Diabetes prediction

Naya College
March 2022

Partners:

Alon Ben Bassat – alonb4040@gmail.com

Liran Kesler – liran249@gmail.com

Hagai Alon – haguy838@gmail.com

Content

Goals

Domain info – What is it diabetes mellitus?

Dataset info

Challenges we had to deal with

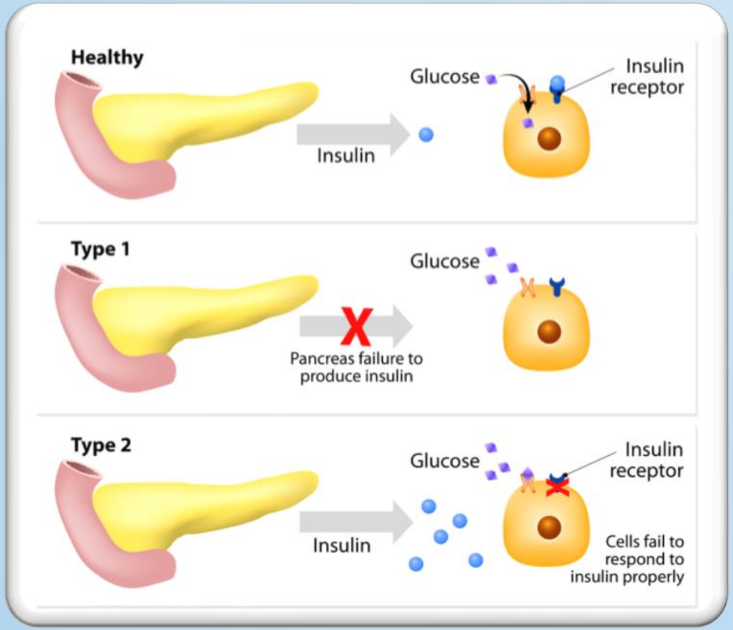
Cleaning data and removing outliers

Feature engineering

Models

Conclusions

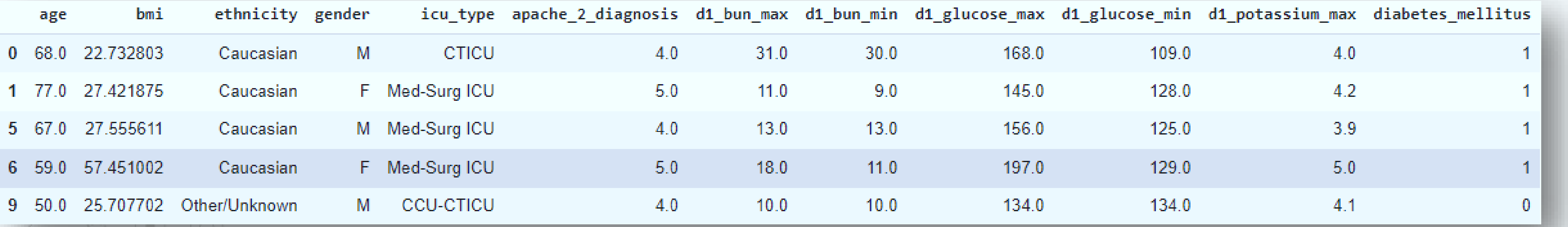
What is it diabetes mellitus ?



Dataset info

- In this project we are going to use a dataset from Kaggle: “WiDS Datathon 2021”.
- The dataset size: (130157 ,180) - each row represents a patient.

Following are few of the most important columns for our analysis:

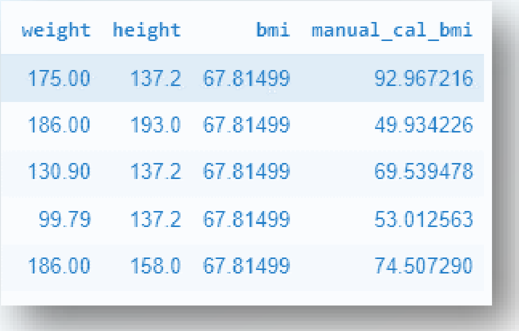


	age	bmi	ethnicity	gender	icu_type	apache_2_diagnosis	d1_bun_max	d1_bun_min	d1_glucose_max	d1_glucose_min	d1_potassium_max	diabetes_mellitus
0	68.0	22.732803	Caucasian	M	CTICU	4.0	31.0	30.0	168.0	109.0	4.0	1
1	77.0	27.421875	Caucasian	F	Med-Surg ICU	5.0	11.0	9.0	145.0	128.0	4.2	1
5	67.0	27.555611	Caucasian	M	Med-Surg ICU	4.0	13.0	13.0	156.0	125.0	3.9	1
6	59.0	57.451002	Caucasian	F	Med-Surg ICU	5.0	18.0	11.0	197.0	129.0	5.0	1
9	50.0	25.707702	Other/Unknown	M	CCU-CTICU	4.0	10.0	10.0	134.0	134.0	4.1	0

Challenges we had to deal with:

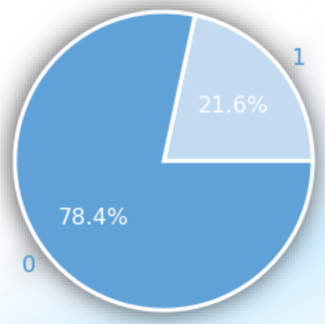
Domain understanding

Mistakes



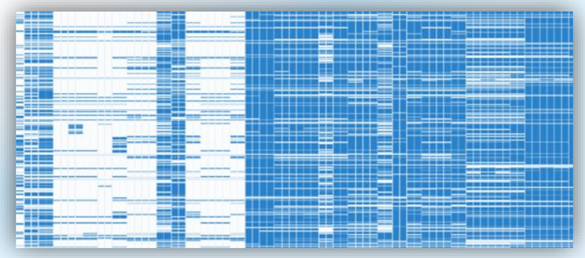
weight	height	bmi	manual_cal_bmi
175.00	137.2	67.81499	92.967216
186.00	193.0	67.81499	49.934226
130.90	137.2	67.81499	69.539478
99.79	137.2	67.81499	53.012563
186.00	158.0	67.81499	74.507290

Imbalanced data



Unlabeled test data

Too many N\A values



Cleaning data, removing outliers and fixing mistakes:

Removed:

- Age = 0
- Threshold of dropping columns - 20% N\A values.
- Columns with no \ negative effect on our models.
- Rows with N\A values – about 20% of the data.

Fixed:

- Creating ‘BMI’ column.
- Exchanging min\max values.
- Imputed N\A values in the ethnicity with “unknown” values.

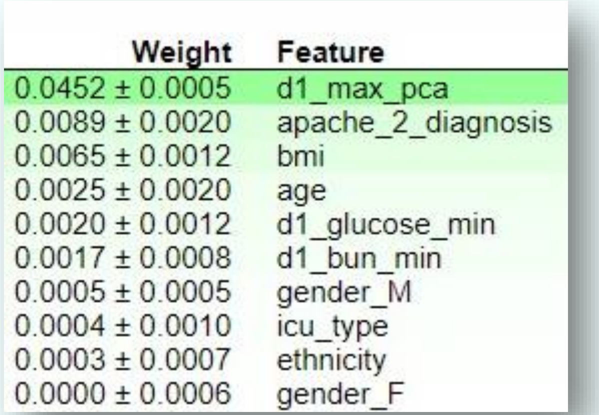
Note:

After these steps we’ve checked that we keep about the same proportion of imbalanced data.

Feature engineering

We've tried many calculated columns, some were designed for specific models. Following are the final calculated columns that we used in our models:

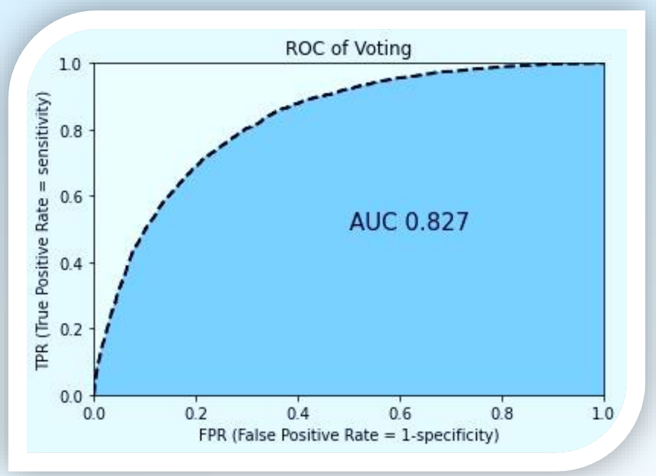
- Grouping and Mapping based on similar diabetes proportion (APACHE_2_diagnosis, ICU_type, Ethnicity)
- PCA on max measures of blood test.
- Select K-best on measures of blood test



Weight	Feature
0.0452 ± 0.0005	d1_max_pca
0.0089 ± 0.0020	apache_2_diagnosis
0.0065 ± 0.0012	bmi
0.0025 ± 0.0020	age
0.0020 ± 0.0012	d1_glucose_min
0.0017 ± 0.0008	d1_bun_min
0.0005 ± 0.0005	gender_M
0.0004 ± 0.0010	icu_type
0.0003 ± 0.0007	ethnicity
0.0000 ± 0.0006	gender_F

Models:

Model	Hyper_parameters	AUC
Random forest	{criterion='entropy', max_depth=10, n_estimators=50}	0.82
Xgboost	{max_depth=10, learning_rate=0.01, n_estimators=200}	0.82
Lgboost	{max_depth=10, learning_rate=0.01, n_estimators=200}	0.83
Adaboost	{learning_rate=0.015, n_estimators=250}	0.82
Logistic regression	{C=100}	0.81
Voting - few models	RND,XGB,LGBM	0.83



The end