Predicting diabetic risk:
Modeling fasting blood glucose level from dietary data

Siegfried Leung

Goal:
This project will be developing predictive models to predict fasting blood glucose level from dietary and basic demographic information. The idea is to provides a tool to perform a virtual type 2 diabetes screening based on one's diet to identify unaware individuals, who might have pre-diabetes or diabetes, for further testing and consultation.

Targeted client:
Health institutions and doctors that would perform health check and disease screening. The predictive model, which will require inputs from simple dietary surveys, will help to identify patients with potential underlying diabetic conditions for further examination. Compared to measuring fasting blood glucose level from a blood drawn, this virtual screening approach, as the very first pass of screening, could provide a quick and non-invasive assessment of the potential diabetic risk.

Data:
The National Health and Nutrition Examination Survey (NHANES) from 2013 to 2014 as curated on Kaggle will be used for this study. The NHANES is an annual study carried out by the National Center for Health Statistics (NCHS), a part of the Centers for Disease Control and Prevention (CDC), to evaluate the health and nutritional status of those reside in the United States, producing the national health statistics. This annual study, consisted of both interviews and physical examinations, surveys 5000 individuals that represents the national population.

The NHANES collects a wide variety of information. The interview portion includes demographic, socioeconomic, dietary and health questionnaires. The examination portion includes laboratory tests as well as surveys of medical, dental, and physiological measurements. This study, focusing on modeling the relationships between diets and diabetes, utilize the following data from the 2013-2014 survey:

1) Laboratory measurements of fasting blood glucose level. This is not included in the data set available on Kaggle, and was directly downloaded from the NHANES site.
2) Dietary and nutrient information from the dietary survey. Specifically, the data includes i) the types and amounts of food and drinks consumed in the 24-hour period before the interview, and the corresponding energy intake, nutrients and other food components, and ii) diet-related habits, such as salt use in food preparation.
3) Basic demographic information, including age and gender.

The selected data is summarized in Table 1. The selected data columns are extracted from various csv/xpt files and merged based on the participant IDs. The data is relative clean as majority of the data is numerical and ready for modeling. Few are categorical data with "refused" and "don't known" categories, which are converted as NaN before modeling. Not all participants participated or answered all parts of the survey, and those with missing responses or NaN in any of the selected data fields are excluded. The final data set includes survey data from 1973 participants.

Table 1: Selected data from NHANES 2013-2014 data set.

Data group	Data label	Description	Variable type
Lab	LBXGLU	Fasting Glucose (mg/dL)	Numerical
Diet	DR1TKCAL	Energy (kcal)	Numerical
Diet	DR1TPROT	Protein (gm)	Numerical
Diet	DR1TCARB	Carbohydrate (gm)	Numerical
Diet	DR1TSUGR	Total sugars (gm)	Numerical
Diet	DR1TFIBE	Dietary fiber (gm)	Numerical
Diet	DR1TTFAT	Total fat (gm)	Numerical
Diet	DR1TSFAT	Total saturated fatty acids (gm)	Numerical
Diet	DR1TMFAT	Total monounsaturated fatty acids (gm)	Numerical
Diet	DR1TPFAT	Total polyunsaturated fatty acids (gm)	Numerical
Diet	DR1TCHOL	Cholesterol (mg)	Numerical
Diet	DR1TATOC	Vitamin E as alpha-tocopherol (mg)	Numerical
Diet	DR1TATOA	Added alpha-tocopherol Vitamin E (mg)	Numerical
Diet	DR1TRET	Retinol (mcg)	Numerical
Diet	DR1TVARA	Vitamin A - RAE (mcg)	Numerical
Diet	DR1TACAR	Alpha-carotene (mcg)	Numerical
Diet	DR1TBCAR	Beta-carotene (mcg)	Numerical
Diet	DR1TCRYP	Beta-cryptoxanthin (mcg)	Numerical
Diet	DR1TLYCO	Lycopene (mcg)	Numerical
Diet	DR1TLZ	Lutein + zeaxanthin (mcg)	Numerical
Diet	DR1TVB1	Thiamin Vitamin B1 (mg)	Numerical
Diet	DR1TVB2	Riboflavin Vitamin B2 (mg)	Numerical
Diet	DR1TNIAC	Niacin (mg)	Numerical
Diet	DR1TVB6	Vitamin B6 (mg)	Numerical
Diet	DR1TFOLA	Total folate (mcg)	Numerical
Diet	DR1TFA	Folic acid (mcg)	Numerical
Diet	DR1TFF	Food folate (mcg)	Numerical
Diet	DR1TFDFE	Folate DFE (mcg)	Numerical
Diet	DR1TCHL	Total choline (mg)	Numerical
Diet	DR1TVB12	Vitamin B12 (mcg)	Numerical
Diet	DR1TB12A	Added vitamin B12 (mcg)	Numerical
Diet	DR1TVC	Vitamin C (mg)	Numerical
Diet	DR1TVD	Vitamin D - D2 + D3 (mcg)	Numerical
Diet	DR1TVK	Vitamin K (mcg)	Numerical
Diet	DR1TCALC	Calcium (mg)	Numerical
Diet	DR1TPHOS	Phosphorus (mg)	Numerical
Diet	DR1TMAGN	Magnesium (mg)	Numerical
Diet	DR1TIRON	Iron (mg)	Numerical
Diet	DR1TZINC	Zinc (mg)	Numerical
Diet	DR1TCOPP	Copper (mg)	Numerical
Diet	DR1TSODI	Sodium (mg)	Numerical
Diet	DR1TPOTA	Potassium (mg)	Numerical
Diet	DR1TSELE	Selenium (mcg)	Numerical
Diet	DR1TCAFF	Caffeine (mg)	Numerical
Diet	DR1TTHEO	Theobromine (mg)	Numerical
Diet	DR1TALCO	Alcohol (gm)	Numerical
Diet	DR1TMOIS	Moisture (gm)	Numerical
Diet	DR1BWATZ	Total bottled water drank yesterday (gm)	Numerical
Diet	DBQ095Z	Type of table salt used	Categorical
Diet	DBD100	How often add salt to food at table	Categorical
Diet	DRQSPREP	Salt used in preparation	Categorical
Diet	DR1TWS	Tap water source	Categorical
Demographic	RIDAGEYR	Age in years	Numerical
Demographic	RIAGENDR	Gender	Categorical

Exploratory data analysis:

Normal fasting blood glucose is <100 mg/dL, whereas values >100 mg/dL are considered as prediabetes (100-125 mg/dL) or diabetes (>125 mg/dL). In this study, data is binary classified into the high fasting blood glucose group (>=100 mg/dL; N=819) or the normal group (<100 mg/dL; N=1154).

The distribution of the fasting blood glucose data is summarized in Table 2 and is plotted in Figure 1. The mean fasting blood glucose is 104 mg/dL which is slight above the 100 mg/dL As shown in Figure 1, the distribution appears to be normal but with a long tail on the right. When grouped the data by gender, the male's distribution shows a sharp peak at 100 mg/dL, while the female's distribution has a shoulder on the left, indicating more females have normal blood glucose measurements than the males.

Table 2. Statistics of fasting blood glucose measurements

count	1973
mean	104.2
std	30.6
min	51
25%	91
50%	97
75%	105
max	405

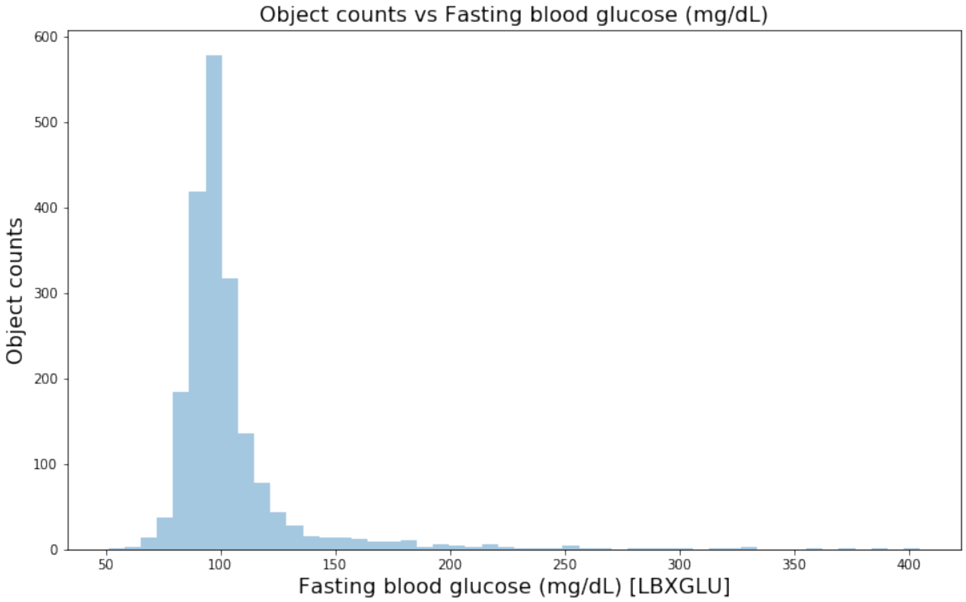


Figure 1. Distribution of fasting blood glucose.

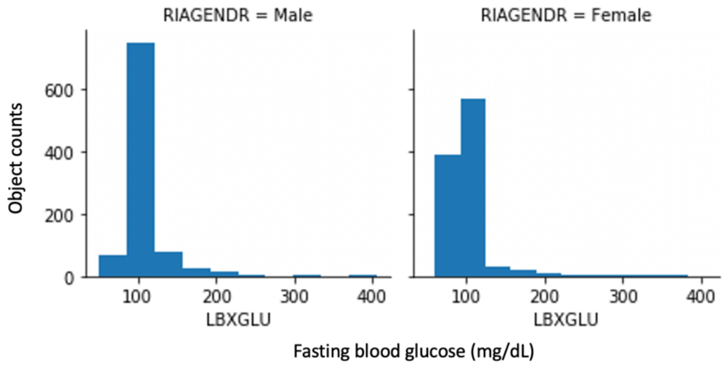


Figure 2. The distributions of fasting blood glucose measurements of males and females.

In Figure 3, fasting blooding glucose values of different age groups are plotted, showing that 1) the participants are evenly distributed across all age groups with a slightly higher population under 20, and 2) all age groups appear to share a similar distribution of fasting blood glucose values, as most measurements are roughly around 100 mg/dL. The detailed boxplot in Figure 4 shows that, as age increases, the mean of fasting blood glucose also increases. The inflection point is around age 40, as the means of fasting blood glucose appear to be above for those above 40 and up.

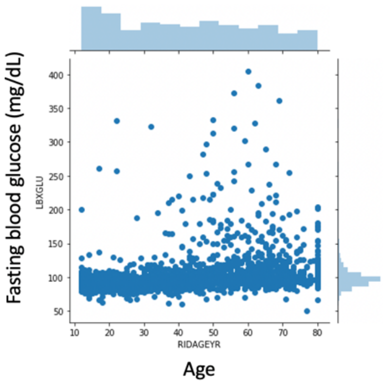


Figure 3. Scatter plot of fasting blood glucose of different age groups.

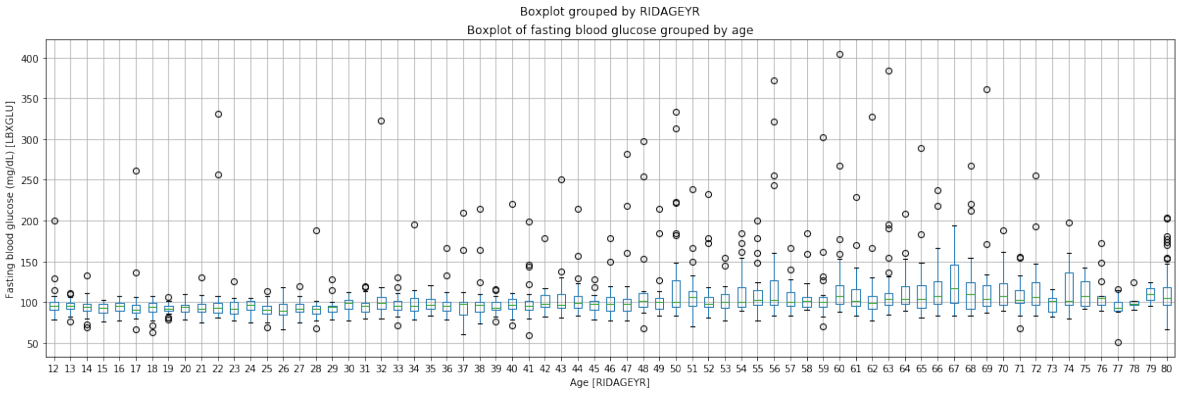


Figure 4. Boxplot of fasting blood glucose measurements grouped by age.

To further examine the relationships between the selected features, the fasting blood glucose measurements and a selected subset of key features, including age, energy intake, and key nutrients consumed, including protein, carbohydrate, sugars and dietary fiber, are plotted against each other in Figure 5. The graph, as colored by the classification mentioned about based on the fasting blood glucose above 100 mg/dL or not, illustrate that there appear to be no clear relationships between most these features. The only exception appears to be age, in which the mean faster blood glucose measurement increases as age increases.

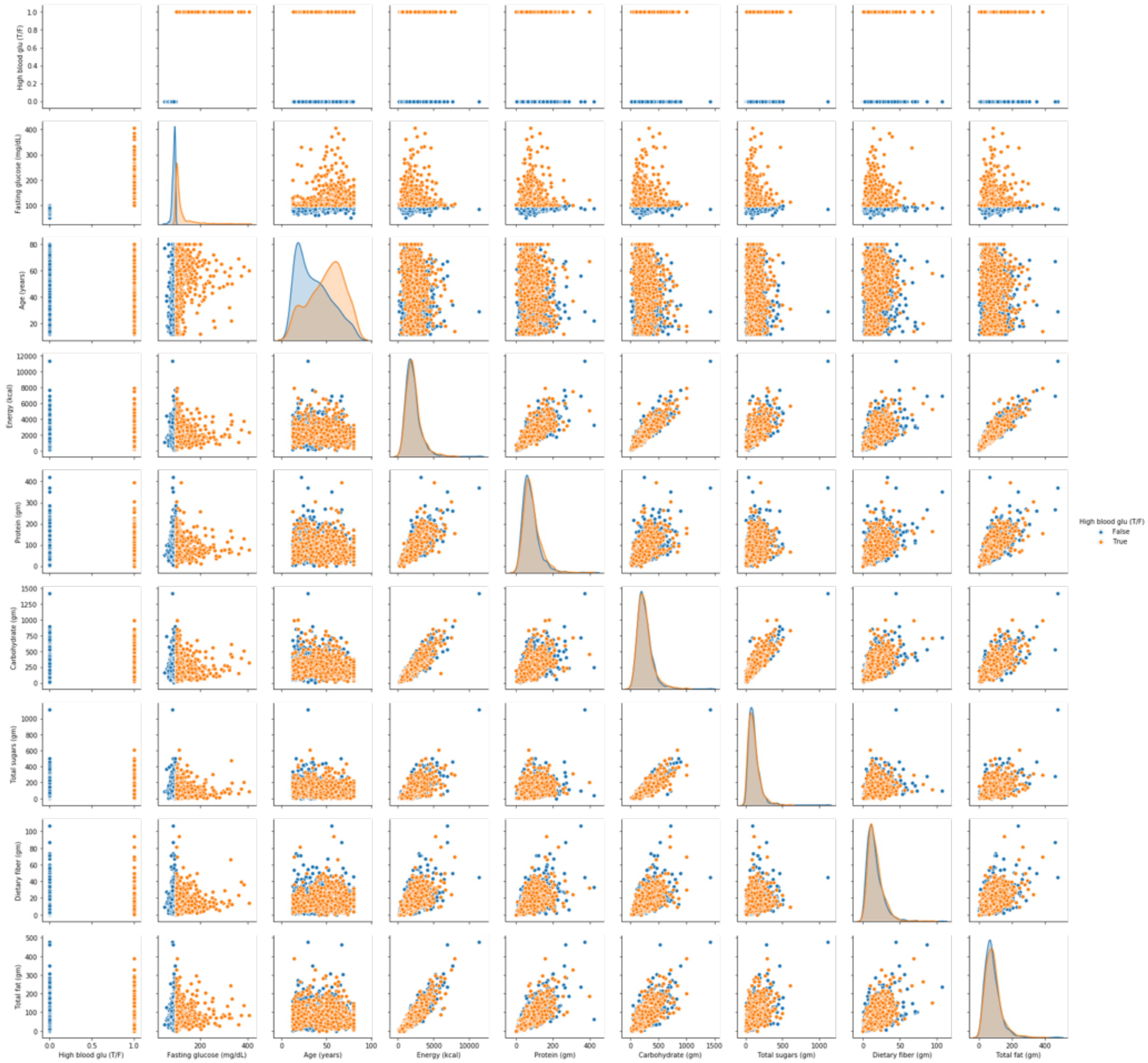
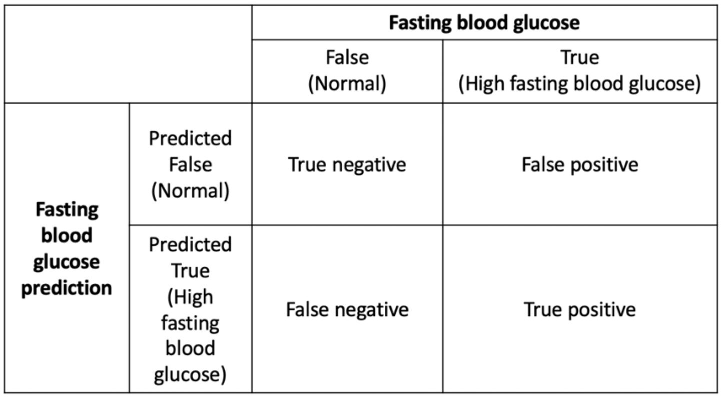


Figure 5. Pair-wise scatter plots of key features.

Classification modeling:

Different machine learning methods from Scikit-learn were employed to develop a classification model for predicting individual's fasting blood glucose level. The utilized methods include random forest, logistic regression, support vector machine, K-nearest neighbors, AdaBoost, gradient tree boosting, and voting classifier. Considering the goal here is to develop a virtual first-pass screening tool to recommend follow-up with physicians, being able to identify as many as true positive (those with high fasting blood glucose being correctly predicted with high fasting blood glucose) would be ideal. False positives (those with normal fasting blood glucose being incorrectly classified as having high fasting blood glucose), while should be minimized, might be acceptable (as long as the false positives are not disproportionally high). In other words, a sensitive model is desired, therefore recall, a measure of the fraction of relevant instances that are correctly predicted, was used as the scoring criterion in model development. The confusion matrices reported are formatted as follow in Figure 6.



		Fasting blood glucose	
		False (Normal)	True (High fasting blood glucose)
Fasting blood glucose prediction	Predicted False (Normal)	True negative	False positive
	Predicted True (High fasting blood glucose)	False negative	True positive

Figure 6. Format of reporting confusion matrix.

The results of different classification methods are summarized in Table 3. Overall, random forest (RF3), support vector machine, and AdaBoost perform best and produce similar recall scores of 0.7 or above for the True class and 0.5 or above for the False class. Gradient tree boosting method gives the highest recall for the True class (high fasting blood glucose) in the test set. However, the recall for the False class (normal fasting blood glucose) is also the lowest (0.09), reflecting the model is indiscriminative in classify most of the test set as the True class.

A series of random forest models were developed to examine features based on feature importance and different sampling methods. As shown by results from RF2 and RF4, re-training the model on features with high importance helps to improve recall for the False class. Interestingly, up-sampling the minority True class (RF6) improves recall for the False class as well, while lowering recall for the True class. Down-sampling the majority class (RF5), on the other hand, shows the opposite effect, i.e. slightly improving recall for the True class relative to RF3 but decreasing recall for the False class.

A voting classifier was developed based on 3 of the best models, including the random forest (RF3), SVM and AdaBoost models. However, the ensemble model did not produce any significant improvement over the individual model.

Table 3. Summary of classification models.

Code	Model	Scaling (0-1)	Number of features	Training: Recall	Test: True Recall	Test: False Recall	Test: Confusion matrix	
RF1	Random forest	No	63	0.65	0.70	0.53	123	108
							49	115
RF2	Random forest	No	20	0.67	0.72	0.61	140	91
	(Top 20 most important feature from RF1)						46	118
RF3	Random forest	Yes	63	0.65	0.71	0.53	122	109
							48	116
RF4	Random forest	Yes	20	0.67	0.74	0.61	140	91
	(Top 20 most important feature from RF3)						42	122
RF5	Random forest	Yes	63	0.74	0.75	0.41	95	136
	(random down-sampling of majority class)						41	123
RF6	Random forest	Yes	63	0.74	0.46	0.77	179	52
	(random up-sampling of minority class)						88	76
LG	Logistic regression	Yes	63	0.51	0.57	0.75	174	57
							71	93
SVM	Support vector machines	Yes	63	0.68	0.79	0.52	121	110
							35	129
KNN	K-nearest neighbors	Yes	63	0.50	0.46	0.61	142	89
							88	76
ADA	AdaBoost	Yes	63	0.63	0.70	0.61	142	89
							49	115
GTB	Gradient tree boosting	Yes	63	0.55	0.88	0.09	21	210
							19	145
Voting	Voting classifier	Yes	63	0.62	0.68	0.65	151	80
	(Random forest (RF3) + SVM + AdaBoost)						53	111

Deep learning:

Neural network models were also developed using Keras and TensorFlow. Overall, relative to the classification models above, the neural network models did not provide better performance, at least for the True class. The recall scores for the True class in the testing set range from 0.45 to 0.66, while the recall scores for the False class are >0.6. The manual optimization by increasing the number of hidden layers and number of epochs did not appear to be very productive, suggesting more thorough and systematic optimization might be needed.

Table 4. Summary of neural network models.

Code	Model	#input neurons	#Epochs	Test: True Recall	Test: False Recall	Test: Confusion matrix	
NN1	1-layer neural network	63	50	0.66	0.64	147	84
						56	108
NN2	1-layer neural network	32	50	0.51	0.75	174	57
						80	84
NN3	3-layer neural network	63	50	0.49	0.63	145	86
	(2 hidden layers with 100 neurons; relu)					84	80
NN4	5-layer neural network	63	50	0.45	0.62	144	87
	(4 hidden layers with 100 neurons; relu)					90	74
NN5	5-layer neutral network	63	100	0.56	0.62	144	87
	(4 hidden layers with 100 neurons; relu)					72	92

Conclusions:

Various classification models were developed to predict fasting blood glucose level from dietary data for type 2 diabetes screening. The best algorithms, including random forest, SVM and AdaBoost, have produced models with recall scores above 0.7 for the high fasting blood glucose group and recall scores above 0.5 for the normal group in the test set. Considering diet is only one of many contributing factors for diabetes, the current models can likely be improved if additional features, such as survey data on lifestyle and medical examination, are considered. In addition, only the survey from 2013 to 2014 was used in this study, and more survey data is available from NHANES, including surveys since 1999. The current data set can be easily expanded to develop the next version of the model.