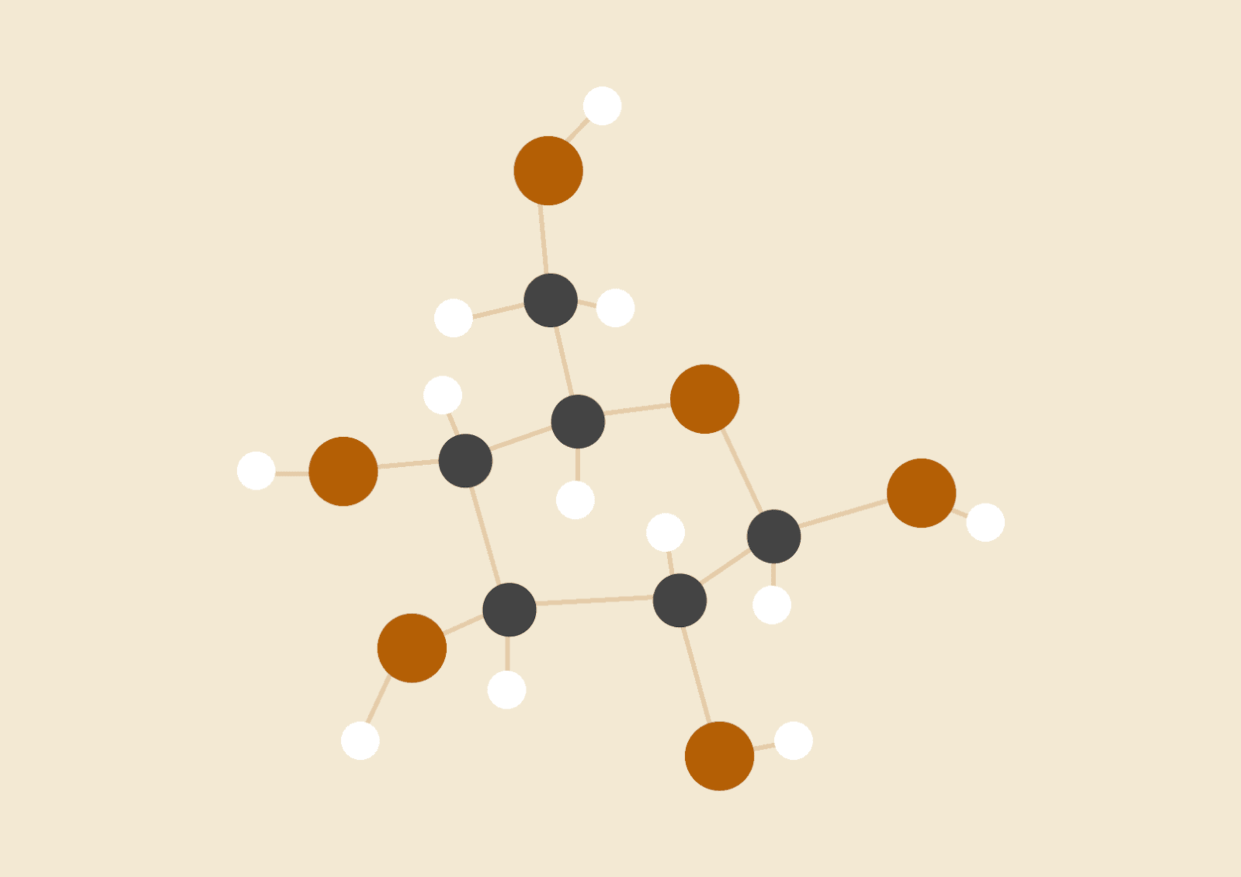
PROJECT REPORT



Gaman Gandi

ES21BTECH11014

INTRODUCTION

I used the Guassian Naive Bayes model to solve the binary classification problem.

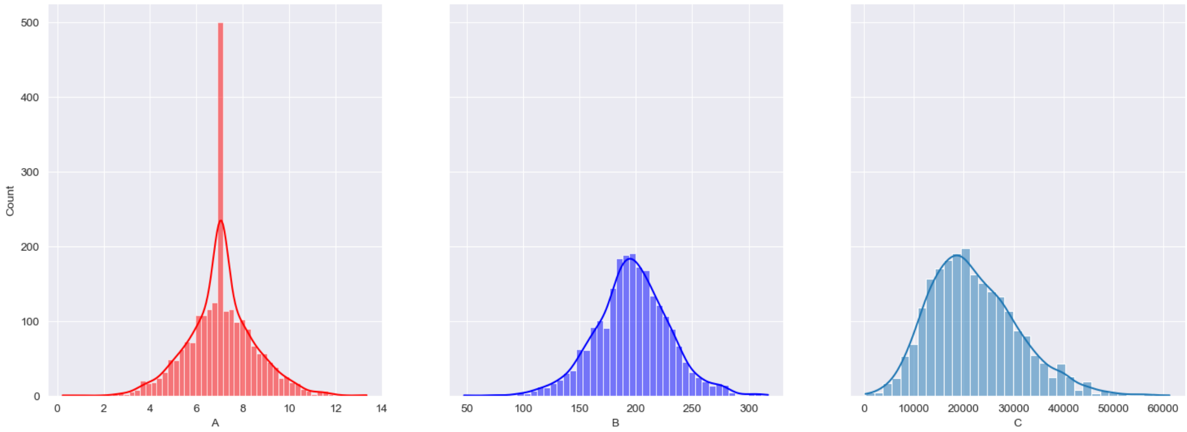
First I tried implementing other models too (logistic regression) but got very less accuracy(around 50% which is very less) so I had to go with Naive Bayes implementation.

PREPROCESSING AND DATA ANALYSIS

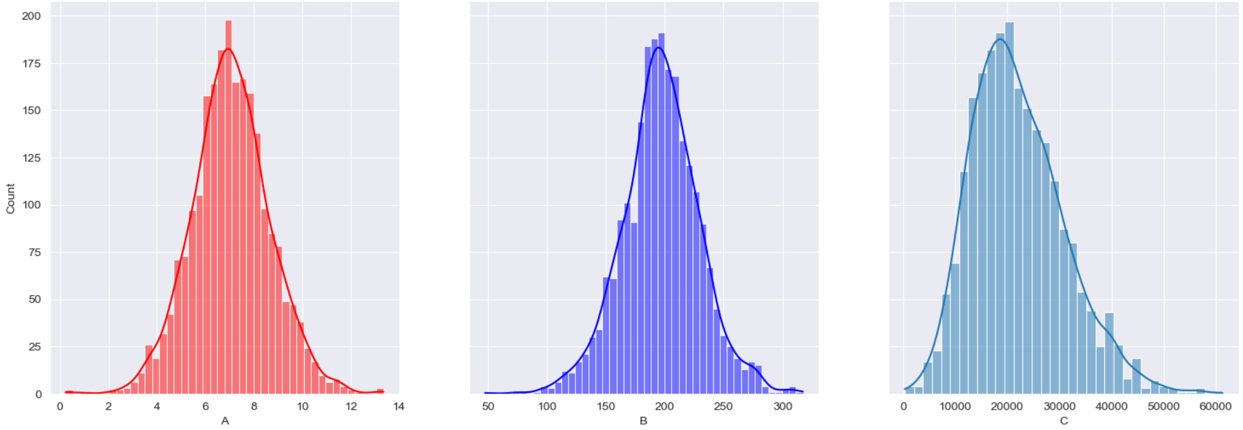
Checked if there are any null values.

Tried to replace the null values with the mean of that column

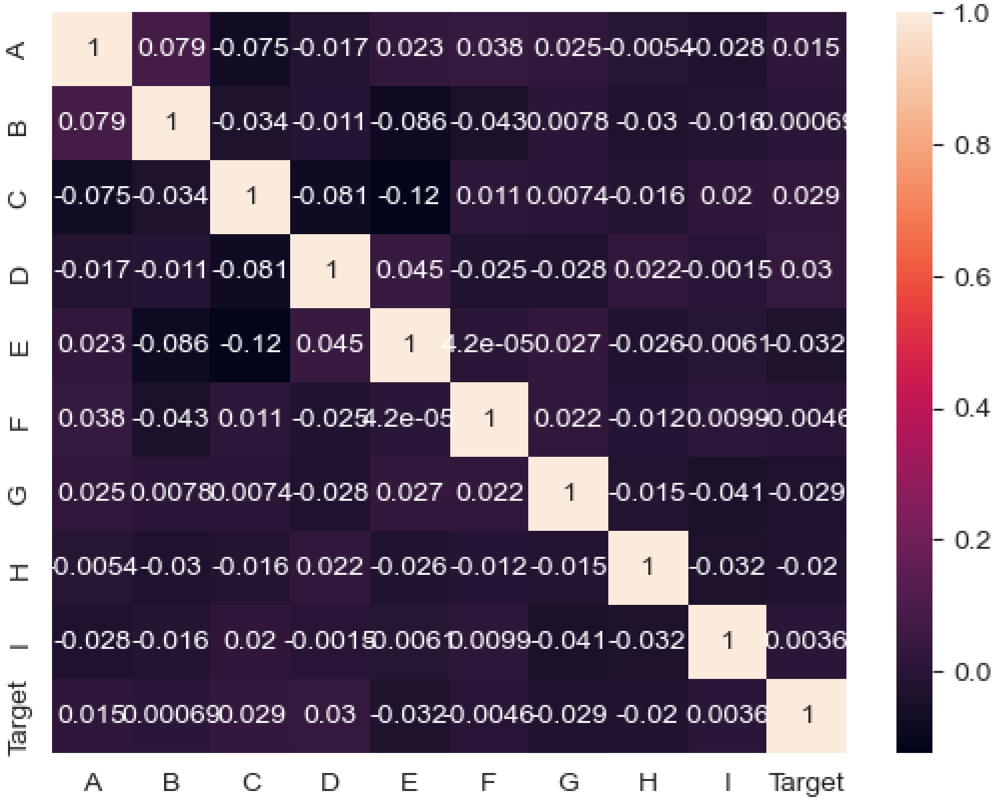
As there are so many null values in some of the features, replacing them with the mean results in accumulation of the data near the mean and the data is no longer following the normal distribution.



So I replaced the null values with the corresponding feature’s previous row value, now the data is following the normal distribution as can be seen below



Plotted the correlation matrix

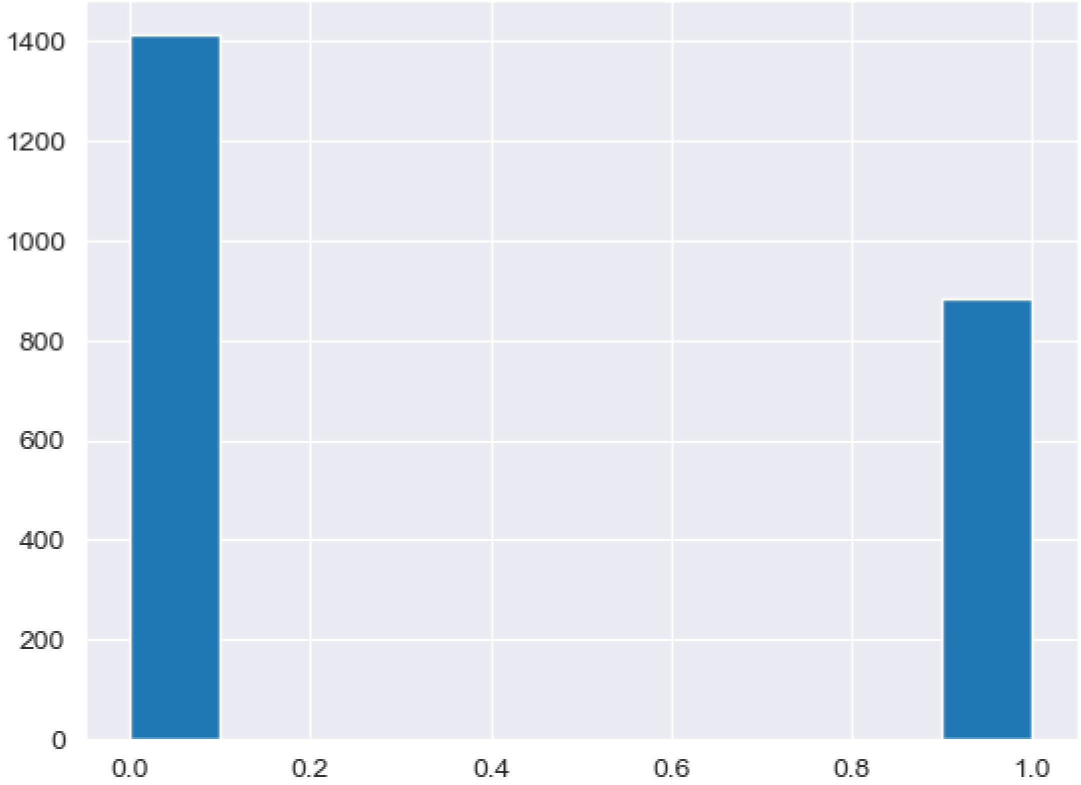


	A	B	C	D	E	F	G	H	I	Target
A	1	0.079	-0.075	-0.017	0.023	0.038	0.025	-0.0054	-0.028	0.015
B	0.079	1	-0.034	-0.011	-0.086	-0.043	0.0078	-0.03	-0.016	0.00069
C	-0.075	-0.034	1	-0.081	-0.12	0.011	0.0074	-0.016	0.02	0.029
D	-0.017	-0.011	-0.081	1	0.045	-0.025	-0.028	0.022	-0.0015	0.03
E	0.023	-0.086	-0.12	0.045	1	4.2e-05	0.027	-0.026	-0.0061	-0.032
F	0.038	-0.043	0.011	-0.025	4.2e-05	1	0.022	-0.012	0.0099	-0.0046
G	0.025	0.0078	0.0074	-0.028	0.027	0.022	1	-0.015	-0.041	-0.029
H	-0.0054	-0.03	-0.016	0.022	-0.026	-0.012	-0.015	1	-0.032	-0.02
I	-0.028	-0.016	0.02	-0.0015	-0.0061	0.0099	-0.041	-0.032	1	0.0036
Target	0.015	0.00069	0.029	0.03	-0.032	-0.0046	-0.029	-0.02	0.0036	1

It can be observed that the correlation between column B and the target is very low and the features are also almost independent of each other.

So I removed the column B from the data (which actually improved the accuracy)

Next I normalized the data.



Number of 1’s are less than the number of 0’s. So i tried to use oversampling (but that didn’t workout so i didn’t do)

MODEL IMPLEMENTATION

Our goal : Given a features(data point) , we have to predict the class to which it belongs to (i.e, in binary classification : 0 or 1)

$P(Y = y \mid X) = ?$

We can easily find this using Bayes theorem :

$$P(Y = y \mid X) = \frac{P(X \mid Y = y) \bullet P(Y = y)}{P(X)}$$

Denominator can be ignored because it doesnt depend on the class

If $P(Y = 0 \mid X) > P(Y = 1 \mid X)$ then the class is 0

Else class is 1.

Using the probability distribution function for gaussian distribution , $P(X \mid Y = y)$ can be evaluated.

I splitted the given training data into train validation split (80 : 20)

I evaluated my model on the validation data and did fine tuning.

ACCURACY

After fine tuning,

I got Accuracy for validation data around 62%

I got Accuracy for test data around 60%