DIABETES DETECTION

Thanks to Springboard mentor – DIPANJAN SARKAR, Data Science Lead – Google, Author

PROBLEM

DIABETIC PATIENTS

According to International Diabetes Foundation, 537 million adults are living with diabetes. And projection is 643 million by 2030

EARLY IDENTIFICATION

Blood test can identify pre-diabetic and diabetic cases. But no preventive method to identify potential candidates

USABILITY

Undiagnosed cases are very often especially in low-and-middle income countries

SOLUTION

EARY IDENTIFICATION

With the genetic and physical measurements, calculate the probability to become diabetic

GLOBAL REACH

Internet based solution will increase the reach

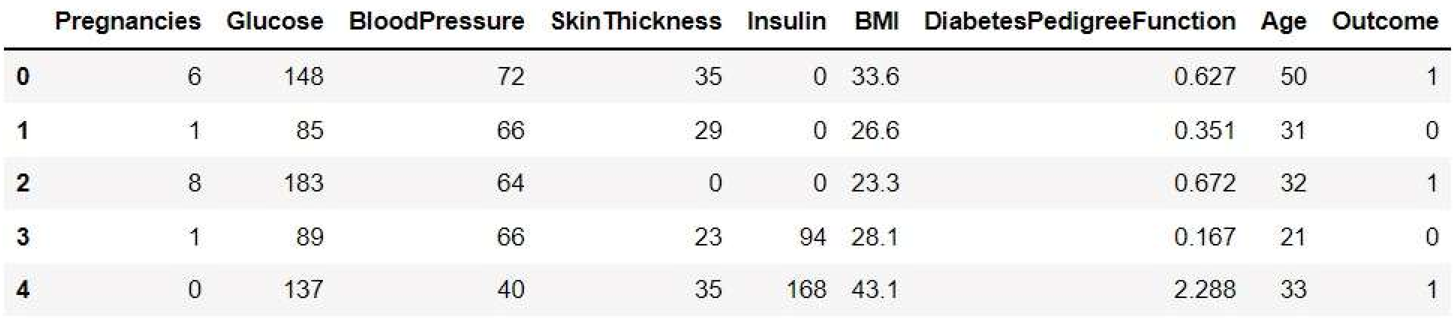
COST SAVINGS

Proactive life-style planning could prevent diabetics

EASY TO USE

Conveniently use from home with an easy-to-use interface

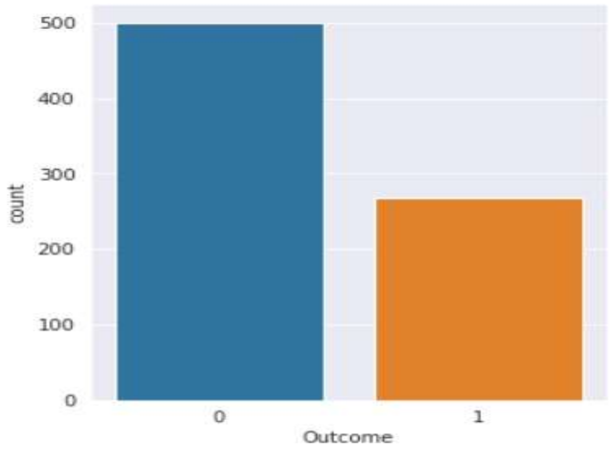
DATA WRANGLING

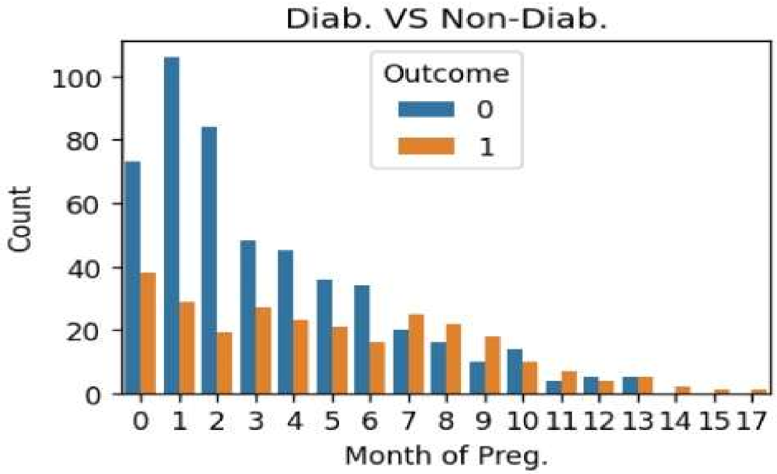


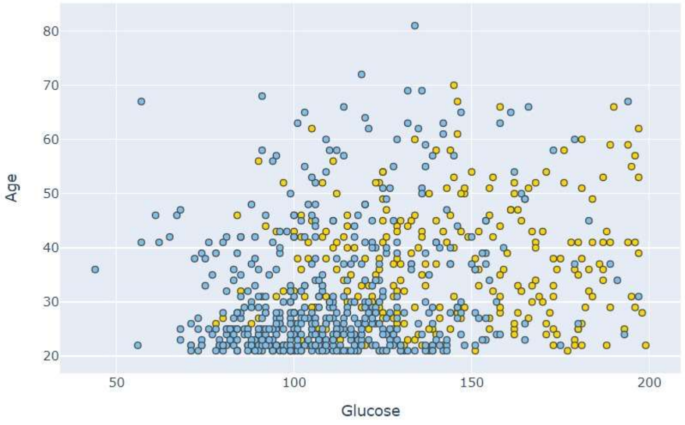
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

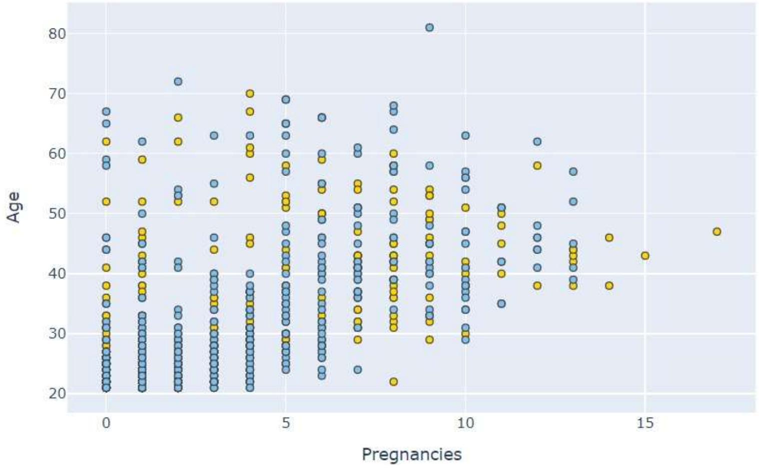
- 769 samples of diabetic and healthy individuals
- All patients are females with minimum of 21 years of age
- This dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases

EXPLORATORY DATA ANALYSIS







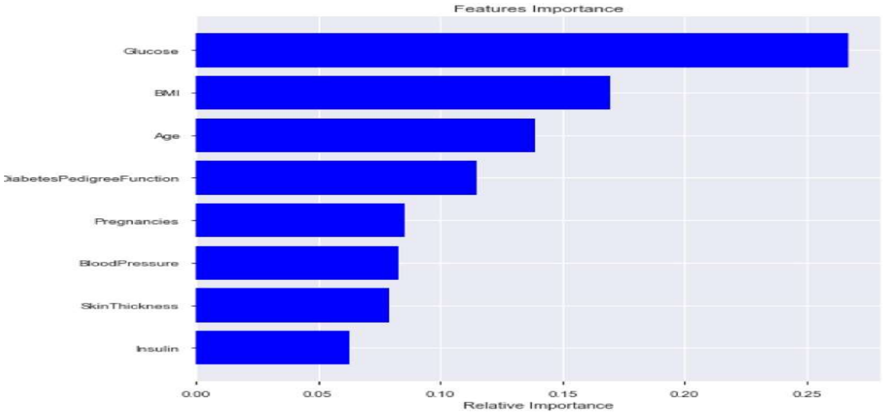


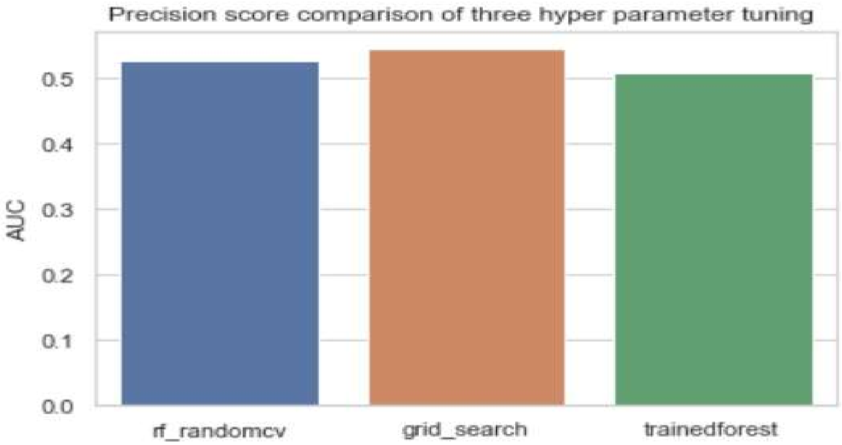
Company A
Product is more expensive

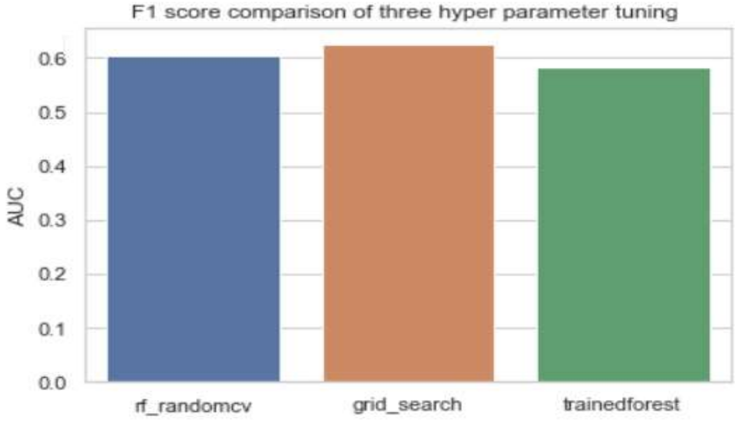
Companies B & C
Product is expensive and inconvenient to use

Companies D & E
Product is affordable, but inconvenient to use

MODELING – RANDOM FOREST

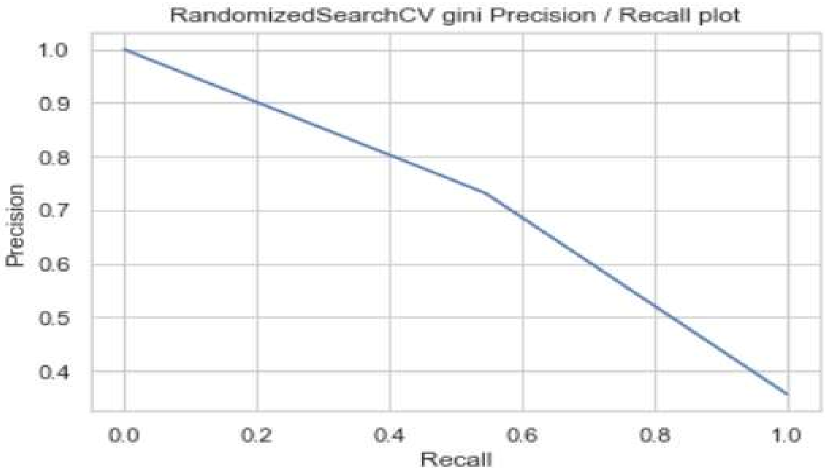


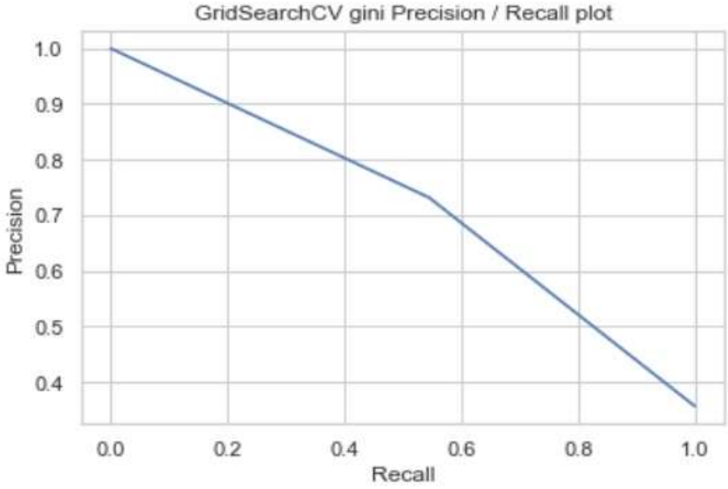


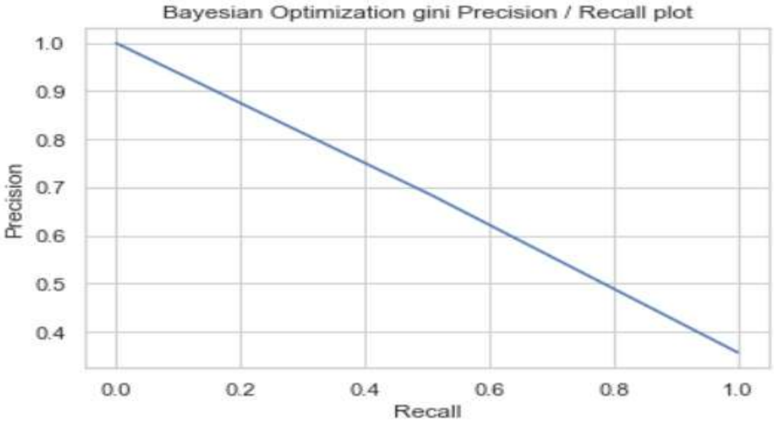


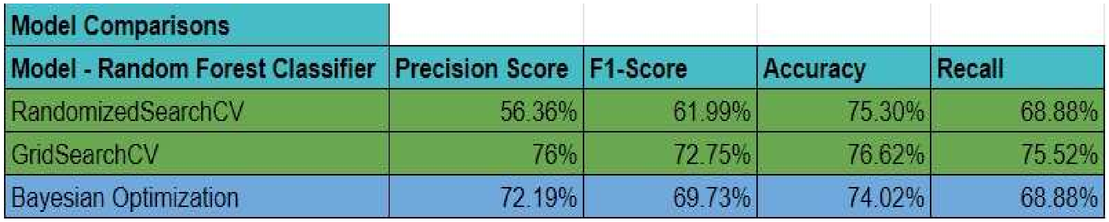
- Three hyperparameter tuning techniques performed in random forest
- GridSearchCV tuning gave better result

PRECISION COMPARISON



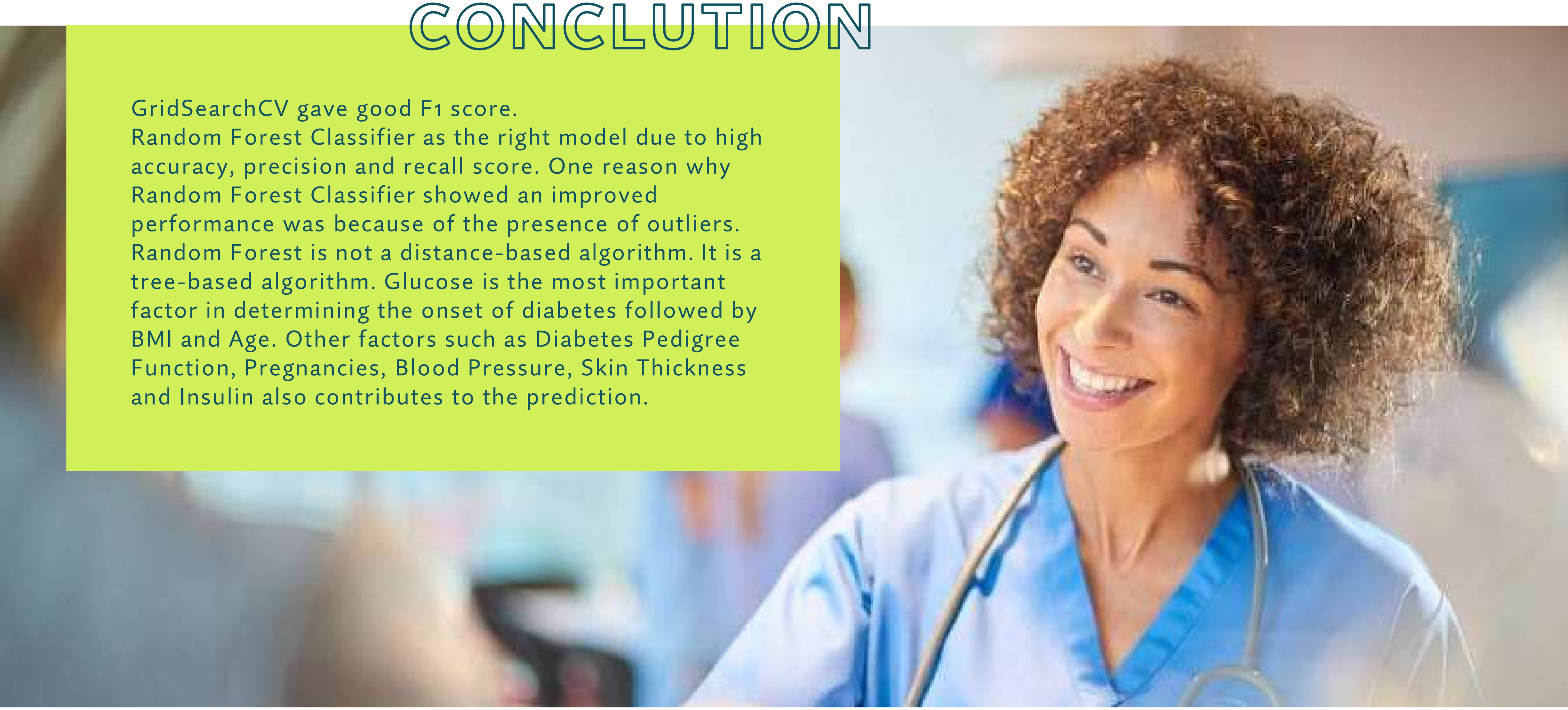






Model Comparisons				
Model - Random Forest Classifier	**Precision Score**	**F1-Score**	**Accuracy**	**Recall**
RandomizedSearchCV	56.36%	61.99%	75.30%	68.88%
GridSearchCV	76%	72.75%	76.62%	75.52%
Bayesian Optimization	72.19%	69.73%	74.02%	68.88%

CONCLUTION



GridSearchCV gave good F1 score.
Random Forest Classifier as the right model due to high accuracy, precision and recall score. One reason why Random Forest Classifier showed an improved performance was because of the presence of outliers. Random Forest is not a distance-based algorithm. It is a tree-based algorithm. Glucose is the most important factor in determining the onset of diabetes followed by BMI and Age. Other factors such as Diabetes Pedigree Function, Pregnancies, Blood Pressure, Skin Thickness and Insulin also contributes to the prediction.

THANK YOU

Archana Robin