Prediction of Diabetes with Machine Learning Classification

BIS 634: Computational Methods for Informatics
Final Project Report
Xinyue Qiu(xq44)

Introduction

Studies have shown that patients with diabetes worldwide have more than doubled during the past 20 years, with a significant increase in type 2 diabetes patients among children and adolescents. [1] Accompanying the global emergence of diabetes, the identification of key risk factors and susceptible populations has been given more attention. In this project, I will try to use a dataset derived from CDC's Behavioral Risk Factor Surveillance System in 2015(BRFSS2015) to identify key risk factors contributing to diabetes and find the optimal classification model to predict diabetes. [2]

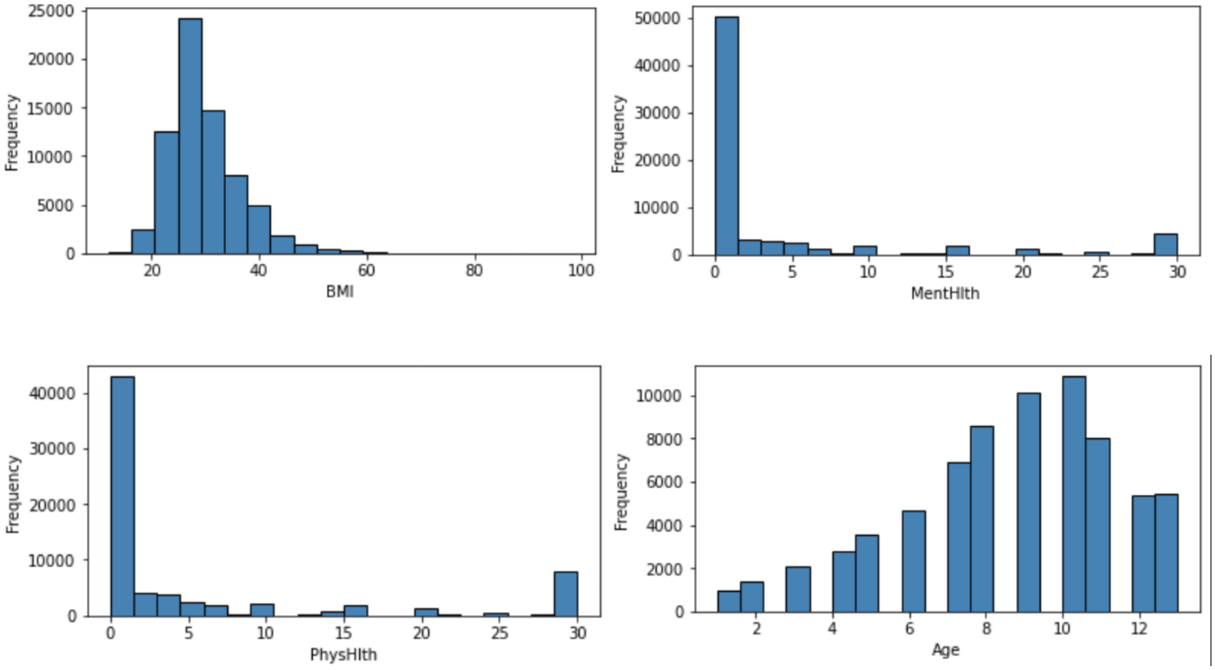
This 'Diabetes, Stroke and Hypertension Prediction' dataset is derived from BRFSS, which is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. [3] The dataset is downloaded from kaggle.com, and it has a license of CC0: public domain. All features included in the dataset are well annotated in metadata, thus the dataset is rather fair in its acquisition. This dataset contains 70,692 survey responses to the CDC's BRFSS2015 and enlisted a small subset of features recorded. In total 17 features including diabetes, are included in this dataset. A detailed description of these variables is included in the Appendix (Table 1). The amount of information provided enables us to conduct a comprehensive analysis of diabetes and study the relationship between diabetes and other health-related information. In the long term, this project aims to correctly identify the patients at risk of diabetes on a large scale, as well as help people identify high-risk health-related factors to further prevent diabetes.

Data Preprocessing

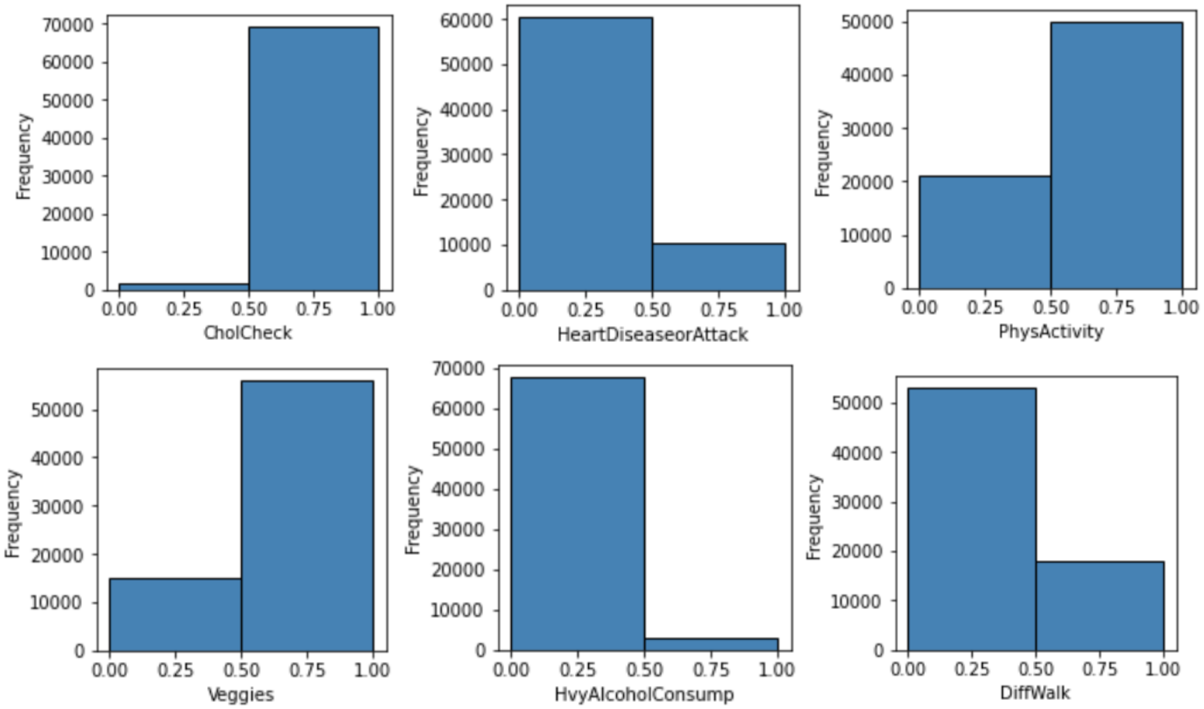
Distribution Check

To analyze the dataset, each feature included in the analysis was visualized and checked using summary statistics and histograms. From the histograms, we can see that the quantitative features 'BMI', 'MenHlth', 'PhysHlth' are right-skewed, while the feature 'age' is slightly left-skewed. The histograms of these variables are attached in Graph 2. As a result, I have used a standard scaler in the 'sklearn' package to scale my quantitative variables. After standardizing quantitative features to their unit variances, their distributions are roughly normal.

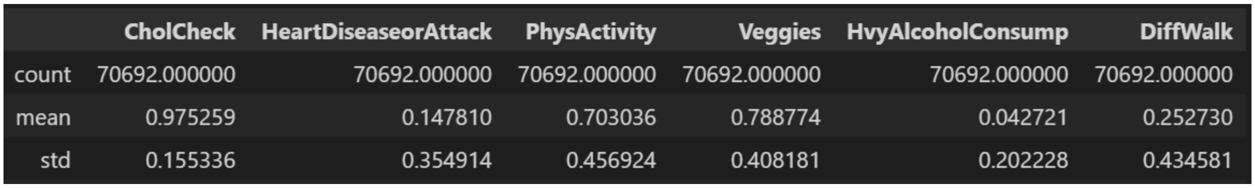
Regarding the binary features, the summary statistics and histograms suggest that features ‘CholCheck’, ‘HeartDiseaseAttack’, ‘PhysActivity’, ‘Veggies’, ‘HvyAlcholConsump’, ‘DiffWalk’ are not strictly balanced. The histograms are attached in Graph 1. From the summary statistics of these variables, attached in Table 2, we can see that the means of these imbalanced variables are far away from 0.5. This indicates these variables are imbalanced. However, given the binary variables are not suitable for standardization, and the dependent variable ‘diabetes’ is already balanced(shown in Graph 3), no standardization was performed for binary features.



Graph 1: Histograms of quantitative features with skewed distribution

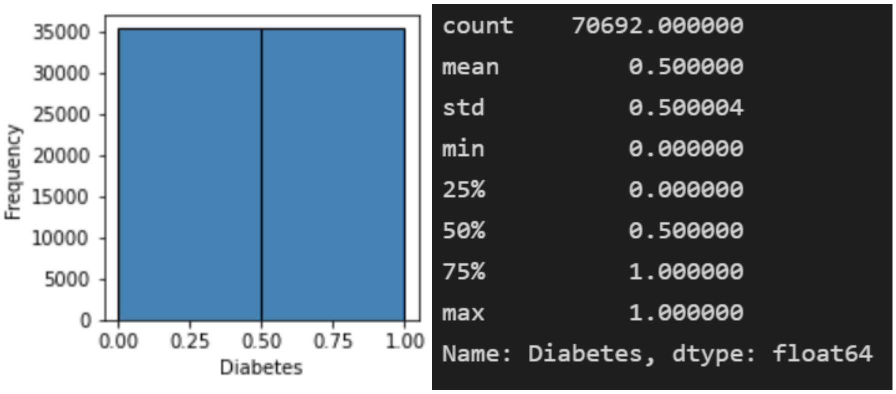


Graph 2: Histograms of Imbalanced Binary Features



	CholCheck	HeartDiseaseorAttack	PhysActivity	Veggies	HvyAlcoholConsump	DiffWalk
count	70692.000000	70692.000000	70692.000000	70692.000000	70692.000000	70692.000000
mean	0.975259	0.147810	0.703036	0.788774	0.042721	0.252730
std	0.155336	0.354914	0.456924	0.408181	0.202228	0.434581

Table 2: Summary Statistics of the Imbalanced Binary Features

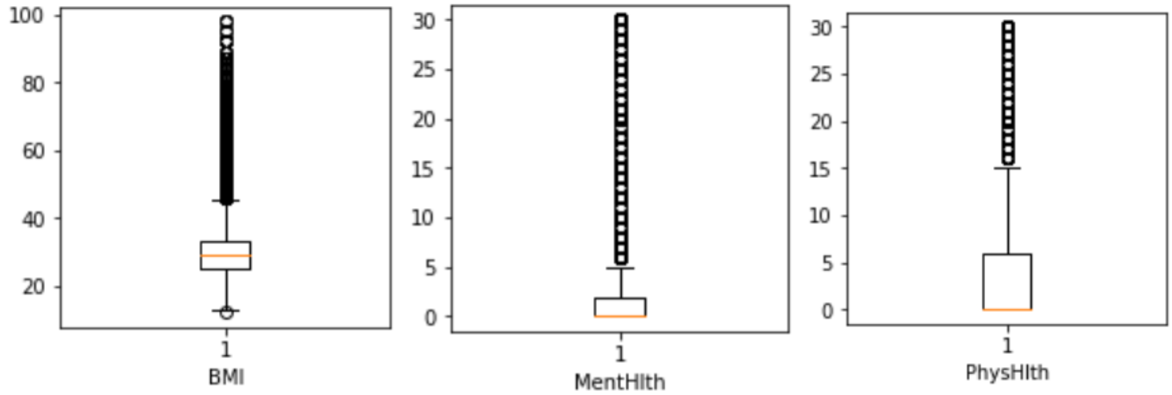


```
count    70692.000000
mean         0.500000
std          0.500004
min          0.000000
25%          0.000000
50%          0.500000
75%          1.000000
max          1.000000
Name: Diabetes, dtype: float64
```

Graph 3: Histogram and Summary Statistics of Diabetes

Outlier Check

To check if outliers are present in the dataset and misleading to the summary statistics, each quantitative feature was visualized using boxplots.



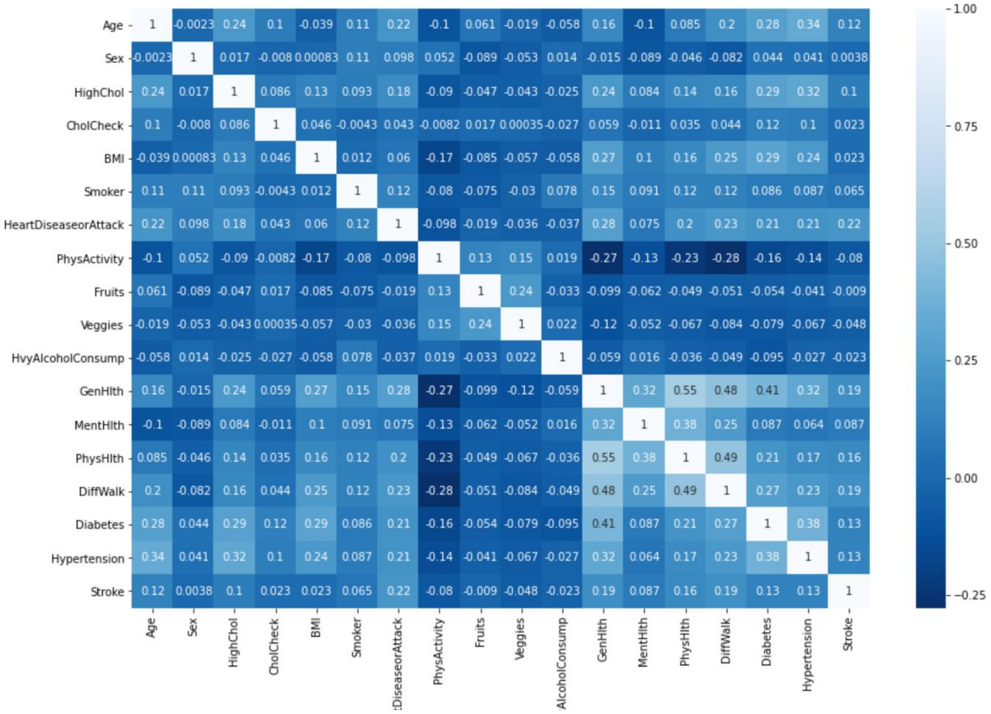
Graph 4: Boxplots of quantitative features identified with outliers

From Graph 4 we can see that features 'BMI', 'MentHlth', 'PhysHlth' all have outliers. This is consistent with the previous observations of their skewed distribution. If we remove these outliers from the dataset, their distributions will appear normal. However, given BMI is highly related to diabetes based on previous research, these outliers have great potential in contributing to the prediction of diabetes. As a result, the outliers of BMI are kept within the dataset. Given that 'MentHlth' and 'PhysHlth' variables are ordinal and the outliers are not falling out of their scale, these outliers are kept as well.

Given the nature of binary variables, there are no outliers detectable from its summary statistics or boxplots. Accordingly, no outliers are removed from the binary variables.

Correlation Check

To avoid multicollinearity, I created a heatmap to check the correlation between independent and dependent variables. In this heatmap(Graph 5), no pairs of correlated variables have a correlation higher than 0.5, which is the usual threshold that indicates multicollinearity. Therefore no variables are removed from the dataset.



	Age	Sex	HighChol	CholCheck	BMI	Smoker	DiseaseorAttack	PhysActivity	Fruits	Veggies	AlcoholConsump	GenHlth	MentHlth	PhysHlth	DiffWalk	Diabetes	Hypertension	Stroke
Age	1	-0.0023	0.24	0.1	-0.039	0.11	0.22	-0.1	0.061	-0.019	-0.058	0.16	-0.1	0.085	0.2	0.28	0.34	0.12
Sex	-0.0023	1	0.017	-0.008	0.00083	0.11	0.098	0.052	-0.089	-0.053	0.014	-0.015	-0.089	-0.046	-0.082	0.044	0.041	0.0038
HighChol	0.24	0.017	1	0.086	0.13	0.093	0.18	-0.09	-0.047	-0.043	-0.025	0.24	0.084	0.14	0.16	0.29	0.32	0.1
CholCheck	0.1	-0.008	0.086	1	0.046	-0.0043	0.043	-0.0082	0.017	0.00035	-0.027	0.059	-0.011	0.035	0.044	0.12	0.1	0.023
BMI	-0.039	0.00083	0.13	0.046	1	0.012	0.06	-0.17	-0.085	-0.057	-0.058	0.27	0.1	0.16	0.25	0.29	0.24	0.023
Smoker	0.11	0.11	0.093	-0.0043	0.012	1	0.12	-0.08	-0.075	-0.03	0.078	0.15	0.091	0.12	0.12	0.086	0.087	0.065
HeartDiseaseorAttack	0.22	0.098	0.18	0.043	0.06	0.12	1	-0.098	-0.019	-0.036	-0.037	0.28	0.075	0.2	0.23	0.21	0.21	0.22
PhysActivity	-0.1	0.052	-0.09	-0.0082	-0.17	-0.08	-0.098	1	0.13	0.15	0.019	-0.27	-0.13	-0.23	-0.28	-0.16	-0.14	-0.08
Fruits	0.061	-0.089	-0.047	0.017	-0.085	-0.075	-0.019	0.13	1	0.24	-0.033	-0.099	-0.062	-0.049	-0.051	-0.054	-0.041	-0.009
Veggies	-0.019	-0.053	-0.043	0.00035	-0.057	-0.03	-0.036	0.15	0.24	1	0.022	-0.12	-0.052	-0.067	-0.084	-0.079	-0.067	-0.048
HvyAlcoholConsump	-0.058	0.014	-0.025	-0.027	-0.058	0.078	-0.037	0.019	-0.033	0.022	1	-0.059	0.016	-0.036	-0.049	-0.095	-0.027	-0.023
GenHlth	0.16	-0.015	0.24	0.059	0.27	0.15	0.28	-0.27	-0.099	-0.12	-0.059	1	0.32	0.55	0.48	0.41	0.32	0.19
MentHlth	-0.1	-0.089	0.084	-0.011	0.1	0.091	0.075	-0.13	-0.062	-0.052	0.016	0.32	1	0.38	0.25	0.087	0.064	0.087
PhysHlth	0.085	-0.046	0.14	0.035	0.16	0.12	0.2	-0.23	-0.049	-0.067	-0.036	0.55	0.38	1	0.49	0.21	0.17	0.16
DiffWalk	0.2	-0.082	0.16	0.044	0.25	0.12	0.23	-0.28	-0.051	-0.084	-0.049	0.48	0.25	0.49	1	0.27	0.23	0.19
Diabetes	0.28	0.044	0.29	0.12	0.29	0.086	0.21	-0.16	-0.054	-0.079	-0.095	0.41	0.087	0.21	0.27	1	0.38	0.13
Hypertension	0.34	0.041	0.32	0.1	0.24	0.087	0.21	-0.14	-0.041	-0.067	-0.027	0.32	0.064	0.17	0.23	0.38	1	0.13
Stroke	0.12	0.0038	0.1	0.023	0.023	0.065	0.22	-0.08	-0.009	-0.048	-0.023	0.19	0.087	0.16	0.19	0.13	0.13	1

Graph 5: Correlation plot of all variables

Questions of Interest

In this project, I intend to answer the following questions:

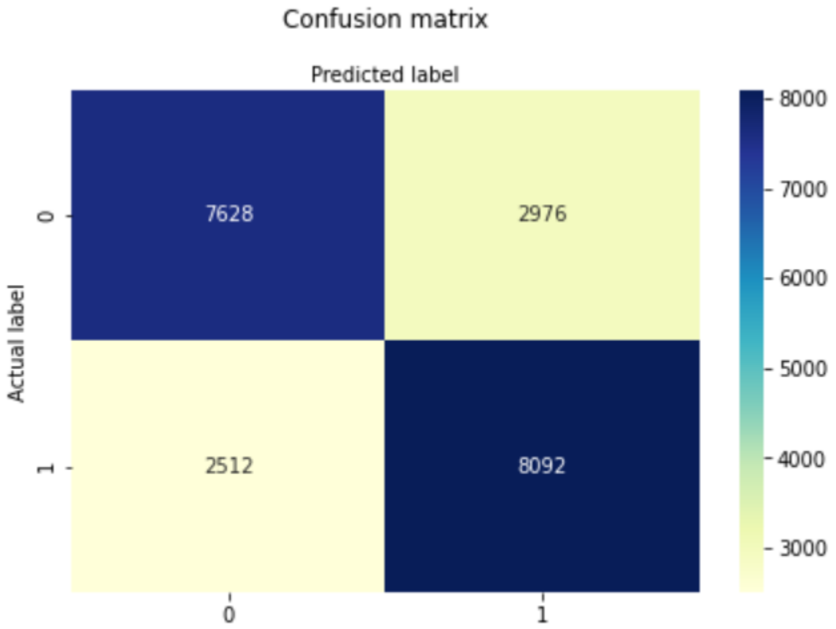
1. How accurate is our prediction of diabetes using machine learning classification algorithms?
2. Which classification model predicts the best regarding accuracy?
3. What are some top features identified as key risk factors for diabetes in the analysis?
4. Does the analysis fit the expectation?

Models and Results

Provided with the analytical goal, this project will aim to conduct a binary classification task. I will use the logistic regression algorithm, random forest algorithm, and k-nearest-neighbor algorithm to predict diabetes.

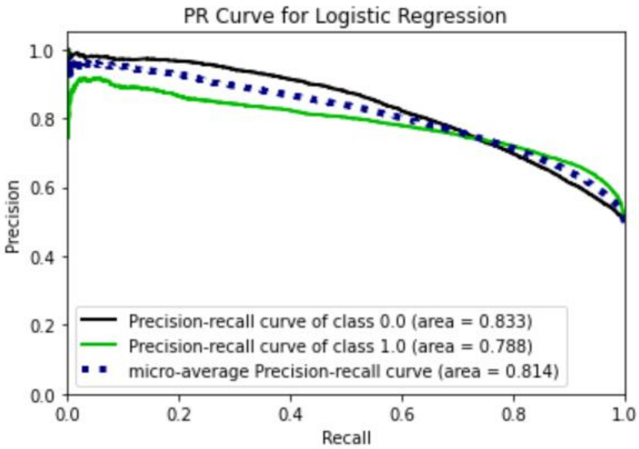
Logistic regression

After splitting the dataset into training and testing datasets, a logistic regression model was implemented to predict diabetes status. The model was validated through a testing dataset. A confusion matrix of a logistic regression model is drawn in Graph 6.



Graph 6: Confusion Matrix of Logistic Regression Model

From this graph, we can calculate the true positive rate, also known as the recall rate, as 76.31%. The model has also yielded a test accuracy of 74.12%, a precision rate of 73.11%, and an F1-score of 0.7358. I then proceeded to draw a precision-recall curve for the logistic regression, attached in Graph 7.

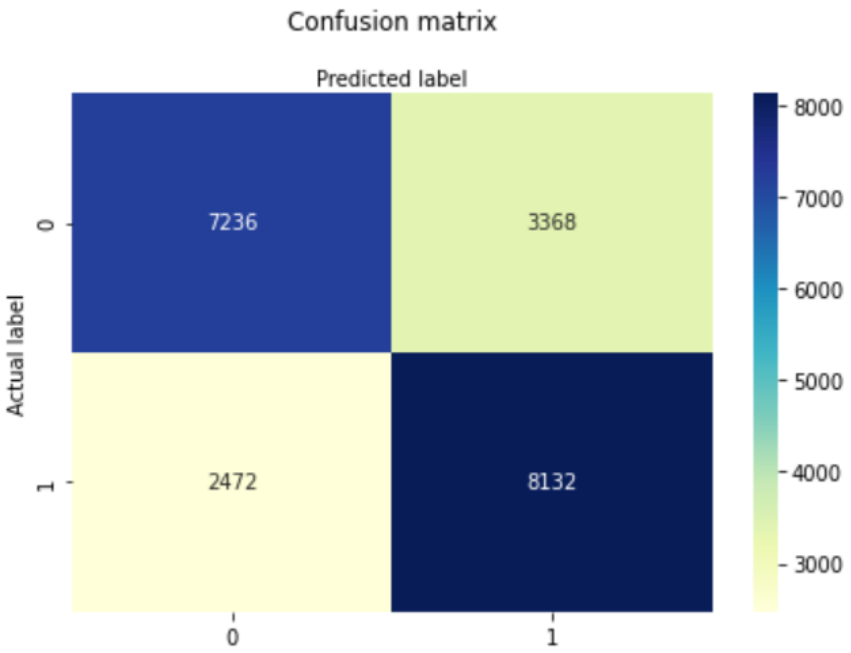


Graph 7: Precision-recall curve of Logistic Regression model

The precision-recall curve shows the tradeoff between precision and recall rate based on changing threshold. [4] In our cases, the tradeoff refers to balancing events of correctly identifying patients with risk of diabetes and avoiding false alarms among patients. From our purposes of providing the population at risk of a proper alarm of diabetes, we would like to balance the precision and recall scores. High scores of both precision and recall score indicate that the classifier is returning accurate results and as well as returning a majority of all positive results. Thus the area covered by PR Curve is negatively proportional to the tradeoff between prediction and recall score. In the logistic regression model, the area covered by the PR curve for the diabetes class is 0.788, and the area of the non-diabetes class is 0.833, which is relatively high. This graph will be further used for comparison with other derived models.

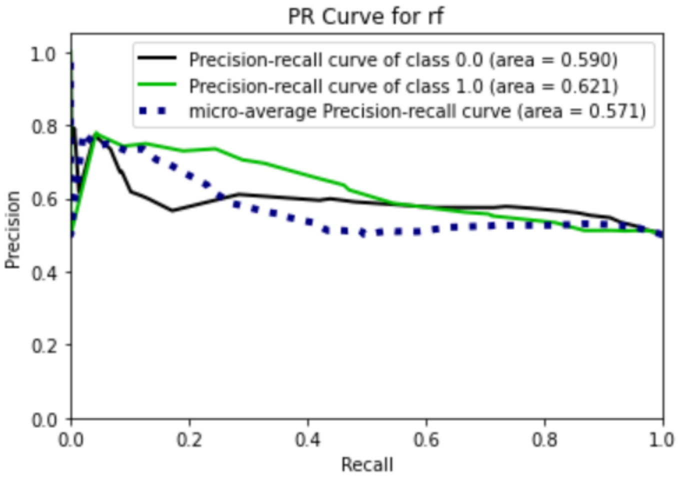
Random Forest Model

Following the same procedure of model fitting, I have fitted a random forest model with the parameter n = 100. The results of this model are presented with a confusion matrix and precision-recall curve in Graphs 8 and 9.



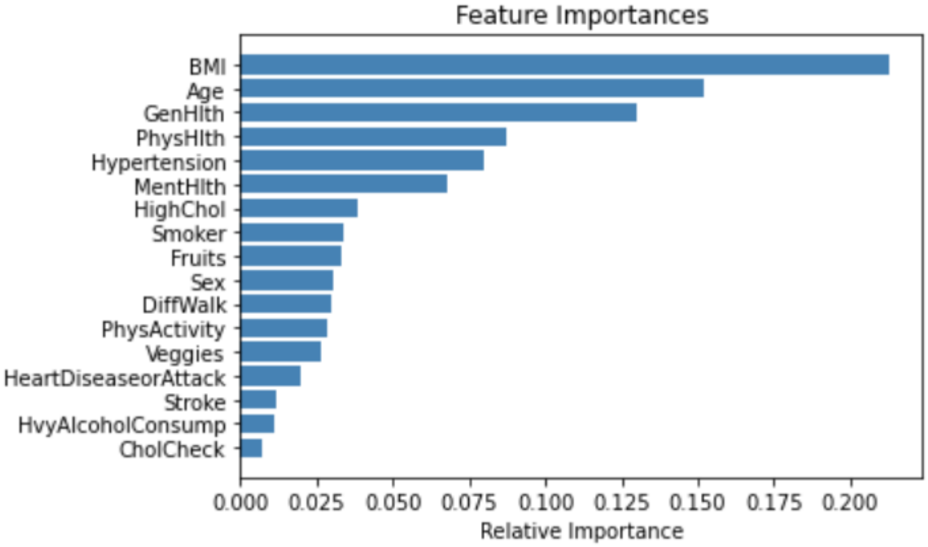
Graph 8: Confusion Matrix of Random Forest Model

From this graph, we know that the model has yielded a test accuracy of 72.20%, a precision rate of 73.32%, a recall rate of 76.34%, and an F1-score of 0.7047. I then proceeded to draw a precision-recall curve for the model, attached in Graph 9.



Graph 9: Precision-Recall Curve of Random Forest Model(n=100)

Compared to the logistic regression model, the random forest model has a relatively smaller area covered by the PR curve. This indicates that the recall rate and precision rate have a potential for high tradeoffs, and a poorer detection of rare events when the model is used on other datasets. To further understand how the algorithm is identifying the key risk factors associated with diabetes, I have plotted the feature importance in Graph 10.



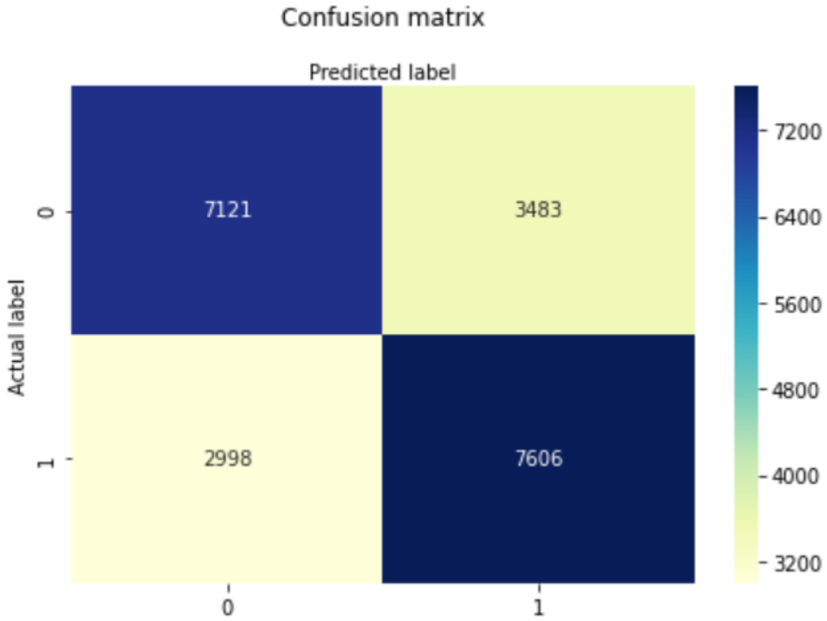
Graph 10: Feature Importance of all Features in Random Forest Model(n=100)

From this plot, we can see the most important features identified in the random forest models are: BMI, age, General Health, Physical Activity, Hypertension, and Mental Health. The numeric values of feature importance are attached in Table 1 inside Appendix. Previous research has confirmed that BMI, age, hypertension, and physical activities are risk factors for diabetes.[5,6] The identification of 'General health' as a key risk factor is also reasonable, as patients diagnosed with diabetes have a higher chance of perceiving themselves with lower general health status.

What is unexpected is to observe mental health serving as a top predictor in the model. Surprisingly, researchers have found that individuals living with type 1 or type 2 diabetes are at increased risk for depression, anxiety, and eating disorder diagnoses. [7] Therefore we can conclude that the key risk factors identified by the analysis are rather consistent with scientific findings on diabetes thus far.

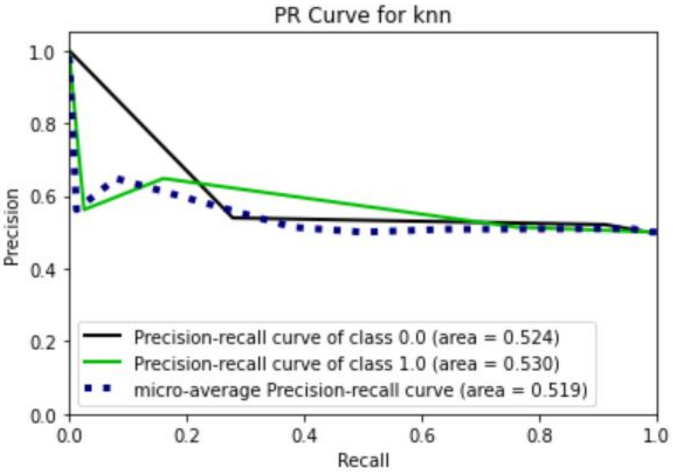
K-Nearest-Neighbors Model

Multiple k have been fitted to KNN and the KNN model with k = 7 was chosen based on its performance. The results of this model are presented with a confusion matrix and precision-recall curve in Graphs 11 and 12.



Graph 11: Confusion matrix of KNN Model(k=7)

From this graph, we know that the model has yielded a test accuracy of 71.98%, a precision rate of 70.29%, a recall rate of 76.14%, and an F1-score of 0.7310. I then proceeded to draw a precision-recall curve for the model, attached in Graph 9.



Graph 12: Confusion matrix of KNN Model(k=7)

Compared to the logistic regression model, the KNN model has a relatively smaller area covered by the PR curve. This indicates that regarding precision and recall scores, the logistic regression model has outperformed the KNN model.

Model Result Summary

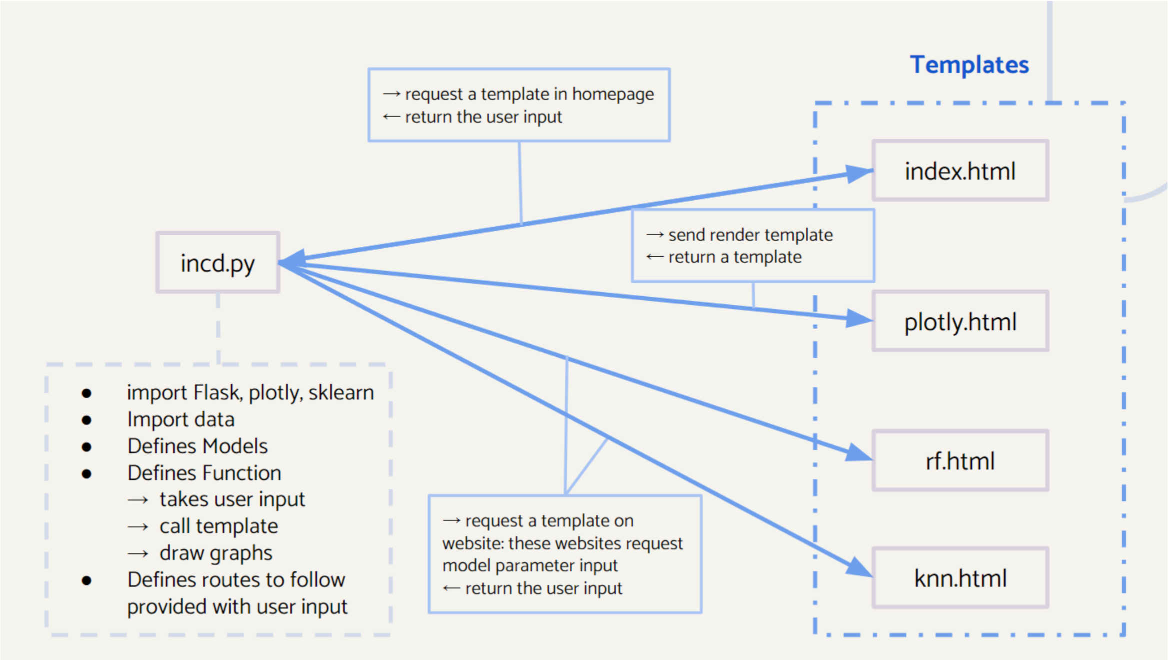
Based on all of the constructed models, I have created Table 3 as a summary of the evaluation metrics. Based on Table 3 we can see that regarding the accuracy, F1 score, and Area under the PR curve, the logistic regression model has outperformed the random forest and KNN model. Thus we can conclude that the logistic regression model has yielded the best prediction of diabetes.

Model name	**Accuracy**	**Precision**	**F-1 score**	**Recall**	**Area under PR (class = 0)**	**Area under PR (class = 1)**
Logistic Regression	74.12%	73.11%	0.7358	76.13%	0.833	0.788
Random Forest	72.20%	73.32%	0.7047	76.34%	0.590	0.621
KNN	71.98%	70.29%	0.7310	76.14%	0.524	0.530

Table 3: Summary of Model Metrics

Server API and Web Front-End

To visualize the results online, I used the packages Flask' and 'plotly' to build up an API that defines the interaction between the user and the website. The whole procedure was realized using py file and HTML files. A rough framework of how the interactions were built was defined in Graph 13 below.



Graph 13: Framework of Server API and Front-end

Based on this graph, I will use a call on analytical results of the random forest model to explain a complete walkthrough of the framework. The 'incd.py' file is the file that defines the routes, functions, request data, and renders the template. This file will import 'Flask', 'plotly', and 'sklearn' packages. It will define all of the constructed models, the analysis about to run upon user selection, and the layout of analytical results.

In this project, I have chosen to present my results using a confusion matrix, and thus my results for each model will be encoded using a confusion matrix and plotted using plotly. The 'incd.py' file will specify which HTML to direct the user to, based on request. Upon the beginning of the request, users will be directed to the home page, and 'incd.py' will automatically locate 'index.html' inside my 'templates' folder, as defined by the route. It will ask users to pick one model to see its analytical results. If the users have selected the model 'Random Forest', the route defined by the py file will direct the user to 'rf.html', which specifies the layout of the website that requires users to implement a parameter: number of trees for random forest model. After selecting the number of trees, users will be directed back to the incd.py file, where it defines a route 'rf/info.html'. When this route is triggered, the local terminal will start fitting the random forest model, calculating the confusion matrix, dumping the results to a JSON file, and then requesting a plotly template for plotting interactive graphs. The plotly template is defined by 'plotly.html'. It will request the calculated JSON file and plot the desired graph as defined in the py file. At the end of this cycle, a heatmap of the confusion matrix will be drawn on the page.

This is how the front end of the website is realized, upon the request of the random forest model results.

Given that the logistic regression model does not require any implementation of parameters, calling the result of logistic regression does not require an additional HTML. Having 'incd.py' and 'plotly.html' is enough to do the job. For the KNN classification model, the routes resemble those of the random forest model, except that it uses 'knn.html' instead of 'rf.html' to request user input.

To further improve the appearance of the website, I used CSS style in HTML to change the background colors and the font sizes of the text.

Conclusion

By the end of this project, I will be able to answer the questions of interest proposed in previous sections. The best classification model identified is the logistic regression model, reaching the highest prediction accuracy of 74.12%. The analysis has shown, key risk factors contributing to diabetes are BMI, age, General Health, Physical Activity, Hypertension, and Mental Health. The discovery of mental health status listed among key risk factors is relatively surprising, but there are emerging researches confirming the diagnosis of diabetes is associated with a higher risk of mental health-related disorders. The results fit into our expectations.

However, there are several restrictions to this project regarding its application. Given the project is limited by the number of features, and the predicted accuracy of the selected logistic regression model is also not as high as expected, its application to real-life diabetes prediction is of limited effectiveness. Furthermore, when establishing the website, I found that the calculation of the random forest and KNN models is pretty time-consuming. To solve this issue, I suggested an ideal range of parameters on the website to prevent users from implementing an overly time-consuming model.

Regardless of these posed limitations, I hope this project can provide a basic exploration of diabetes regarding health-related information and insights into key risk factors associated with diabetes.

Reference

1. Zimmet, P.Z. *et al.* (2014) "Diabetes: A 21st Century Challenge," *The Lancet Diabetes & Endocrinology*, 2(1), pp. 56–64. Available at: https://doi.org/10.1016/s2213-8587(13)70112-8.
2. Chuks, P. (2022) *Diabetes, hypertension and stroke prediction*, *Kaggle*. Available at: https://www.kaggle.com/datasets/prosperchuks/health-dataset/code (Accessed: December 16, 2022).
3. *CDC - BRFSS* (2022) *Centers for Disease Control and Prevention*. Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/brfss/index.html (Accessed: December 16, 2022).
4. *Precision-recall* (no date) *scikit*. Available at: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html#:~:text=The%20precision%2Drecall%20curve%20shows,a%20low%20false%20negative%20rate. (Accessed: December 16, 2022).
5. *Diabetes risk factors* (2022) *Centers for Disease Control and Prevention*. Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/diabetes/basics/risk-factors.html (Accessed: December 16, 2022).
6. “Diabetes, hypertension, and kidney disease in the Pima Indians” (2004) *The Kidney and Hypertension in Diabetes Mellitus*, pp. 879–892. Available at: https://doi.org/10.3109/9780203326916-55.
7. Ducat, L., Philipson, L.H. and Anderson, B.J. (2014) “The mental health comorbidities of diabetes,” *JAMA*, 312(7), p. 691. Available at: https://doi.org/10.1001/jama.2014.8040.

Appendix.

Variable Summary

Variable Name	Variable Type	Information	Feature importance	Dependent / independent
Diabetes	Categorical (binary)	Participant Diabetes Status	-	Dependent
BMI	Quantitative	Participant Body Mass Index	0.210060	Independent

Age	Quantitative	13-level age categories:
		1=18-24;
		9 = 60-64;
		13 = 80 or older
			0.152212	Independent

GenHlth	Ordinal	General health self assessment (scale 1-5):
		1 = excellent;
		2 = very good;
		3 = good;
		4 = fair;
		5 = poor
			0.122406	Independent

PhysHlth	Ordinal	General physical health self assessment (scale 1-5):
		1 = excellent;
		2 = very good;
		3 = good;
		4 = fair;
		5 = poor
			0.088283	Independent

Hypertension	Categorical (binary)	History of hypertension: 0 = no; 1 = yes	0.085381	Independent
MentHlth	Quantitative	Total days of	0.067734	Independent

		poor mental health in past 30 days (scale 1-30)		
HighChol	Categorical (binary)	Participant high Cholesterol status: 0 = no high cholesterol; 1 = high cholesterol	0.042057	Independent
Smoker	Categorical (binary)	Participant Smoking History: 0 = have not smoked 100 cigarettes in your life; 1 = have smoked 100 cigarettes	0.033824	Independent
Fruits	Categorical (binary)	Consume fruit 1 or more times per day: 0 = no; 1 = yes	0.033270	Independent
Sex	Categorical (binary)	Participant's gender: 1= male; 0 = female.	0.030674	Independent
DiffWalk	Categorical (binary)	Serious difficulty walking or climbing stairs: 0 = no; 1 = yes	0.029323	Independent
PhysActivity	Categorical (binary)	Physical Activity in past 30 days not including job: 0 = no; 1 = yes	0.028681	Independent
Veggies	Categorical (binary)	Consume vegetables 1 or more times per day: 0 = no; 1 = yes	0.026203	Independent

HeartDiseaseorAttack	Categorical (binary)	Coronary heart disease or myocardial infarction: 0 = no; 1 = yes	0.020922	Independent
Stroke	Categorical (binary)	History of stroke: 0 = no; 1 = yes	0.011991	Independent
HvyAlcoholConsump	Ordinal	adult men >=14 drinks per week and adult women>=7 drinks per week: 0 = no 1 = yes	0.011073	Independent
CholCheck		Participant’s cholesterol check status: 0 = no check in 5 years; 1 = yes cholesterol check in 5 years	0.006734	Independent

Table 1: Variable Description and Feature Importance