Applications of Machine Learning in Cancer, Blood Pressure and Diabetes Prediction

Aziz Koyuncu

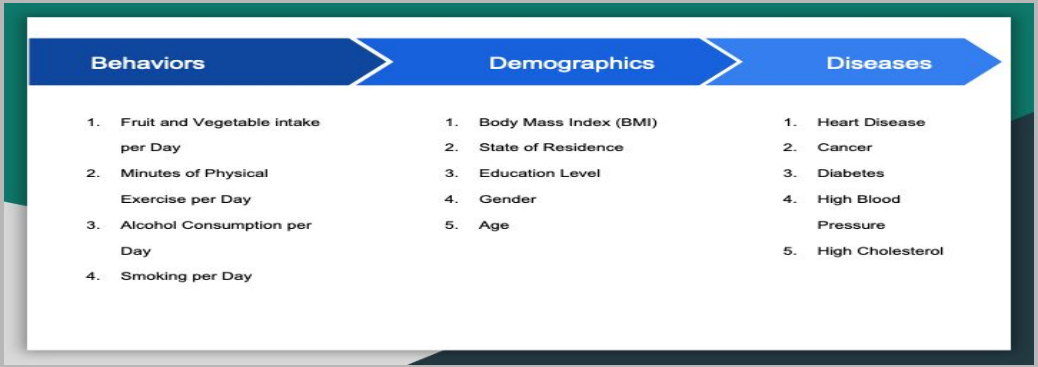
Purpose

The purpose of this project is to utilize ML algorithms to predict cancer, blood pressure, and diabetes based on behaviors, demographics and diseases.

Data (csv)

CDC

CENTERS FOR DISEASE CONTROL AND PREVENTION



Behaviors

1. Fruit and Vegetable intake per Day
2. Minutes of Physical Exercise per Day
3. Alcohol Consumption per Day
4. Smoking per Day

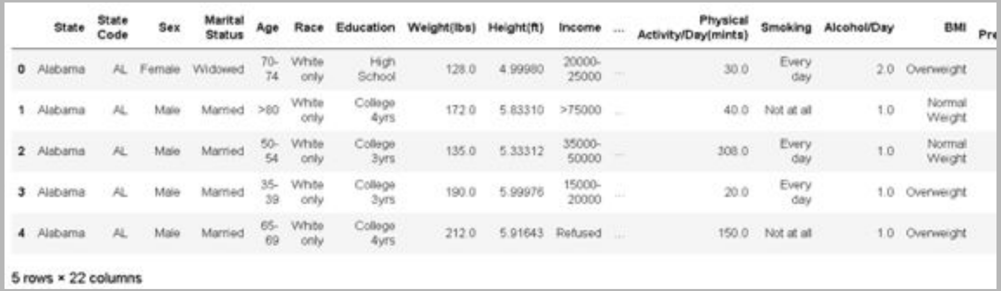
Demographics

1. Body Mass Index (BMI)
2. State of Residence
3. Education Level
4. Gender
5. Age

Diseases

1. Heart Disease
2. Cancer
3. Diabetes
4. High Blood Pressure
5. High Cholesterol

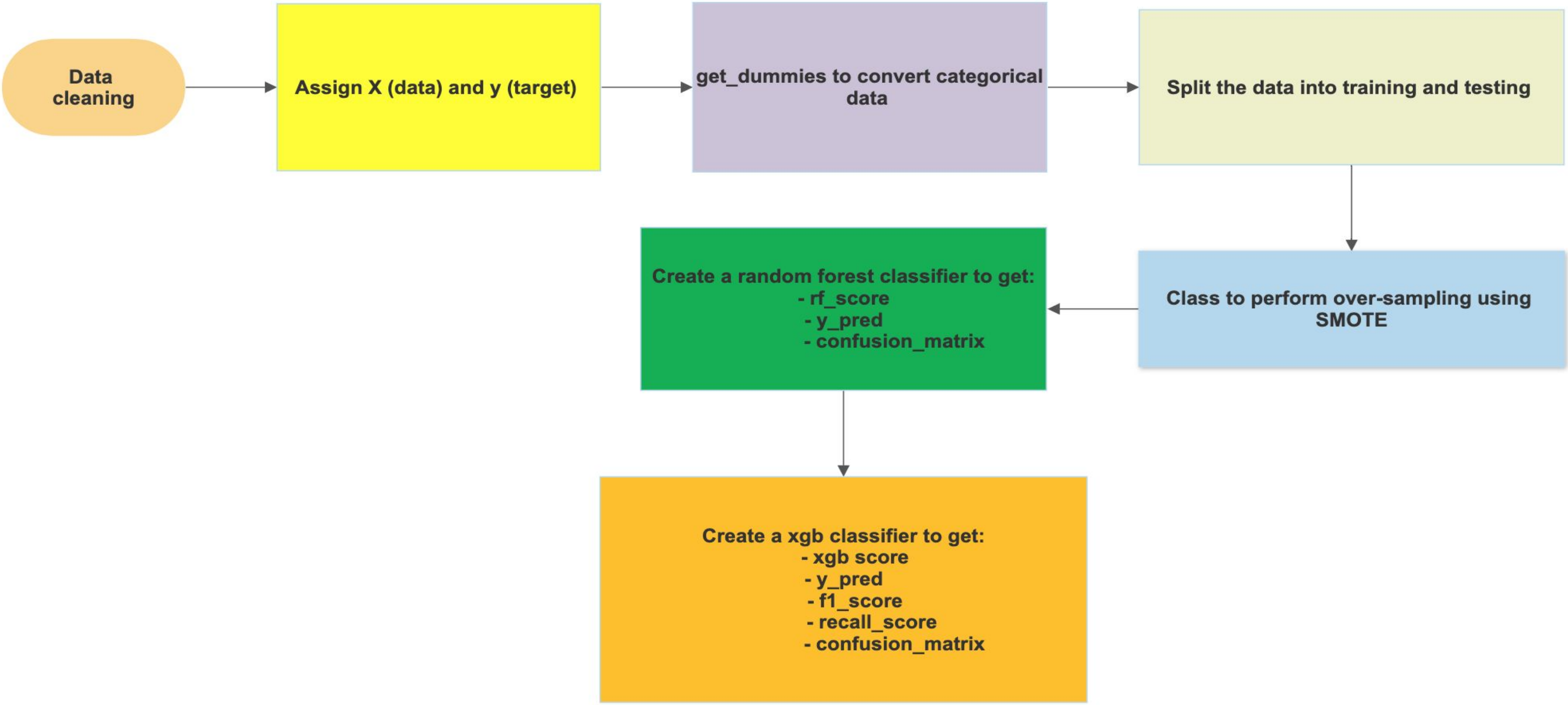
kaggle



	State	State Code	Sex	Marital Status	Age	Race	Education	Weight(lbs)	Height(ft)	Income	...	Physical Activity/Day(mints)	Smoking	Alcohol/Day	BMI	Pre
0	Alabama	AL	Female	Widowed	70-74	White only	High School	128.0	4.99980	20000-25000	...	30.0	Every day	2.0	Overweight	
1	Alabama	AL	Male	Married	>80	White only	College 4yrs	172.0	5.83310	>75000	...	40.0	Not at all	1.0	Normal Weight	
2	Alabama	AL	Male	Married	50-54	White only	College 3yrs	135.0	5.33312	35000-50000	...	308.0	Every day	1.0	Normal Weight	
3	Alabama	AL	Male	Married	35-39	White only	College 3yrs	190.0	5.99976	15000-20000	...	20.0	Every day	1.0	Overweight	
4	Alabama	AL	Male	Married	65-69	White only	College 4yrs	212.0	5.91643	Refused	...	150.0	Not at all	1.0	Overweight	

5 rows × 22 columns

Methodology



Data cleaning

Assign X (data) and y (target)

get_dummies to convert categorical data

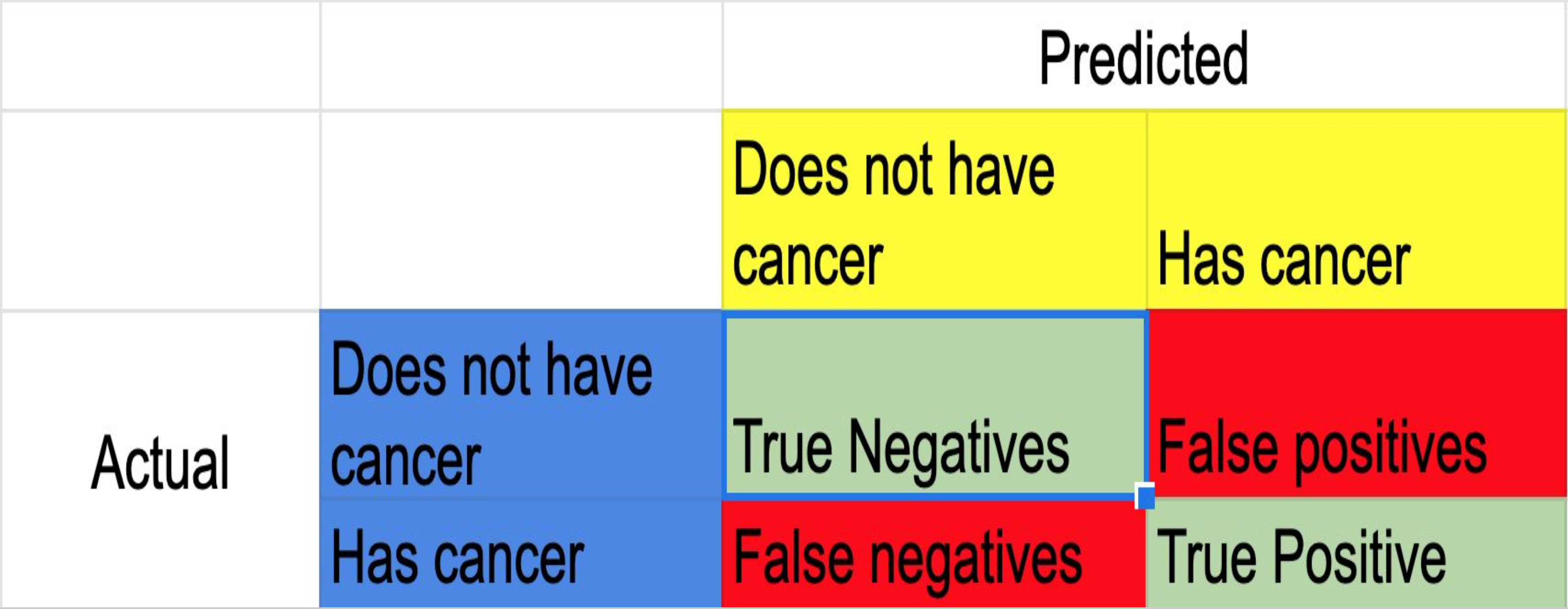
Split the data into training and testing

Class to perform over-sampling using SMOTE

Create a random forest classifier to get:
- rf_score
- y_pred
- confusion_matrix

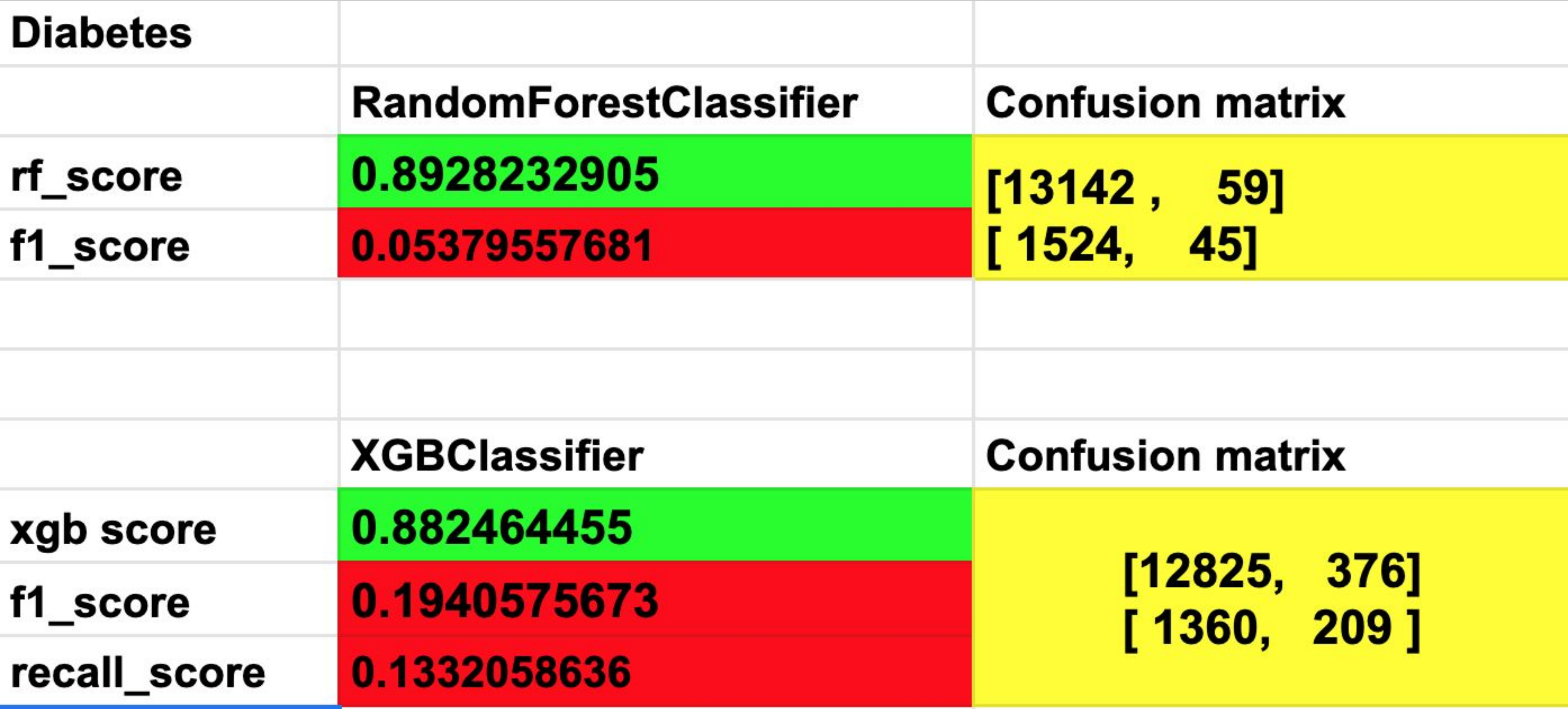
Create a xgb classifier to get:
- xgb score
- y_pred
- f1_score
- recall_score
- confusion_matrix

Confusion Matrix



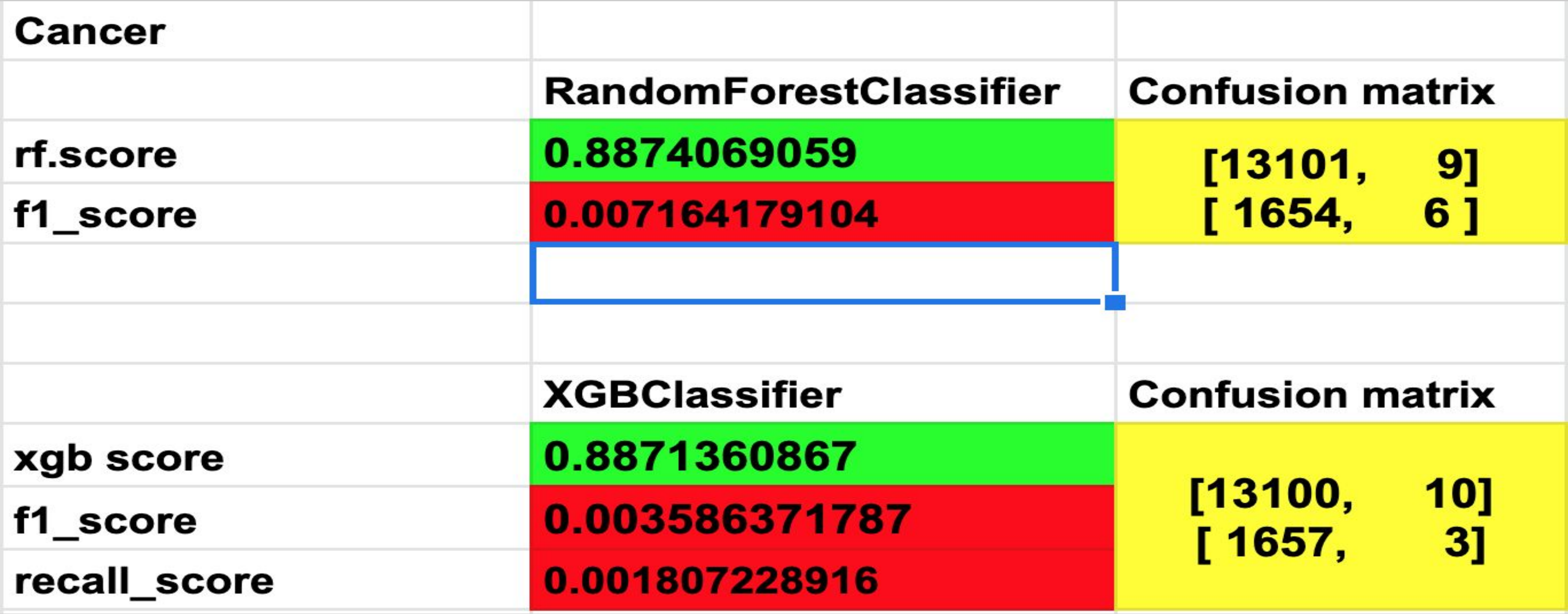
		Predicted	
		Does not have cancer	Has cancer
Actual	Does not have cancer	True Negatives	False positives
	Has cancer	False negatives	True Positive

Predicting diabetes



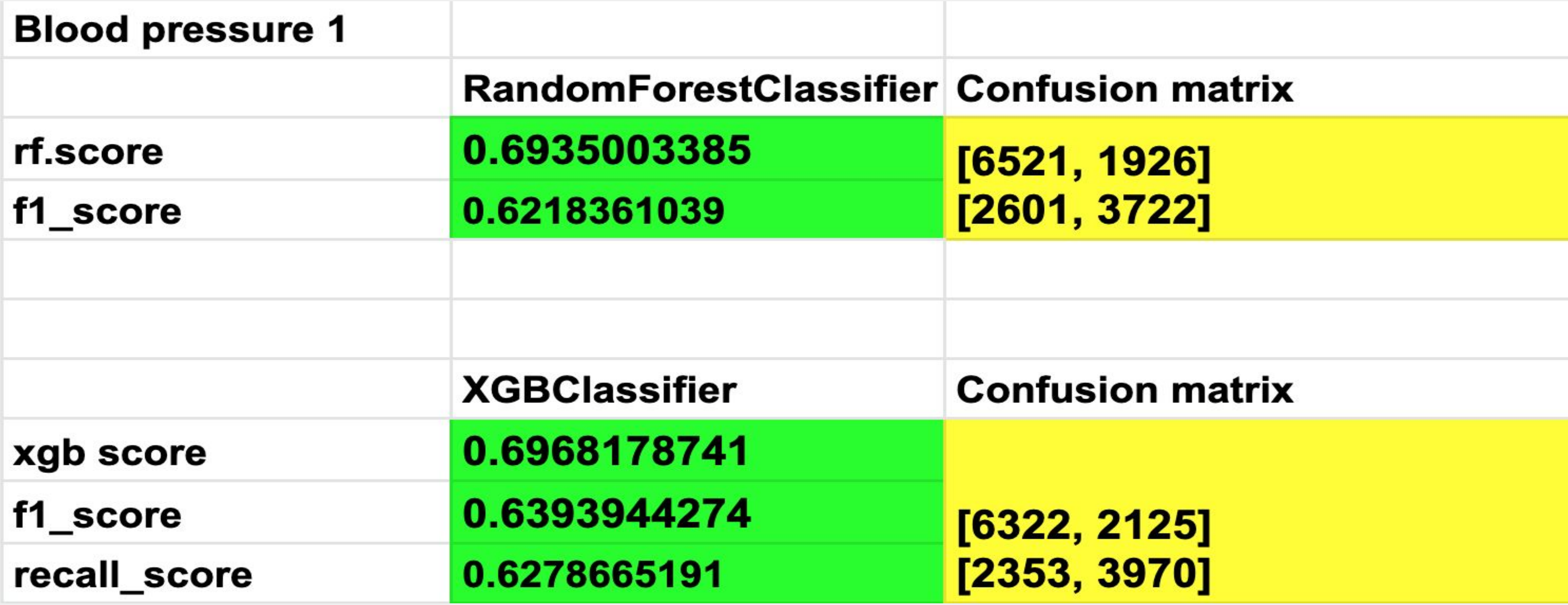
Diabetes		
	RandomForestClassifier	Confusion matrix
rf_score	0.8928232905	[13142 , 59] [1524, 45]
f1_score	0.05379557681	
	XGBClassifier	Confusion matrix
xgb score	0.882464455	[12825, 376] [1360, 209]
f1_score	0.1940575673	
recall_score	0.1332058636	

Predicting cancer



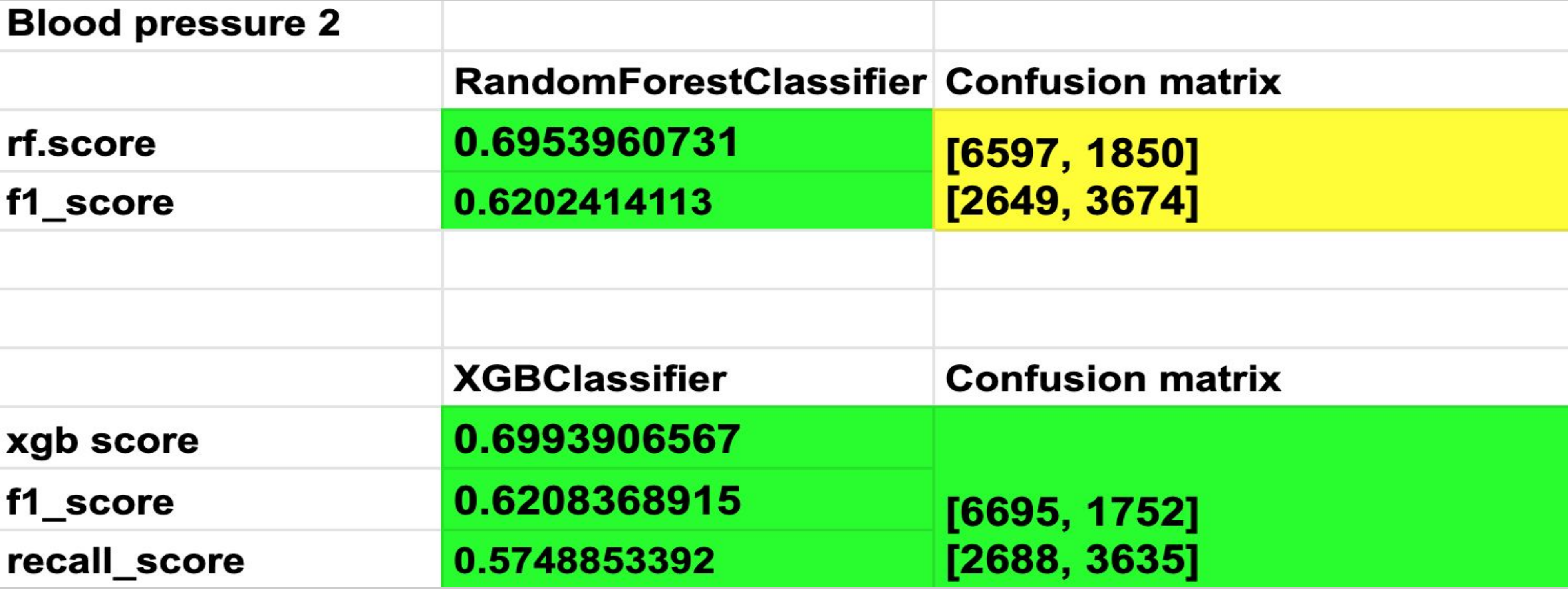
Cancer		
	RandomForestClassifier	Confusion matrix
rf.score	0.8874069059	[13101, 9] [1654, 6]
f1_score	0.007164179104	
	XGBClassifier	Confusion matrix
xgb score	0.8871360867	[13100, 10] [1657, 3]
f1_score	0.003586371787	
recall_score	0.001807228916	

Predicting blood pressure 1



Blood pressure 1		
	RandomForestClassifier	Confusion matrix
rf.score	0.6935003385	[6521, 1926] [2601, 3722]
f1_score	0.6218361039	
	XGBClassifier	Confusion matrix
xgb score	0.6968178741	[6322, 2125] [2353, 3970]
f1_score	0.6393944274	
recall_score	0.6278665191	

Predicting blood pressure 2



Blood pressure 2		
	RandomForestClassifier	Confusion matrix
rf.score	0.6953960731	[6597, 1850] [2649, 3674]
f1_score	0.6202414113	
	XGBClassifier	Confusion matrix
xgb score	0.6993906567	[6695, 1752] [2688, 3635]
f1_score	0.6208368915	
recall_score	0.5748853392	

Tools used

=> Random Forest (The **random forest** is a classification algorithm consisting of many decisions trees)

Gradient Boosting algorithms:

XGBoost

https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

Highly imbalanced, it was an issue to predict. (Over sampling). Prototype selection since I already had my data. Cancer free versus cancer.

A **confusion matrix** is a table that is often used to describe the performance of a classification model (or “classifier”)