whereas cells in darker colors show negative correlation. This correlation matrix helps explain how lifestyle markers are related to each other. As shown in Figure 3, physical activity is negatively correlated with health markers (health markers in this dataset denote the number of days the patient's health was not good). Hence, the presence of physical activity will influence good physical and mental health. Higher education and income groups also correlate with absence of high blood pressure, cholesterol, and low BMI. These are only rough measures and do not indicate full causal relationships. Table 2 depicts the mean and standard deviation for each feature, as shown below.

Table 2 summarizes the central tendency and variability of each lifestyle marker in the dataset by reporting the mean and standard deviation.

Procedure

The original dataset was divided randomly into three mutually exclusive subsets[9]: a training set (80%), a validation set (10%), and a held-out test set (10%). The training set was used for model selection using cross-validation, the validation set was used for model operating point selection (threshold tuning) based on Receiver Operating Characteristic (ROC) analysis, and the testing set was used finally only once for reporting performance metrics. All these dataset splits were performed using stratified sampling[10] to preserve the proportion of positive (prediabetes/diabetes) and negative cases across subsets. Feature standardization[11,12] (zero mean, unit variance) was fitted on the training data and applied to the validation and test sets to avoid data leakage. We use default hyperparameters from scikit-learn library for all model cross-validations or training.

In this project, there were four main phases in the experimental project as shown in Figure 4.

Phase 1: 10-fold stratified cross validation

In this phase, multiple ML classifiers were evaluated using 10-fold stratified cross-validation[13–15] (StratifiedKFold) with AUC (Area Under the ROC Curve) as the scoring metric on the training dataset. An AUC[16,17] close to 1 means the model effectively separates prediabetic/diabetic and non-diabetic cases well, while an AUC near 0.5 implies classification is almost random. Stratified cross validation was used so that each fold has a similar proportion of prediabetic/diabetic and non-diabetic cases. Since the dataset is class-imbalanced (13.9% positive), model performance was evaluated using AUC. Scikit-learn, an API library for ML classifiers, was used for this purpose. Table 3 lists the eight different classifiers evaluated along with their AUC scores in ranked order. The Linear SVC model ranked highest with an AUC = 82.3% and was selected as the best model for this dataset and classification task. While Linear SVC, Logistic Regression[18], and AdaBoost[19] models performed similarly, Linear SVC (top-ranked) was selected for simplicity and interpretability. For all these models, we use default hyperparameters from scikit-learn.

Phase 2: Training

The Linear SVC model is then trained on the entire training dataset. For obtaining class probabilities, CalibratedClassifierCV was used with default parameters. The CalibratedClassifierCV uses cross-validation internally to both estimate the parameters of a classifier and subsequently calibrate a classifier. This is required since the LinearSVC model predicts only class labels and does not return class probabilities.

Phase 3: Validation

In this phase, the trained Linear SVC model from Phase 2 makes predictions on the validation dataset. A Receiver Operating Characteristic (ROC) curve, depicting the relationship between the True Positive Rate (TPR) vs. the False Positive Rate (FPR), for the entire range of probability scores, is plotted for the validation dataset predictions. For this dataset, the Yes points (prediabetes/diabetes diagnosed) will be the positives and the No points (not diagnosed) will be the negatives. The True Positive Rate or Recall is defined as the ratio of the number of true positives (data points predicted correctly) by the model to the total number of positive data points. The False Positive Rate is defined as the ratio of the number of false positives (negative data points the model classified as positive), to the total number of negative data points. From the ROC plot, the optimal threshold is chosen as the operating point[20] of the model. This threshold represents a probability value, for classifying data points as positive with probability score above the threshold and Negative for data points scoring below the threshold.

A conservative threshold of 0.1175 was chosen for the Linear SVC model, as shown by the red point in the ROC plot in Figure 5. A threshold of 0.1175 for the model corresponds to a true positive rate of 80.31% and a false positive rate of 31.71% on the validation dataset.

Phase 4: Testing

In this phase, the trained Linear SVC model from Phase 2 makes predictions on the held-out testing dataset. The threshold selected from the validation set in Phase 3 is then applied without modification to the held-out test set to generate the final confusion matrix and performance metrics. The test set

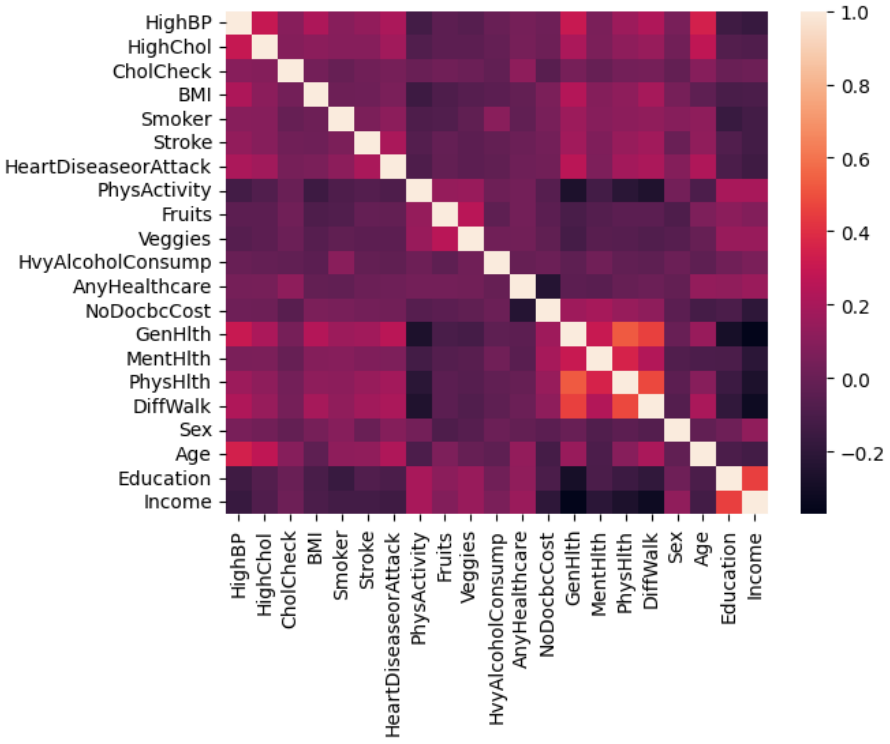

Fig. 3 Pearson Correlation Matrix between the 21 lifestyle markers.

contains 25,368 individuals (3,534 with prediabetes/diabetes and 21,834 without).

Results

Linear SVC had the highest AUC, making it the optimal model for this project as shown in Table 3 above. A confusion matrix was created to depict the results based on the Linear SVC predictions on the testing dataset, as shown in Table 4 below. The test set contains 25,368 individuals (3,534 with prediabetes/diabetes and 21,834 without).

Confusion matrices are utilized to depict the results of a machine learning classifier by comparing its predicted labels to the actual labels. The diagonal cells are the number of data points the model was able to correctly identify as positive or negative. The off-diagonal cells are the number of data points the model was not able to correctly predict. Overall, the model successfully achieved an accuracy of $\sim 70\%$. In addition, the Linear SVC model was also able to achieve a 80.62% recall, which is more important, since the cost of classifying a non diabetic patient as prediabetic/diabetic is much less than the cost of classifying a prediabetic/diabetic patient as non diabetic.

While prioritizing recall is most important for this project, the potential costs of false positives must be acknowledged as well. Potential costs include unnecessary anxiety for users, additional medical consultations, and possible strain on healthcare resources. Therefore, the model output should be interpreted strictly as a preliminary screening signal rather than a diagnostic result. In practice, any high-risk classification would need confirmation through standard clinical tests. The threshold value implemented reflects a deliberate trade-off favoring sensitivity over specificity to reduce missed cases, but different deployment contexts could justify alternative thresholds depending on acceptable false positive rates.

We also check how well calibrated are the model prediction probabilities as compared to the true outputs in the held-out testing dataset. The calibration curve[21] is as shown below in Fig. 6.

From Fig. 6, we find that the model scores are well-calibrated upto probability = 0.5 after which the model is over-predicting. For our classification context, we are using a conservative threshold of 0.1175 only to flag prediabetic/diabetic cases. We are not using these model prediction probabilities as risk scores.

Performance is also evaluated on different subsets of the testing dataset:

1. Young adults: For this subset, the age groups 1 and 2 corresponding to populations with ages 18-24 and 25-29

Table 2 Feature Means (with standard deviations)

Feature	Training dataset (n=202944)	Validation dataset (n=25368)	Testing dataset (n=25368)
High Blood Pressure	0.429089 (0.494947)	0.427507 (0.494727)	0.429793 (0.495056)
High Cholesterol	0.424156 (0.494215)	0.423802 (0.494169)	0.424156 (0.494224)
Cholesterol Check	0.962561 (0.189835)	0.96468 (0.184591)	0.961526 (0.192341)
BMI	28.384303 (6.614076)	28.35872 (6.620453)	28.390492 (6.553833)
Smoker	0.442541 (0.496689)	0.441343 (0.496557)	0.450016 (0.497505)
Stroke	0.04044 (0.196989)	0.042179 (0.201002)	0.040011 (0.195989)
Heart Disease or Attack	0.094218 (0.292133)	0.096026 (0.294633)	0.092085 (0.289151)
Physical Activity	0.756036 (0.429472)	0.759618 (0.427324)	0.757529 (0.428586)
Fruits	0.634116 (0.481678)	0.635919 (0.481181)	0.633712 (0.481799)
Vegetables	0.811229 (0.391328)	0.809997 (0.392311)	0.814372 (0.388813)
Heavy Alcohol Consumption	0.056863 (0.231581)	0.053532 (0.225096)	0.053532 (0.225096)
Any Health Care	0.950986 (0.215897)	0.950173 (0.217591)	0.95246 (0.212795)
No Doctor Because of Cost	0.084477 (0.278102)	0.084989 (0.278871)	0.080968 (0.272792)
General Health	2.511038 (1.06769)	2.511944 (1.072901)	2.513679 (1.070379)
Mental Health	3.186037 (7.409267)	3.151766 (7.392514)	3.207663 (7.461807)
Physical Health	4.246462 (8.721475)	4.216138 (8.686586)	4.232971 (8.721354)
Difficulty Walking	0.168253 (0.374092)	0.165721 (0.371837)	0.17049 (0.37607)
Gender	0.439949 (0.496382)	0.444458 (0.496915)	0.439372 (0.49632)
Age	8.028776 (3.053709)	8.047304 (3.058033)	8.043677 (3.05454)
Education	5.051921 (0.984684)	5.04147 (0.993025)	5.047501 (0.9872)
Income	6.054931 (2.069851)	6.038789 (2.085603)	6.060509 (2.067018)

Table 3 ML classifiers evaluated with their cross-validation AUC scores.

ML Classifier Model	AUC
Linear SVC	82.30%
Logistic Regression	82.26%
AdaBoost	82.20%
Random Forest	79.84%
Naive Bayes	78.41%
QDA	78.06%
Nearest Neighbors	76.43%
Decision Tree	59.75%

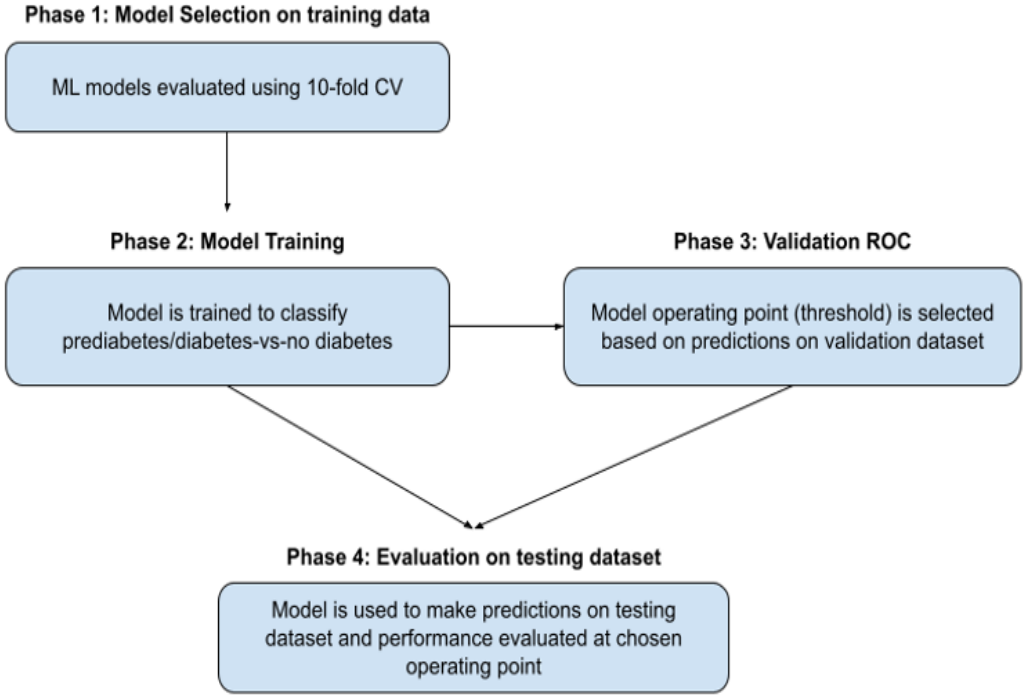

Fig. 4 Experimental phases.

Table 4 Confusion matrix computed on the testing dataset predictions

		Predicted label		**Accuracy**	**70.08%**
		No	**Yes**	**Total**	**Recall**
True label	**No**	14928	6906	21834	68.37%
	Yes	685	2849	3534	80.62%
	Total	15613	9755	25368	
	Precision	95.61%	29.21%		

respectively in the testing dataset are evaluated. There are 1,286 cases here with only 2% labeled as prediabetics/diabetics. The Linear SVC model still achieves a recall of ∼42% for the positive class and an overall accuracy of 95%.

2. Men: For this subset, the male population in the testing dataset is evaluated. There are 11,208 cases here with 15.27% of prediabetics/diabetics. Recall for the positive class is ∼81%, which is very similar to that of the entire testing dataset (=80.62%). Overall accuracy is 67.58%, which is slightly less than overall accuracy of entire testing dataset (=70.08%).

3. Women: For this subset, the female population in the testing dataset is evaluated. There are 14,160 cases here with 12.87% of prediabetics/diabetics. Recall for the positive class is ∼81%, which is very similar to the male population recall and that of the entire testing dataset (=80.62%). Overall accuracy is 72.06%, which is more than overall accuracy of male population (=67.58%) and slightly more than that of the entire testing dataset (= 70.08%).

SHAP Analysis

A SHapley Additive exPlanations (SHAP) plot was created to explain which lifestyle markers most strongly influenced the model prediction output, as shown in Figure 7. SHAP

Table 5 Confusion matrix computed on the young adult predictions in testing dataset

		Predicted label		Accuracy	95.10%
		No	**Yes**	**Total**	**Recall**
True label	**No**	1213	49	1262	96.12%
	Yes	14	10	24	41.67%
	Total	1227	59	1286	
	Precision	98.86%	16.95%		

Table 6 Confusion matrix computed on the male predictions in testing dataset

		Predicted label		Accuracy	67.58%
		No	**Yes**	**Total**	**Recall**
True label	**No**	6193	3303	9496	65.22%
	Yes	331	1381	1712	80.67%
	Total	6524	4684	11208	
	Precision	94.93%	29.48%		

Table 7 Confusion matrix computed on the female predictions in testing dataset

		Predicted label		Accuracy	72.06%
		No	**Yes**	**Total**	**Recall**
True label	**No**	8735	3603	12338	70.80%
	Yes	354	1468	1822	80.57%
	Total	9089	5071	14160	
	Precision	96.11%	28.95%		

Table 8 Sample rows from the validation dataset

HighBP	BMI	GenHlth	Gender	Age	Label
1	28	5	1	9	1
1	34	2	0	10	1
1	50	4	0	8	1

Table 9 SHAP feature contributions corresponding to Table 8 rows

HighBP	BMI	GenHlth	Gender	Age	Max factor
0.0808	-0.0351	0.3077	0.0341	0.0278	GenHlth
0.0808	0.0724	-0.0989	-0.0257	0.0545	HighBP
0.0808	0.3592	0.1721	-0.0257	0.0011	BMI

is a library that can attribute the overall contribution of each lifestyle marker to the final prediction. General health, BMI, and High BP are the strongest contributors for the final decision from the Linear SVC model. The purpose of this SHAP[22–24] plot is to quantitatively measure the contributions each lifestyle marker had on the prediction of the model, essentially providing a ranked list of the most influential markers. For General health, the range of values correspond to 1 = excellent, 2 = very good, 3 = good, 4 = fair and 5 = poor. According to the SHAP plot, higher BMI[25], poorer general health (higher values), and HighBP[26] (=1) increased the predicted model score, whereas lower BMI, HighBP (=0), and

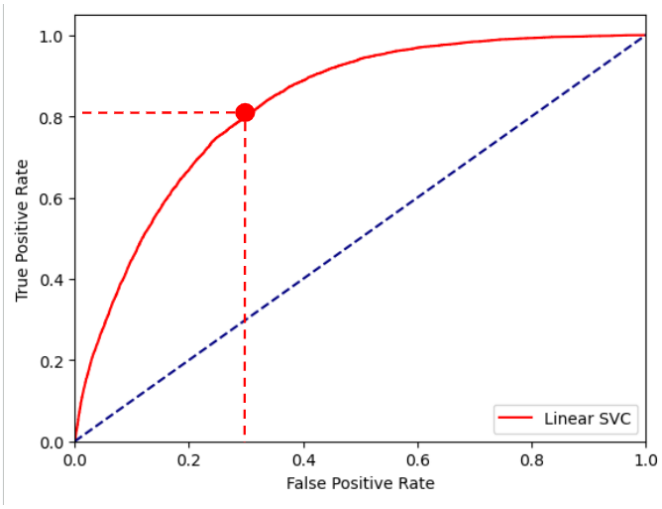

Fig. 5 Receiver Operating Characteristic (ROC) curve plotted for the validation dataset predictions

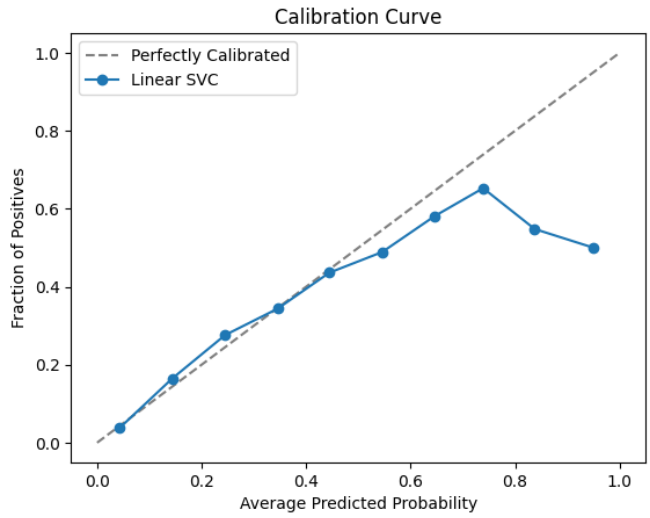

Fig. 6 Calibration curve for model predictions on testing dataset

better general health (lower values) decreased it.

Here are some data examples from the validation dataset in the following Table 8 for some current prediabetic/diabetic cases. Table 9 shows the corresponding SHAP feature contributions.

All of the code was developed in Python on Google's Colab notebook and is available on GitHub.

Web Application

The web application created utilizes a Flask server hosting the trained Linear SVC model in the back-end. Users enter their

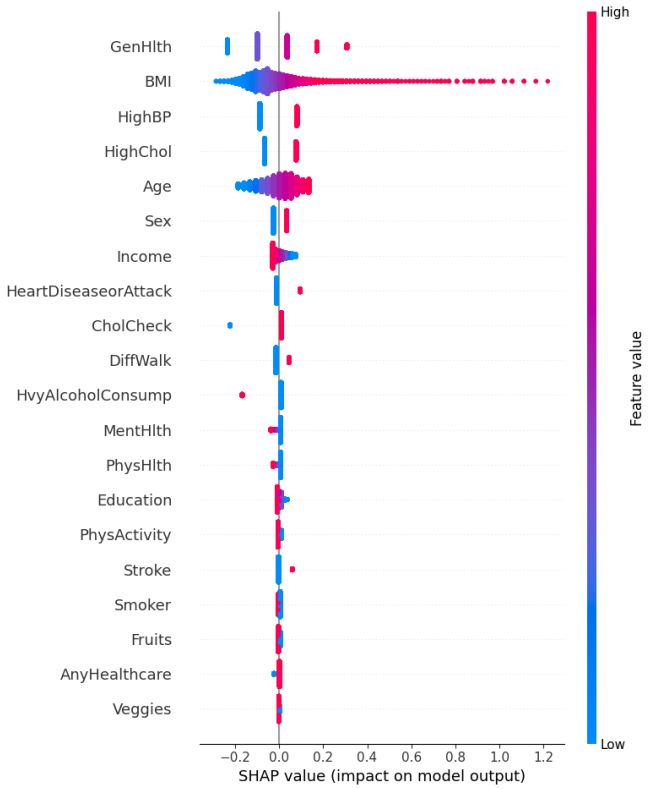

Fig. 7 SHAP plot showing the ranked list of feature importances from the model output

lifestyle markers, which are then passed to the model for predictions. Based on the lifestyle markers of the user, the model outputs "HIGH" if the prediction is above the chosen threshold of 0.1175 and "LOW" if the prediction is below the chosen threshold, as shown in Figure 8. While this web application aims to provide users with their classification of prediabetes/diabetes, it does not aim to provide medical advice or replace professional medical diagnosis. This is only for preliminary screening use only. This web application is deployed on AWS and it does not log or store any users' data (form entries) or IP addresses. The user inputs are only passed to the model for prediction purposes.

Discussion

In conclusion, the Linear SVC model was able to achieve an 80.62% prediabetes/diabetes classification recall on the held-out testing dataset, comprising 10% of the total dataset. While this was a single-dataset and multiple-split evaluation conducted, external validation is future work. A trivial baseline classifying all individuals as non-diabetic would achieve ∼86% accuracy but 0% recall for the positive class, demon-

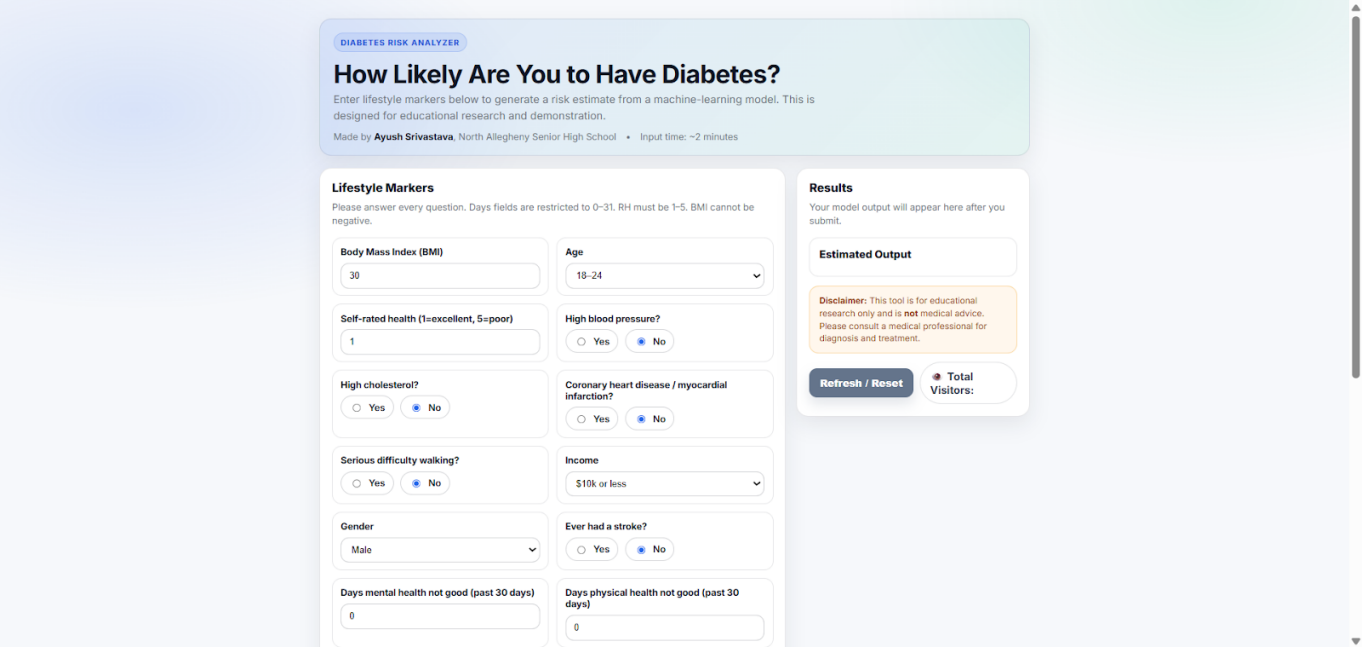

Fig. 8 Web application created using the Linear SVC model and Flask server.

strating the need for a recall-oriented model.

A web application was created where the trained Linear SVC model can make a prediabetes/diabetes classification using the user's input lifestyle markers. This project met all the objectives of an accurate model, a mathematical analysis, and the web application.

The first objective, which was to train a machine learning classifier to accurately differentiate between positive cases and negative cases, was achieved through the selection of the best classifier, the training procedure of the model, and the optimal threshold chosen.

The second objective, which was to analyze and explain which lifestyle markers were the most correlated with prediabetes/diabetes, was achieved by computing the SHAP plot and the correlation matrix. Both of these diagrams assist in the analysis and explanation of the model's predictions.

Lastly, the web application can be utilized to make quick, robust, and accessible prediabetes/diabetes classifications. Our model is not replacing any official test; it is an extra early warning using lifestyle only. In real clinics, a high-risk flag would still need a blood test. Also, this model is mostly suited to populations similar to the CDC dataset, and its performance on other countries is still unknown or a future work. Some future goals with this project are to evaluate this model on more global diabetes datasets, such as the PIMA Indians Diabetes Database.

References

1 A. Devi, P. Sasireka, K. Kovardhani, G. Premalatha, T. Sivasakthi and R. Subash, 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, pp. 1519–1522,.

2 W. Doorn, Y. Foreman, N. Schaper, H. H. Savelberg, A. Koster, C. J. Kallen, A. Wesselius, M. T. Schram, R. M. Henry and P. C. Dagnelie, *PloS one*, **16**, year.

3 C. Hayes and A. Kriska, *Journal of the American Dietetic Association*, **108**, year.

4 A. Mujumdar and V. Vaidehi, *2nd International Conference on Recent Trends in Advanced Computing ICRTAC -DISRUP - TIV INNOVATION*, **165**, 292–299.

5 H. Lai, H. Huang, K. Keshavjee, A. Guergachi and X. Gao, *BMC endocrine disorders*, **19**, 1–9.

6 N. Abdulhadi and A. Al-Mousa, 2021 international conference on information technology (ICIT.

7 M. M. Mijwil and M. Aljanabi, *Baghdad Science Journal*, **21**, 24.

8 N. Nanayakkara, *Diabetologia*, **64**, 275–287.

9 T. Burzykowski, *American journal of orthodontics and dentofacial orthopedics*, **164**, 295–297.

10 M. Khushi, *Ieee Access*, **9**, 109960–109975.

11 N. M. N. Mathivanan, *Malaysian Journal of Computing*, **10**, 2159–2175.

12 K. M. Sujon, *IEEE access*, **12**, 135300–135314.

13 L. A. Yates, *Ecological monographs*, **93**, 1557.

14 S. Bates, T. Hastie and R. Tibshirani, *Journal of the American Statistical Association*, **119**, 1434–1445,.

15 L.-b. Sweet, *Artificial Intelligence for the Earth Systems*, **2**, 230026.

16 T. Polo and H. Miot, *J Vasc Bras*, **19**, 20200186.

17 T. Yang and Y. Ying, *ACM computing surveys*, **55**, 1–37.

18 E. C. Zabor, *International Journal of Radiation Oncology* Biology* Physics*, **112**, 271–277.

19 A. Shahraki, M. Abbasi and Haugen, *Engineering Applications of Artifi-*

cial Intelligence, **94**, 103770.
20 M. Hassanzad and K. Hajian-Tilaki, *BMC medical research methodology*, **24**, 84.
21 P. C. Austin, F. E. Harrell, Jr and D. Klaveren, *Statistics in Medicine*, **39**, 2714–2742.
22 E. Mosca, Proceedings of the 29th international conference on computational linguistics.
23 A. V. Ponce-Bobadilla, *Clinical and translational science*, **17**, 70056.
24 H. Wang, *Journal of Big Data*, **11**, 44.
25 P. Chandrasekaran and R. Weiskirchen, *International journal of molecular sciences*, **25**, 1882.
26 M. W. Naseri, H. A. Esmat and M. D. Bahee, *Annals of Medicine and Surgery*, **78**, year.