PREDICTING DIABETIC RISK

MODELING FASTING BLOOD GLUCOSE LEVEL FROM DIETARY DATA

SIEGFRIED LEUNG

HOW DO WE DIAGNOSE TYPE II DIABETES?

- Checking fasting blood glucose or A1c by blood test

HOW DO WE VIRTUAL SCREEN?

Concept:

- As a first pass, can we predict those with high fasting blood glucose?
- Can we do so from one's diet?

Goal:

- Develop a predictive classification model for fasting blood glucose from dietary data

- This can be a non-invasive tool to help alert those affected to follow up with physicians
- Or help doctors identify potential diabetes patients early on

DATA FOR THIS STUDY

Organized data set on Kaggle:
https://www.cdc.gov/nchs/nhanes/index.htm

National Center for Health Statistics

National Health and Nutrition Examination Survey

Overview

Centers for Disease Control and Prevention
National Center for Health Statistics

- The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States.
- The survey is unique in that it combines interviews and physical examinations:
  - ❖ The interview includes demographic, socioeconomic, dietary, and health-related questions
  - ❖ The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests
  - ❖ The survey examines a nationally representative sample of about 5,000 persons each year.
- NHANES is a major program of the National Center for Health Statistics (NCHS). NCHS is part of the Centers for Disease Control and Prevention (CDC) and is responsible for producing vital and health statistics for the nation

DATA FOR THIS STUDY

National Health and Nutrition Examination Survey

NHANES 2013-2014

- Demographics Data
- Dietary Data
- Examination Data
- Laboratory Data
- Questionnaire Data
- Limited Access Data

Basic demographic data
(Age and gender)

Dietary and nutrient information from the dietary survey

i) 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
ii) Diet-related habits, such as salt use in food preparation.

Laboratory measurements of fasting blood glucose

Official: https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2013
Kaggle (organized): https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey

SELECTED DATA (1/2)

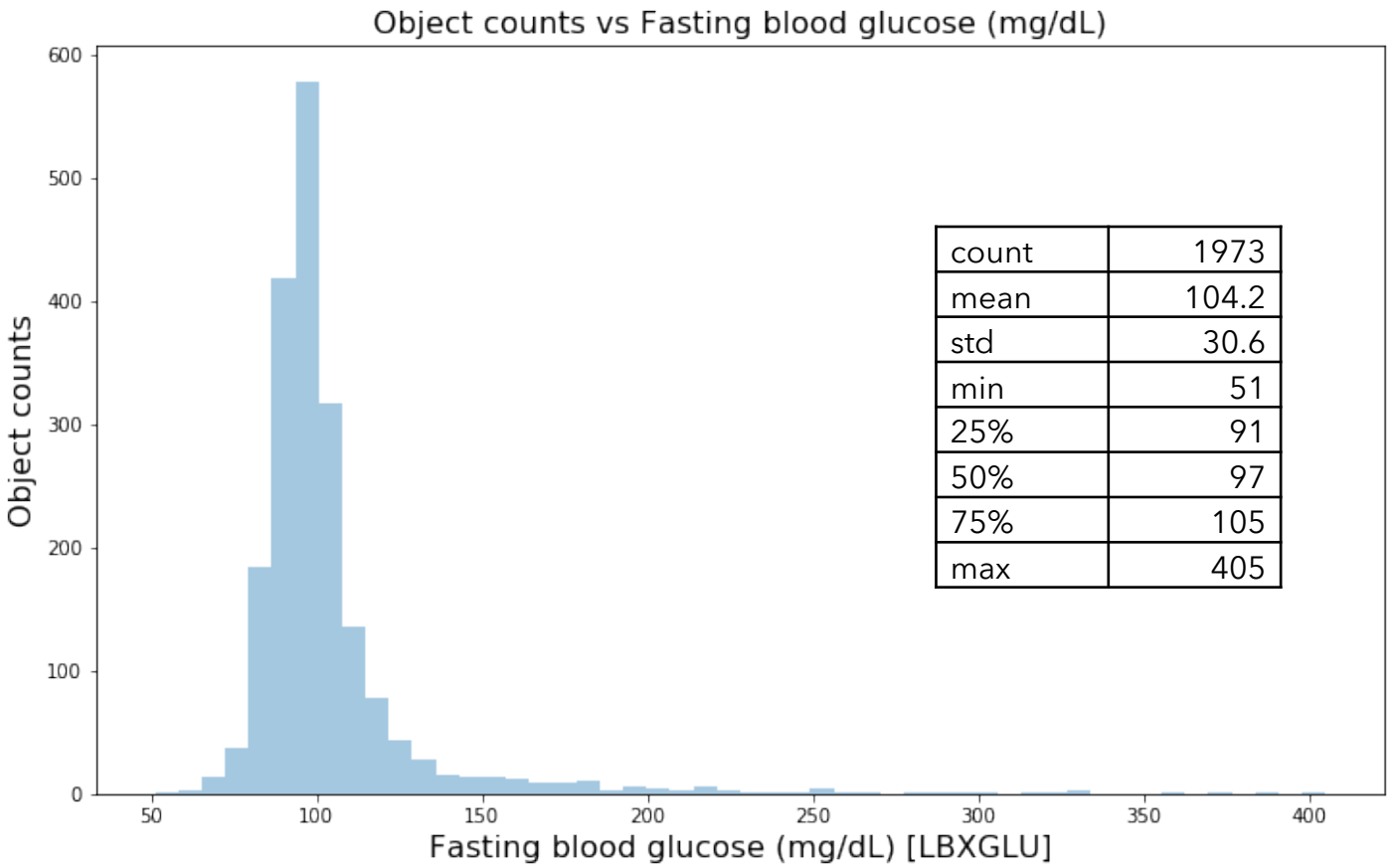
Data group	Data label	Description	Variable type
Lab	LBXGLU	Fasting Glucose (mg/dL)	Numerical
Diet	DR1TKCAL	Energy (kcal)	Numerical
Diet	DR1TPROT	Protein (gm)	Numerical
Diet	DR1TCARB	Carbohydrate (gm)	Numerical
Diet	DR1TSUGR	Total sugars (gm)	Numerical
Diet	DR1TFIBE	Dietary fiber (gm)	Numerical
Diet	DR1TTFAT	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

SELECTED DATA (2/2)

Data group	Data label	Description	Variable type
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
Demography	RIDAGEYR	Age in years	Numerical
Demography	RIAGENDR	Gender	Categorical

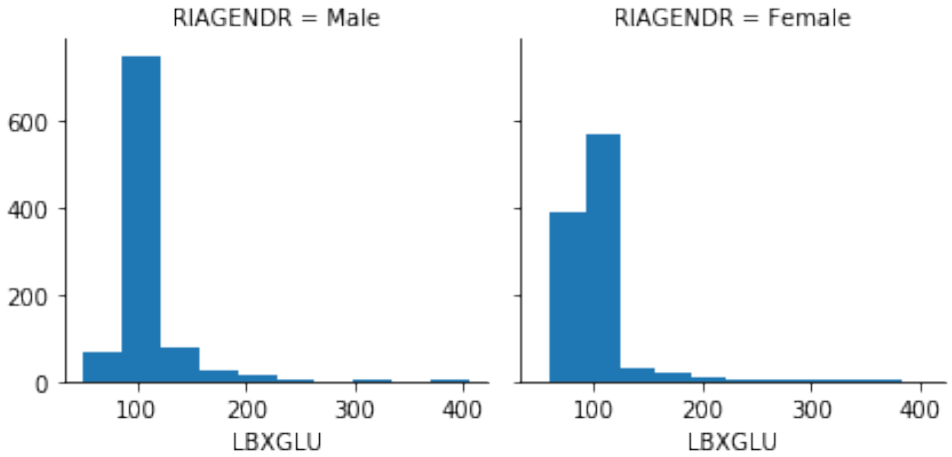
EXPLORATORY DATA ANALYSIS

FASTING BLOOD GLUCOSE



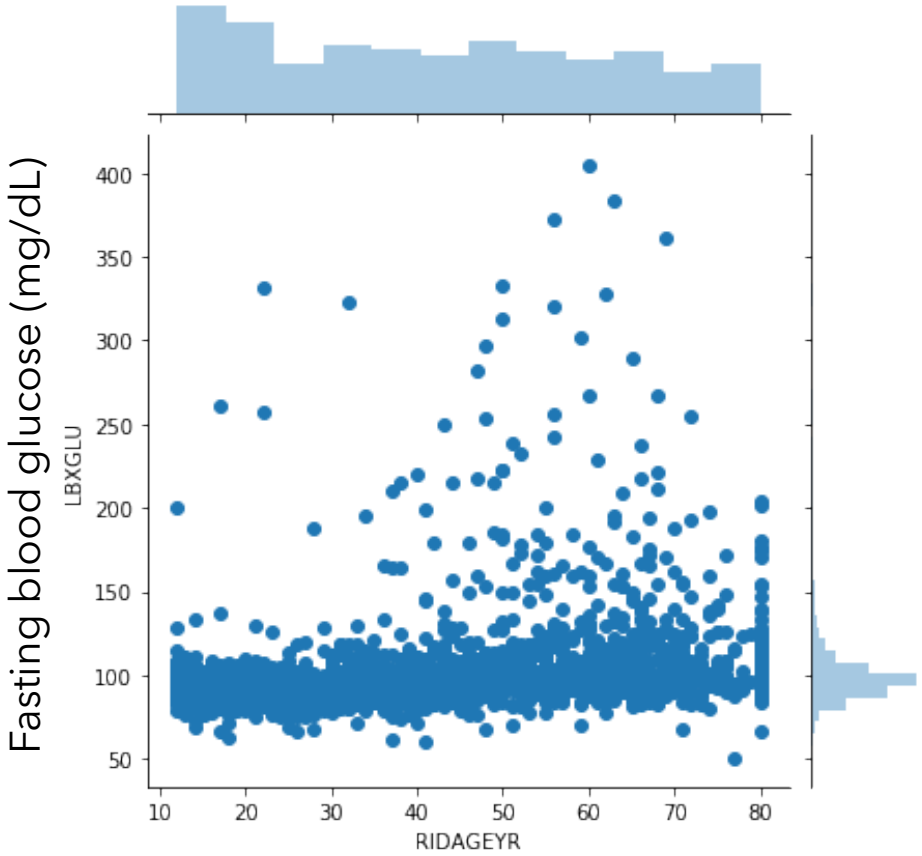
count	1973
mean	104.2
std	30.6
min	51
25%	91
50%	97
75%	105
max	405

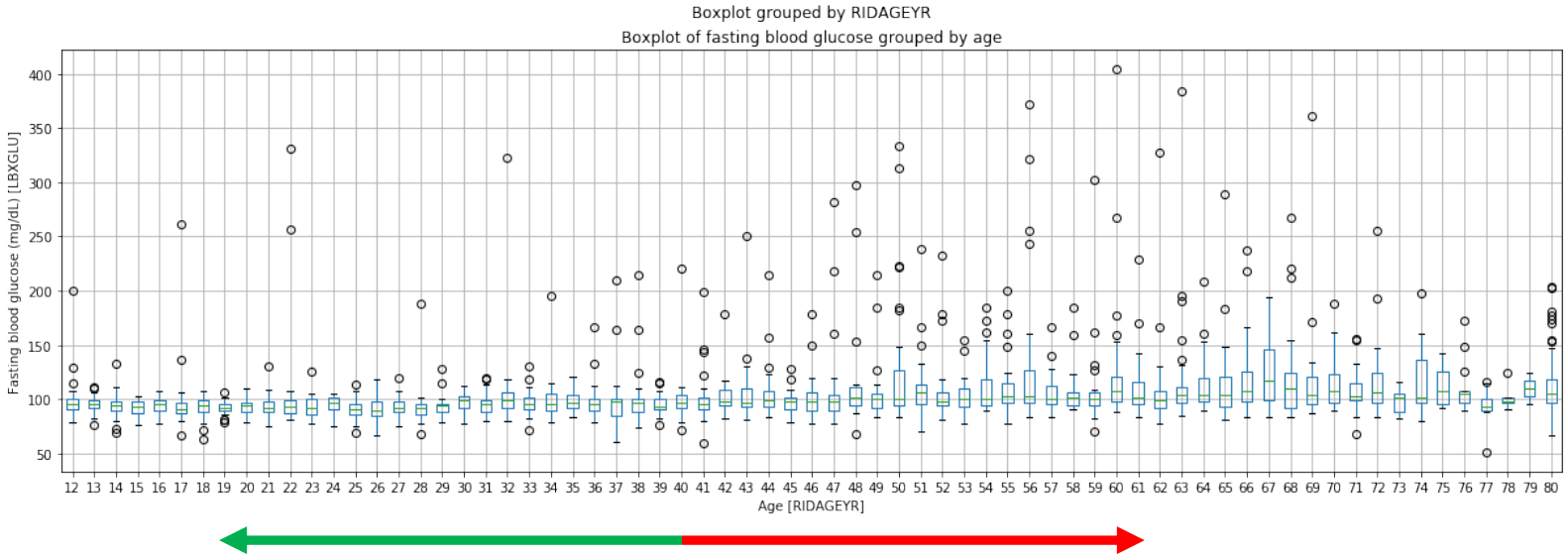
Group	Fasting blood glucose (mg/dL)	High fasting blood glucose Classification
Normal	<100	No
Pre-diabetes	100-125	Yes
Diabetes	>126	



EXPLORATORY DATA ANALYSIS

FASTING BLOOD GLUCOSE VS. AGE





Mean < 100

Mean > 100

EXPLORATORY DATA ANALYSIS

KEY FEATURES

MACHINE LEARNING: CLASSIFICATION MODEL

- Random Forest
- Logistic regression
- Support vector machine
- K-nearest neighbors
- AdaBoost
- Gradient tree boosting
- Voting classifier

- Scoring: Recall
- As a

		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

MACHINE LEARNING: CLASSIFICATION MODELS

- Random Forest
- Logistic regression
- Support vector machine
- K-nearest neighbors
- AdaBoost
- Gradient tree boosting
- Voting classifier

(Scikit-learn)

- Scoring parameter: $Recall = \frac{tp}{tp+fn}$
- Aim: as a screening tool, catch as many positive as possible, as long as not too many false positive

		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

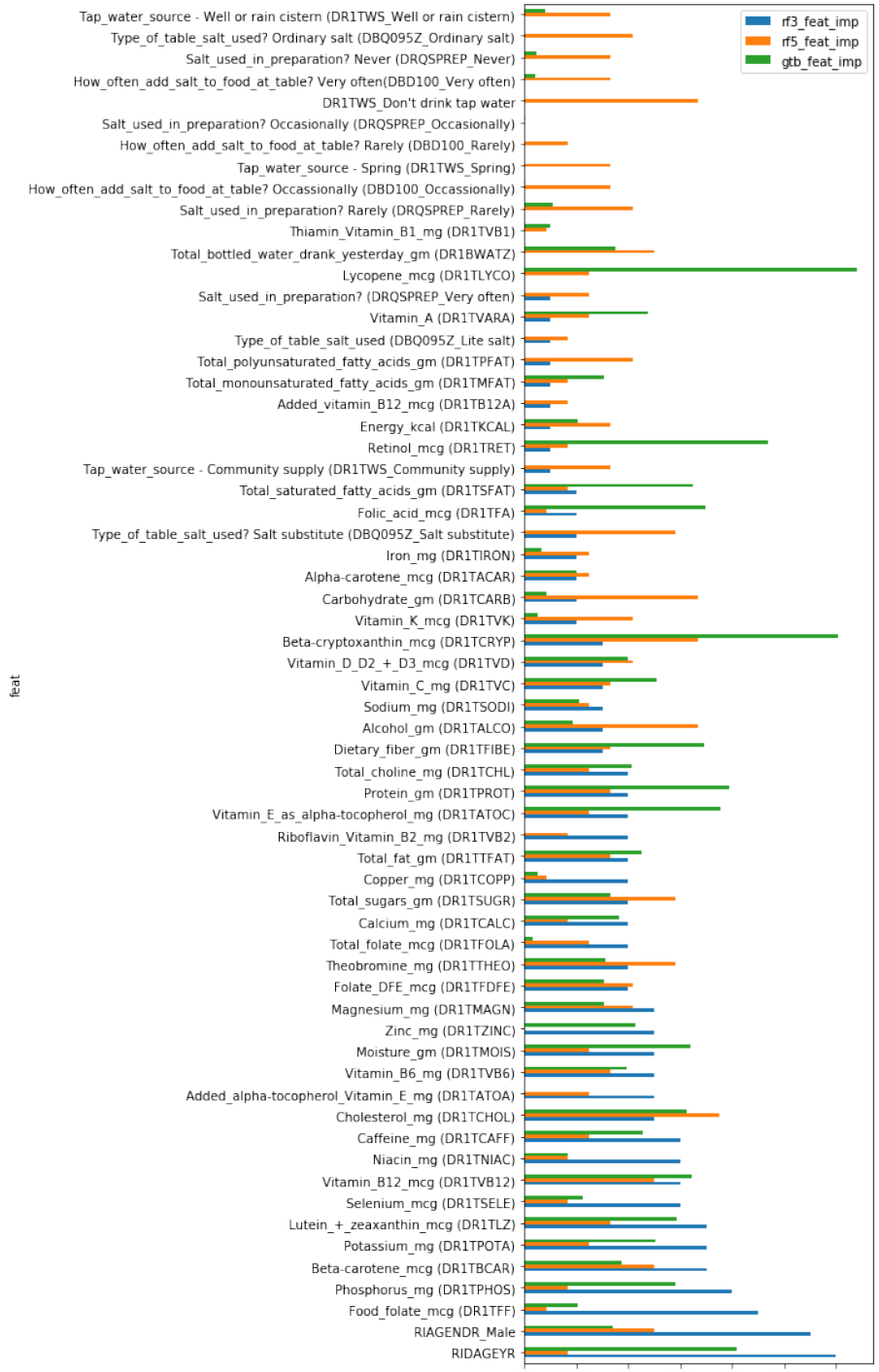
ML: CLASSIFICATION MODELS

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

ML: 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 (Top 20 most important feature from RF1)	No	20	0.67	0.72	0.61	140	91
							46	118
RF3	Random forest	Yes	63	0.65	0.71	0.53	122	109
							48	116
RF4	Random forest (Top 20 most important feature from RF3)	Yes	20	0.67	0.74	0.61	140	91
							42	122
RF5	Random forest (random down-sampling of majority class)	Yes	63	0.74	0.75	0.41	95	136
							41	123
RF6	Random forest (random up-sampling of minority class)	Yes	63	0.74	0.46	0.77	179	52
							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 (Random forest (RF3) + SVM + AdaBoost)	Yes	63	0.62	0.68	0.65	151	80
							53	111

FEATURE IMPORTANCE



IMPORTANT FEATURES

DEEP LEARNING: 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

Keras/TensorFlow

NEXT

Beyond diet, whare are we missing?

- Expanding features from NHANES, such as lifestyle, socioeconomic, health, medication data, etc.

More data!

- Expand data set with NHANES data from 1999 to present

SUMMARY

- Developed classification models to predict high fasting blood glucose level based on one's diet/age/gender – a potential virtual screening tool for type 2 diabetes
- Best models (Recall >0.7):
    - Random forest
    - SVM
    - AdaBoost