Title: Predict diabetes based on diagnostic measures

1. Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation and the benefits that your analysis provides to the business or stakeholders of this data.

- The objective of the analysis is to develop a machine learning model for diabetes prediction given the diagnostic measures so patients can receive timely proper treatment to minimise the risk of developing other complications. It will be helpful as well to healthcare professionals in developing a health strategy and raise public awareness about diabetes.

2. Brief description of the data set you chose, a summary of its attributes, and an outline of what you are trying to accomplish with this analysis.

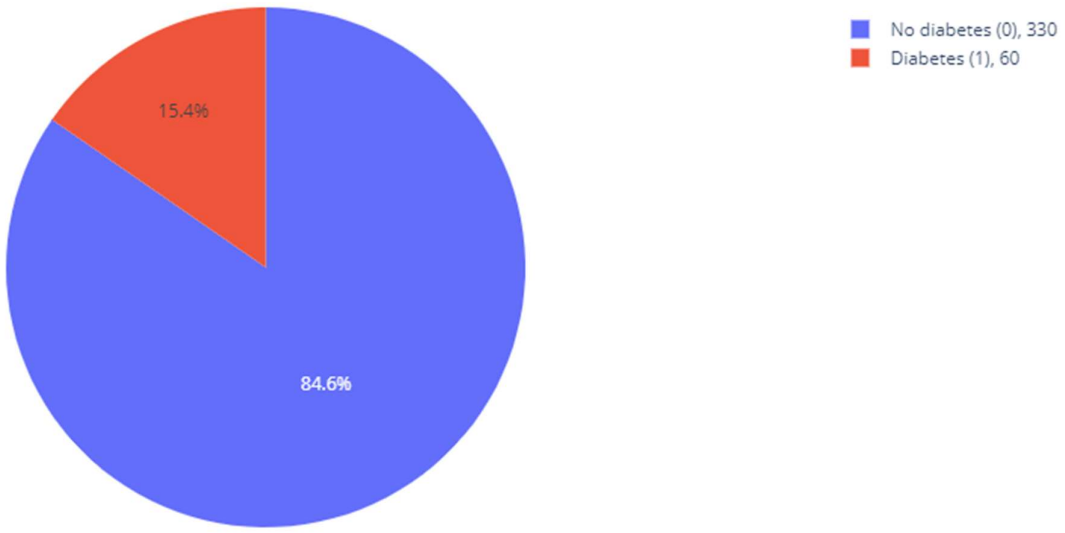
- The data was uploaded to Kaggle by Houcem Benmansourand. It's called "Predict diabetes based on diagnostic measure" version one and originally from the National Institute of Diabetes and Digestive and Kidney Diseases. It is accessible via this URL, https://www.kaggle.com/houcembenmansour/predict-diabetes-based-on-diagnostic-measures.
- The dataset contains 390 entries from 390 different patients and 16 data columns in CSV file format.
- The 16 data columns are the patient number, cholesterol, glucose, HDL cholesterol, cholesterol/HDL ratio, age, gender (male or female), height, weight, BMI, systolic blood pressure, diastolic blood pressure, waist size, hip size, waist/hip ratio and diabetes status (diabetes or no diabetes). All are numerical values except for gender and diabetes status which are dichotomous.
- The web-based, interactive environment Jupyter Notebook via Anaconda Navigator individual edition i.e., Anaconda3 2021.11 (Python 3.9.7 64-bit) was used to perform the analysis.
- The CSV data file was read into Jupyter Notebook to conduct the analysis. Data cleanliness was assessed and appropriate action was taken accordingly to ensure data quality. Exploratory data analysis then be performed to study the data characteristics through summary statistics and graphical representation. Data pre-process such as feature encoding and six classification models were assessed.

3. Brief summary of data exploration and actions taken for data cleaning and feature engineering.

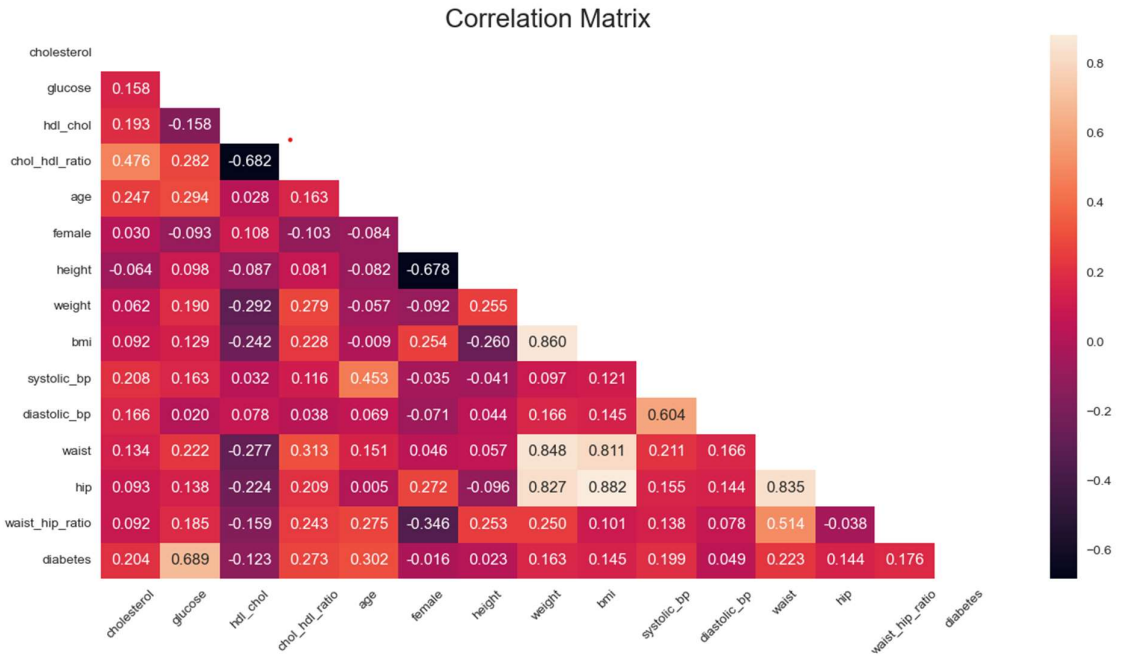
- There are 390 observation rows from 390 different patients.
- Following is the last five entries of the data:

patient_number	cholesterol	glucose	hdl_chol	chol_hdl_ratio	age	gender	height	weight	bmi	systolic_bp	diastolic_bp	waist	hip	waist_hip_ratio	diabetes
386	227	105	44	5.2	83	female	59	125	25.2	150	90	35	40	0.88	No diabetes
387	226	279	52	4.3	84	female	60	192	37.5	144	88	41	48	0.85	Diabetes
388	301	90	118	2.6	89	female	61	115	21.7	218	90	31	41	0.76	No diabetes
389	232	184	114	2.0	91	female	61	127	24.0	170	82	35	38	0.92	Diabetes
390	165	94	69	2.4	92	female	62	217	39.7	160	82	51	51	1.00	No diabetes

- There are no duplicates in the data.
- There are no missing values from the data.
- Customer number will not add value to diabetes prediction hence it was dropped from the analysis dataset.
- The gender column is renamed to female before it's label-encoded to 0 and 1 where 0 means male and 1 means female to avoid confusion.
- Diabetes column is label-encoded to 0 and 1 where 0 refers to no diabetes and 1 refers to yes to diabetes.
- We have an imbalanced dataset because both classes are not equally distributed among all observations. That is, the proportion of no diabetes (84.6% or 330 patients) is very much higher than the proportion of diabetes (15.4% or 60 patients) as depicted by the pie chart below.



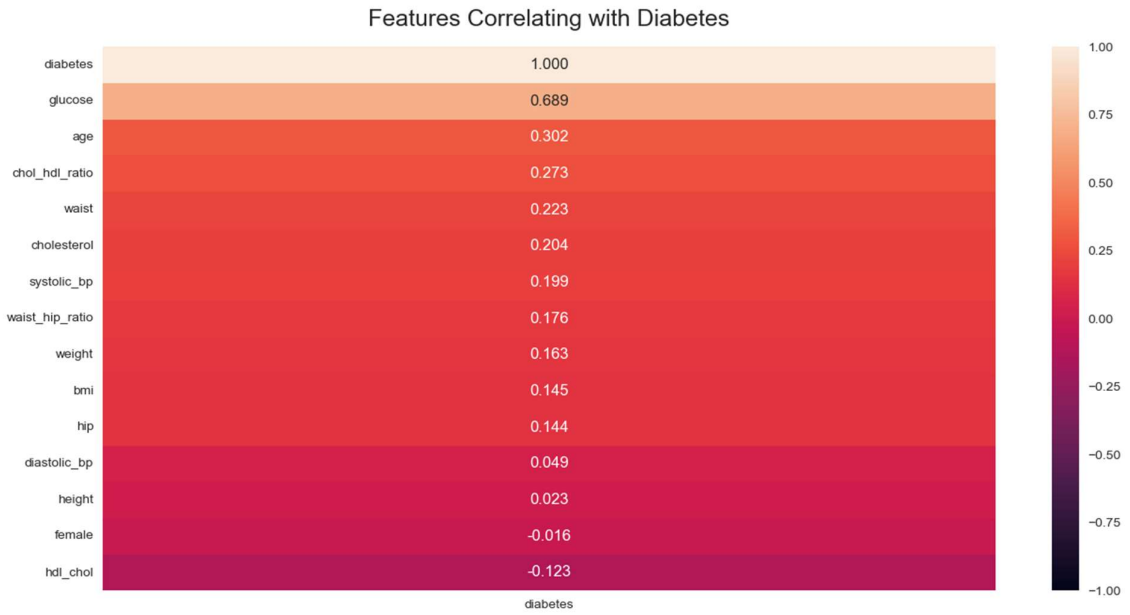
- The following is the correlation matrix of the data:



High correlations are found between the following features:

- Glucose and diabetes (0.689)
- Cholesterol-to-HDL ratio and HDL cholesterol (-0.682)
- Female and height (-0.678)
- Weight with hip (0.827), waist (0.848) and BMI (0.86)
- BMI with hip (0.882) and waist (0.811)
- Systolic BP and diastolic BP (0.604)
- Waist and hip (0.835)

Looking at this correlations, it makes sense to have anthropometric measurements (weight, hip, waist and BMI) correlating with each other. Glucose has been the standard biomarker for diabetes and it shows here as well. The rest of the biomarkers are useful indicators for monitoring patients at risk for diabetes in healthcare.



The above shows the correlation coefficient of each feature on diabetes in descending order. Glucose is most correlated following by age with diabetes.

- The following table is the summary statistics by diabetes group. We can see that the average of glucose is much more elevated for diabetes patients (194.2) than patients without diabetes (91.6). On average, diabetes patients are older (58.4) and weighted heavier (192.8) with higher cholesterol (228.6).

Feature	Statistics	No diabetes	Diabetes	Total
cholesterol	**count**	330	60	390
	mean	203.35	228.60	207.23
	std	41.08	56.53	44.67
	min	78	115	78
	25%	175.25	195.75	179
	50%	199	219	203
	75%	226.75	249.75	229
	max	347	443	443
glucose	**count**	330	60	390
	mean	91.56	194.17	107.34
	std	26.87	77.44	53.80
	min	48	60	48
	25%	79	120	81
	50%	86.5	186	90
	75%	97	241.25	107.75
	max	371	385	385
HDL	**count**	330	60	390
	mean	51.17	45.28	50.27
	std	17.23	16.85	17.28

	min	12	23	12
	25%	40	33.75	38
	50%	47	42	46
	75%	59.75	55	59
	max	120	114	120
cholesterol/HDL ratio	**count**	330	60	390
	mean	4.32	5.64	4.52
	std	1.43	2.63	1.74
	min	1.5	2	1.5
	25%	3.2	4	3.2
	50%	4.1	5.2	4.2
	75%	5.3	7	5.4
	max	10.6	19.3	19.3
age	**count**	330	60	390
	mean	44.66	58.4	46.77
	std	16.11	13.12	16.44
	min	19	26	19
	25%	32	50.75	34
	50%	42	59.5	44.5
	75%	55.75	65.25	60
	max	92	91	92
height	**count**	330	60	390
	mean	65.91	66.17	65.95
	std	3.94	3.82	3.92
	min	52	59	52
	25%	63	63	63
	50%	66	67	66
	75%	69	69	69
	max	76	75	76
weight	**count**	330	60	390
	mean	174.60	192.83	177.41
	std	39.84	40.34	40.41
	min	99	123	99
	25%	146	166.5	150.25
	50%	170	189	173
	75%	195.75	211	200
	max	325	320	325
bmi	**count**	330	60	390
	mean	28.37	31.02	28.78
	std	6.58	6.29	6.60
	min	15.2	21.5	15.2
	25%	23.6	25.95	24.1
	50%	27.5	30.2	27.8
	75%	31.78	33.8	32.28
	max	55.8	51.4	55.8
systolic BP	**count**	330	60	390
	mean	135.2	147.77	137.13
	std	22.76	20.50	22.86

	min	90	100	90
	25%	120	138	122
	50%	132	145	136
	75%	142	160	148
	max	250	200	250
diastolic BP	**count**	330	60	390
	mean	83.01	84.85	83.29
	std	13.60	12.91	13.50
	min	48	50	48
	25%	75	77.25	75
	50%	82	87	82
	75%	90	92.5	90
	max	124	118	124
waist	**count**	330	60	390
	mean	37.32	40.88	37.87
	std	5.60	5.75	5.76
	min	26	30	26
	25%	33	36	33
	50%	37	40.5	37
	75%	41	44	41
	max	53	56	56
hip	**count**	330	60	390
	mean	42.65	44.9	43.00
	std	5.64	5.476297	5.66
	min	30	37	30
	25%	39	41	39
	50%	42	44.5	42
	75%	46	48	46
	max	64	62	64
waist/hip ratio	**count**	330	60	390
	mean	0.88	0.91	0.88
	std	0.07	0.08	0.07
	min	0.68	0.75	0.68
	25%	0.83	0.87	0.83
	50%	0.87	0.91	0.88
	75%	0.92	0.95	0.93
	max	1.14	1.14	1.14

The proportion of either no diabetes or diabetes is only differed slightly between males and females. Overall, we have a larger proportion of females (58.46%) in the data.

	No diabetes		**Diabetes**		**Total**	
	n	%	n	%	n	**%**
Male	136	83.95	26	16.05	162	41.54
Female	194	85.09	34	14.91	228	58.46
Total	330	84.62	60	15.38	390	100

4. Summary of training at least three different classifier models, preferably of different nature in explainability and predictability. For example, you can start with a simple logistic regression as a baseline, adding other models or ensemble models. Preferably, all your models use the same training and test splits, or the same cross-validation method.

- A stratified train-test split with 80% and 20% train and test set respectively was used due to imbalanced number of diabetes (15.4%) and non-diabetes (84.6%) before starting training the logistic models, random forests and Naïve Bayes classifiers.

  Initial classification model was the weighted logistic model with L2 regularization by specifying the class weight equals to balanced was fitted. Because all available features are to be used for diabetes prediction regardless of if there is multicollinearity, logistic regression with L2 regularization was preferred over of L1 regularization. Using L1 regularization will reduce the number of input features due to collinearity for having a weight of zero. To compare, the hybrid resampling method of both upsampling and downsampling of SMOTE and Tomek's link method was applied for logistic regression with L2 regularization. The former, weighted logistic model with L2 regularization performs better according to their evaluation metrics.

  Tree-based classifier was explored. One being the weighted random forest using Gini was compared with random forest with SMOTE and Tomek's sampling method. The former gives better evaluation metrics.

  Lastly, the Naïve Bayes classifier. It was compared against that with SMOTE and Tomek's sampling method. Naïve Bayes classifier without SMOTE and Tomek's sampling method is the best among all the classifiers based on the evaluation metrics.

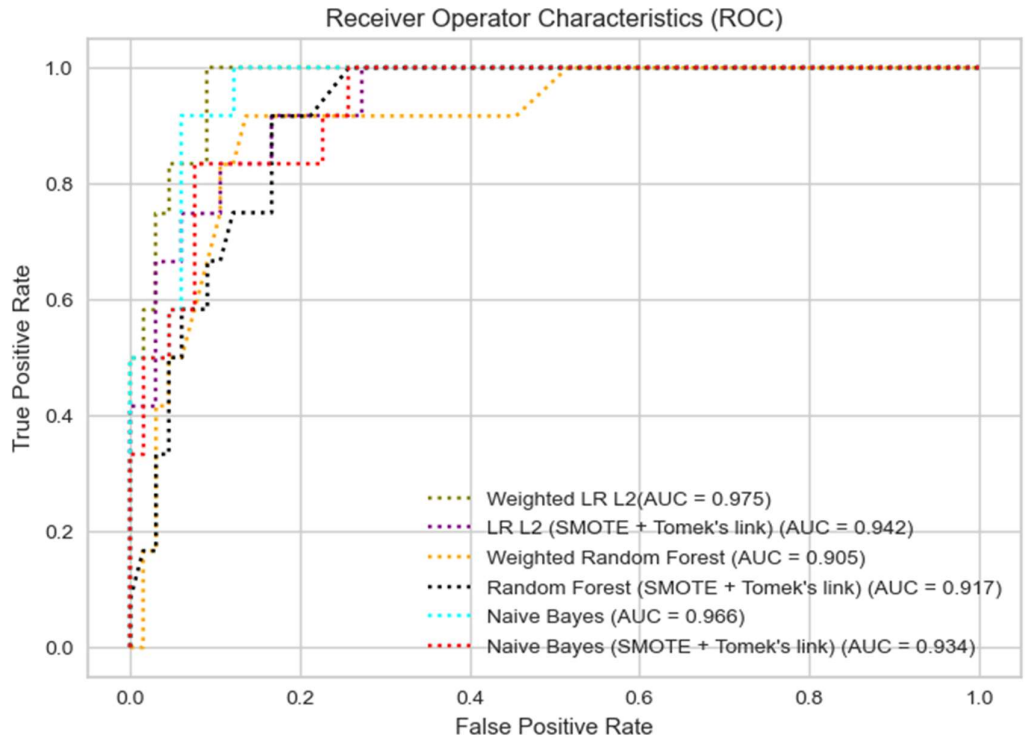
5. A paragraph explaining which of your classifier models you recommend as a final model that best fits your needs in terms of accuracy and explainability.

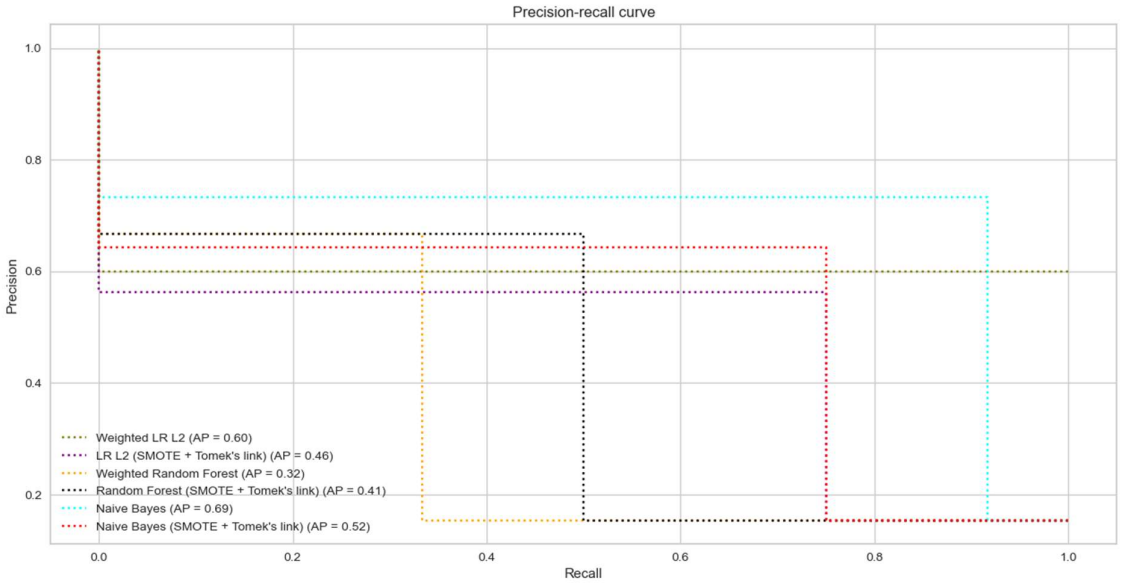
Confusion matrix of all the six classifiers:

Classifier			Predicted: No diabetes	Predicted: Diabetes
Weighted LR with L2 regularization	**Actual**	**No diabetes**	58	8
		Diabetes	0	12
Logistic Regression L2 regularization (SMOTE+Tomek's link)	**Actual**	**No diabetes**	59	7
		Diabetes	3	9
Weighted Random Forest (Gini)	**Actual**	**No diabetes**	64	2
		Diabetes	8	4
Random Forest (SMOTE + Tomek's link)	**Actual**	**No diabetes**	63	3
		Diabetes	6	6
Naïve Bayes	**Actual**	**No diabetes**	62	4
		Diabetes	1	11
Naïve Bayes (SMOTE + Tomek's link)	**Actual**	**No diabetes**	61	5
		Diabetes	3	9

Performance metrics:

Classifier	Accuracy	Precision	Recall	F1 score	AUC	AUPRC
Weighted LR with L2 regularization	0.897	0.600	1.000	0.750	0.975	0.600
Logistic Regression L2 regularization (SMOTE+Tomek's link)	0.846	0.563	0.750	0.643	0.942	0.460
Weighted Random Forest	0.872	0.667	0.333	0.444	0.905	0.325
Random Forest (SMOTE + Tomek's link)	0.885	0.667	0.500	0.571	0.917	0.410
Naïve Bayes	0.936	0.733	0.917	0.815	0.966	0.685
Naïve Bayes (SMOTE + Tomek's link)	0.897	0.643	0.750	0.692	0.934	0.521





Based on the performance metrics and both ROC and precision-recall plots above of all the six classifiers above, Naïve Bayes shows as the best classifier. It has the highest area under the precision-recall curve (AUPRC) and F1 score given the data is not balanced. Accuracy (0.936) is also the best despite not a recommended performance metric for data imbalance. It also gives the least misclassifications (false negative=1 and false positive=4).

6. Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your classifier model.

- Diabetes is a very common disease nowadays that can happen to anyone at any age. There are three main types, Type 1, Type 2 or gestational diabetes. Type I is genetic related whereas Type 2 is largely diet related and gestational diabetes is pregnancy related.

  Without timely proper diabetes treatment, patients will be at risk of developing health complications regardless of any types. Upon exploring the six classification models all considering the same features, Naïve Bayes shows as the best model for diabetes prediction according to the evaluation metrics with the highest AUPRC and least number of misclassifications from statistical standpoint. However, weighted logistic regression with L2 regularization model manages to predict all diabetes cases correctly and it is important in the healthcare industry because patient safety is the utmost concern. But the misclassification of non-diabetes to diabetes may cost burden to the healthcare system to some degree.

7. Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model after adding specific data features that may help you achieve a better explanation or a better prediction.

- Naïve Bayes model assumes independence between features. Further investigation on this aspect will be needed to assess its impact on the model's prediction performance. Moreover, we need to consider more features such as diet, lifestyle and other laboratory measures to introduce more information if they will improve the model in achieving zero false negative from patient safety perspective with reduced number of false positive from healthcare burden perspective. One will favour over the other depending on the business needs. Other classification algorithms such as neural network or K-nearest neighbours will be worth looking into.

References

1. https://www.kaggle.com/houcembenmansour/predict-diabetes-based-on-diagnostic-measures.
2. https://www.coursera.org/learn/supervised-machine-learning-classification/home/welcome
3. https://inria.github.io/scikit-learn-mooc/python_scripts/metrics_classification.html
4. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.PrecisionRecallDisplay.html#sklearn.metrics.PrecisionRecallDisplay
5. https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html
6. https://www.cdc.gov/diabetes/basics/diabetes.html
7. https://towardsdatascience.com/demystifying-roc-and-precision-recall-curves-d30f3fad2cbf#:~:text=The%20precision%2Drecall%20(PR)%20curve%20plots%20the%20precision%20versus,recall%20and%20a%20high%20precision.
8. https://www.justintodata.com/machine-learning-model-evaluation-metrics/
9. https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/