EDA and Comparision of Seed and Synthetic Data (Pima-indians-diabetes)

~Hitesh Kumar (hiteshkumar111@outlook.com)

Seed Data

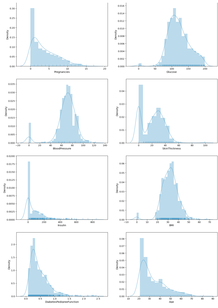
Shape of Seed data: 768 x 9

	Features		
0	Pregnancies	768 non-null	int64
1	Glucose	768 non-null	int64
2	BloodPressure	768 non-null	int64
3	SkinThickness	768 non-null	int64
4	Insulin	768 non-null	int64
5	BMI	768 non-null	float64
6	DiabetesPedigreeFunction	768 non-null	float64
7	Age	768 non-null	int64
8	Outcome	768 non-null	int64

<---dependent Variable

Distribution of all variables in Seed data

Distribution of Independent variables

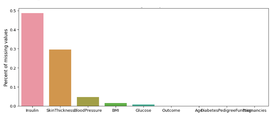


From here we can see that there are many zero values which are in pregnancies, Glucose,SkinThickness, Blood pressure, Insuline, BMI variable.

In case of Pregnancy variable zeros are explainable.

In other other case they are not, these are data discrepancy, So we will treat them as Missing values.

Missing Values



Missing values Percentages

	Total	Percent
Insulin	374	48.7%
SkinThickness	227	29.6%
BloodPressure	35	4.6%
BMI	11	1.4%
Glucose	5	0.65%

Synthetic data

Shape of Seed data: 1301 x 9

Missing values

Pregnancies	0
Glucose	0
BloodPressure	0
SkinThickness	0
Insulin	1
BMI	0
DiabetesPedigreeFunction	5
Age	0
Outcome	0

#This Synthetic Missing is generated using Gretel Api

There are total of 6 missing value in synthetic data.

1 in Insulin variable and 5 in DiabetesPedigreeFunction variable

Seed Data vs Synthetic Data

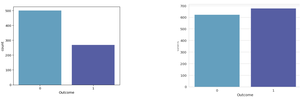
-respectively

Number of training lines duplicated in synthetic lines	0
Mean Absolute Error between training and synthetic correlation matrices	0.0853
Average field Jenson-Shannon distributional distance	0.1827

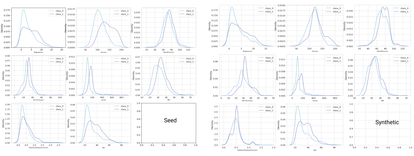
Note: distance scores range from 0 (no difference) to 1 (maximally different)

Field Name	Unique Values	Distance Score
DiabetesPedigreeFunction	517	0.4320
Insulin	187	0.2114
SkinThickness	50	0.2060
Glucose	135	0.1616
Outcome	2	0.1479
BloodPressure	46	0.1418
Age	52	0.1149
Pregnancies	17	0.1148
BMI	247	0.1138

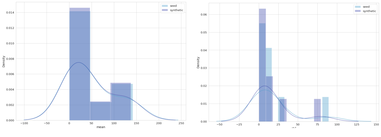
Distribution of Traget variable



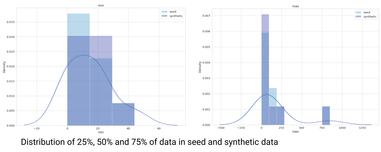
Distribution of other variables variable



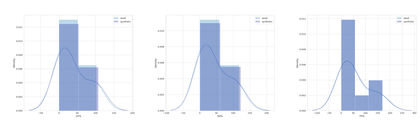
Distribution of Mean and Std. deviation in seed and synthetic data



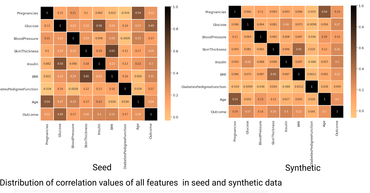
Distribution of Min and Max in seed and synthetic data



Distribution of 25%, 50% and 75% of data in seed and synthetic data



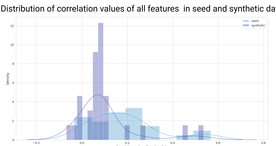
Correlation between features in Seed and synthetic data



Seed

Synthetic

Distribution of correlation values of all features in seed and synthetic data



There's less correlation in synthectic data as compared to seed.

As the distribution of correlation values in synthetic data is less spread-out and more centred around close to zero.

Findings

The seed data was imbalanced data but in syntetic data minority class is now a bit more than majority class of seed data and now it much more balanced data.

Distribution of Mean, std. deviation, min and max over all features is quite same, as they overlap each other.

#Distribution of 25%, 50% and 75% of data over all features in seed and synthetic data is also quite same.

There are some diffrences pin correlation between features of two seed and synthetic data but for more part its is quite comparable, for e.g there is a correlation of +0.58 between insulin and glucose in seed data but in the synthetic data is significantly low.

Jensen-Shannon score(0.1827) is also low which means there's less diffrence between seed and synthetic data.