Introduction:

The project’s main objective is to fit the Logistic model on respective datasets to estimate the relationship between the dependentvariable & one or more independent variables. Mathematical methods like logistic regression are applied to get future outcomes of thedataset and to behave accordingly to get the best out of it.

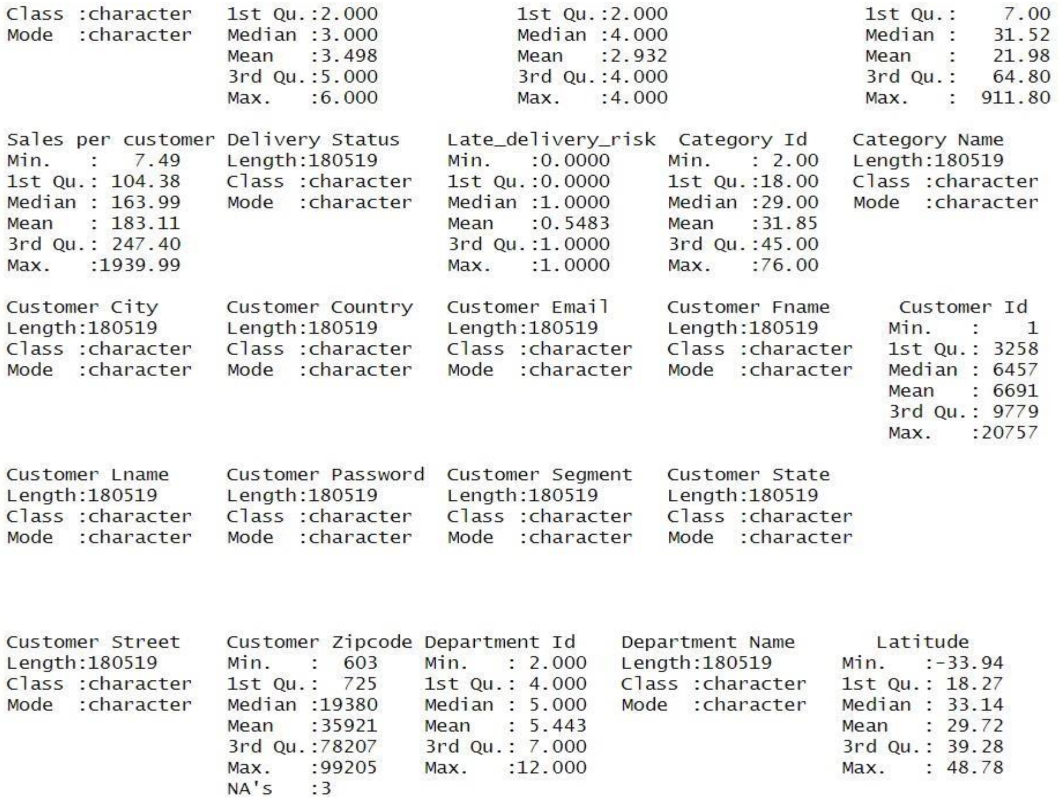
NOTE - All the missing values and incorrect observations are corrected with the process of data cleaning on both of the datasets

Objective:

The whole motive of the Project is centered to get future outcomes of the dataset by logistics we are showing effects of smoking on health parameters. On general grounds we always heard that “smoking is injurious to health” but to what degree? Beingthe real question. From the logistic model, we are actually getting how it is dependent on different health parameters on mathematical grounds. So firstly we will start with the Summary of the dataset Programmed in R!!!

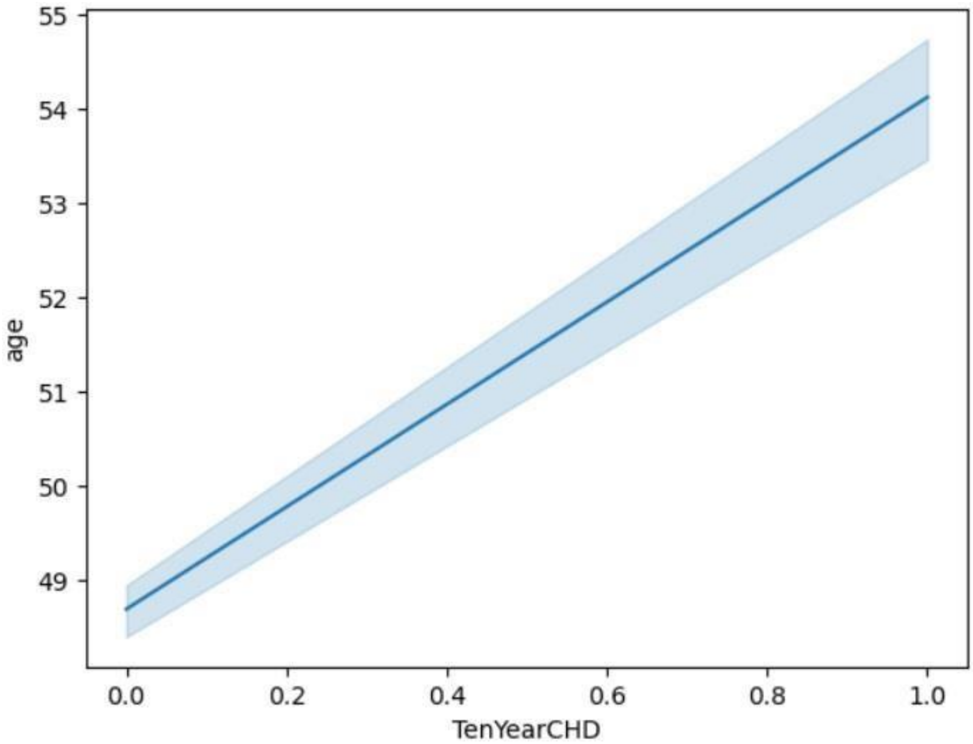
Summary of Dataset:

CHD Dataset (Summary)

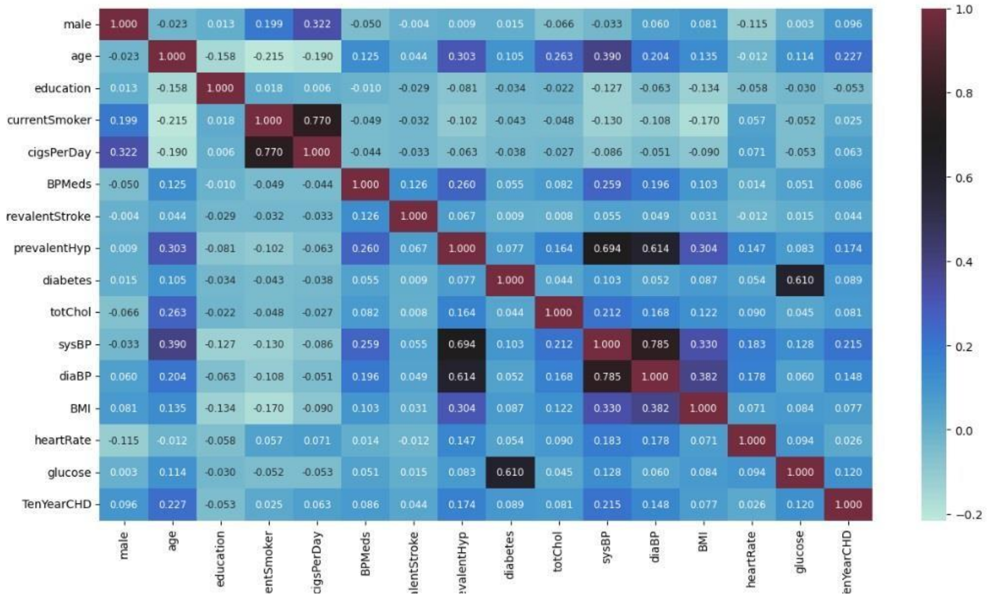


```
 Class :character   1st Qu.:2.000            1st Qu.:2.000               1st Qu.:    7.00
 Mode  :character   Median :3.000            Median :4.000               Median :   31.52
                    Mean   :3.498            Mean   :2.932               Mean   :   21.98
                    3rd Qu.:5.000            3rd Qu.:4.000               3rd Qu.:   64.80
                    Max.   :6.000            Max.   :4.000               Max.   :  911.80

 Sales per customer Delivery Status    Late_delivery_risk  Category Id    Category Name
 Min.   :   7.49    Length:180519      Min.   :0.0000      Min.   : 2.00  Length:180519
 1st Qu.: 104.38    Class :character   1st Qu.:0.0000      1st Qu.:18.00  Class :character
 Median : 163.99    Mode  :character   Median :1.0000      Median :29.00  Mode  :character
 Mean   : 183.11                       Mean   :0.5483      Mean   :31.85
 3rd Qu.: 247.40                       3rd Qu.:1.0000      3rd Qu.:45.00
 Max.   :1939.99                       Max.   :1.0000      Max.   :76.00

 Customer City      Customer Country   Customer Email     Customer Fname       Customer Id
 Length:180519      Length:180519      Length:180519      Length:180519      Min.   :    1
 Class :character   Class :character   Class :character   Class :character   1st Qu.: 3258
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 6457
                                                                             Mean   : 6691
                                                                             3rd Qu.: 9779
                                                                             Max.   :20757

 Customer Lname     Customer Password  Customer Segment   Customer State
 Length:180519      Length:180519      Length:180519      Length:180519
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


 Customer Street    Customer Zipcode Department Id    Department Name       Latitude
 Length:180519      Min.   :  603    Min.   : 2.000   Length:180519      Min.   :-33.94
 Class :character   1st Qu.:  725    1st Qu.: 4.000   Class :character   1st Qu.: 18.27
 Mode  :character   Median :19380    Median : 5.000   Mode  :character   Median : 33.14
                    Mean   :35921    Mean   : 5.443                      Mean   : 29.72
                    3rd Qu.:78207    3rd Qu.: 7.000                      3rd Qu.: 39.28
                    Max.   :99205    Max.   :12.000                      Max.   : 48.78
                    NA's   :3
```

Exploratory Data Analysis [EDA]:

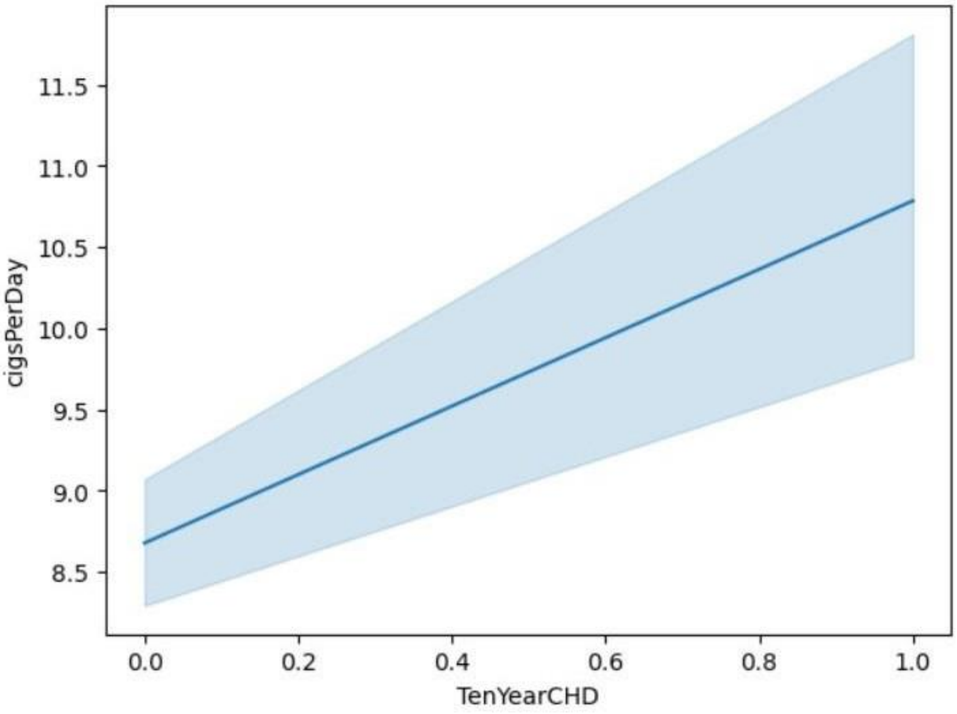


The above graph is showing how ten-year Chronic Heart disease (CHD) is varying with age. According to Graph and even according to medical science, both factors are linearly dependent on each other. Mathematically, age tends to infinity , CHD is also tending to infinity.



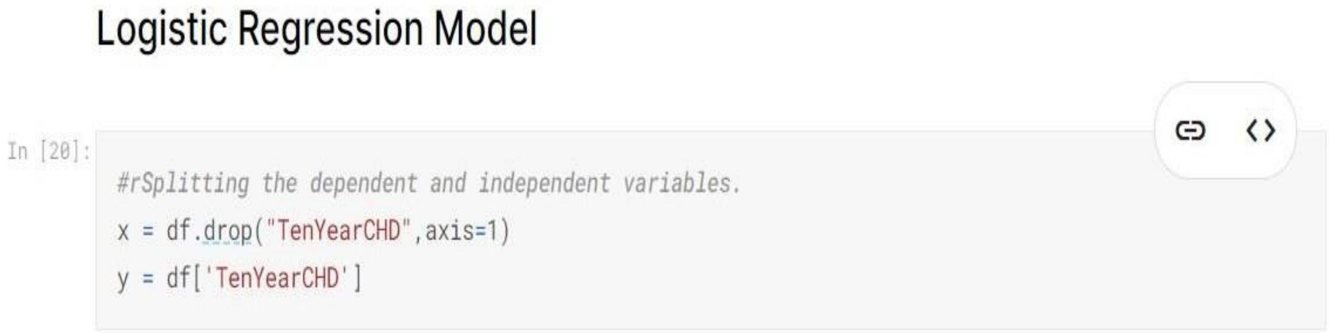
	male	age	education	entSmoker	cigsPerDay	BPMeds	lentStroke	evalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	enYearCHD
male	1.000	-0.023	0.013	0.199	0.322	-0.050	-0.004	0.009	0.015	-0.066	-0.033	0.060	0.081	-0.115	0.003	0.096
age	-0.023	1.000	-0.158	-0.215	-0.190	0.125	0.044	0.303	0.105	0.263	0.390	0.204	0.135	-0.012	0.114	0.227
education	0.013	-0.158	1.000	0.018	0.006	-0.010	-0.029	-0.081	-0.034	-0.022	-0.127	-0.063	-0.134	-0.058	-0.030	-0.053
currentSmoker	0.199	-0.215	0.018	1.000	0.770	-0.049	-0.032	-0.102	-0.043	-0.048	-0.130	-0.108	-0.170	0.057	-0.052	0.025
cigsPerDay	0.322	-0.190	0.006	0.770	1.000	-0.044	-0.033	-0.063	-0.038	-0.027	-0.086	-0.051	-0.090	0.071	-0.053	0.063
BPMeds	-0.050	0.125	-0.010	-0.049	-0.044	1.000	0.126	0.260	0.055	0.082	0.259	0.196	0.103	0.014	0.051	0.086
revalentStroke	-0.004	0.044	-0.029	-0.032	-0.033	0.126	1.000	0.067	0.009	0.008	0.055	0.049	0.031	-0.012	0.015	0.044
prevalentHyp	0.009	0.303	-0.081	-0.102	-0.063	0.260	0.067	1.000	0.077	0.164	0.694	0.614	0.304	0.147	0.083	0.174
diabetes	0.015	0.105	-0.034	-0.043	-0.038	0.055	0.009	0.077	1.000	0.044	0.103	0.052	0.087	0.054	0.610	0.089
totChol	-0.066	0.263	-0.022	-0.048	-0.027	0.082	0.008	0.164	0.044	1.000	0.212	0.168	0.122	0.090	0.045	0.081
sysBP	-0.033	0.390	-0.127	-0.130	-0.086	0.259	0.055	0.694	0.103	0.212	1.000	0.785	0.330	0.183	0.128	0.215
diaBP	0.060	0.204	-0.063	-0.108	-0.051	0.196	0.049	0.614	0.052	0.168	0.785	1.000	0.382	0.178	0.060	0.148
BMI	0.081	0.135	-0.134	-0.170	-0.090	0.103	0.031	0.304	0.087	0.122	0.330	0.382	1.000	0.071	0.084	0.077
heartRate	-0.115	-0.012	-0.058	0.057	0.071	0.014	-0.012	0.147	0.054	0.090	0.183	0.178	0.071	1.000	0.094	0.026
glucose	0.003	0.114	-0.030	-0.052	-0.053	0.051	0.015	0.083	0.610	0.045	0.128	0.060	0.084	0.094	1.000	0.120
TenYearCHD	0.096	0.227	-0.053	0.025	0.063	0.086	0.044	0.174	0.089	0.081	0.215	0.148	0.077	0.026	0.120	1.000

Correlation matrix for the 16 different variables with each other.



The relationship between cigs/day & ten-year chronic heart disease is also linear but has some amount of homoscedasticity around

2. Fitting Logistic Regression On Dataset

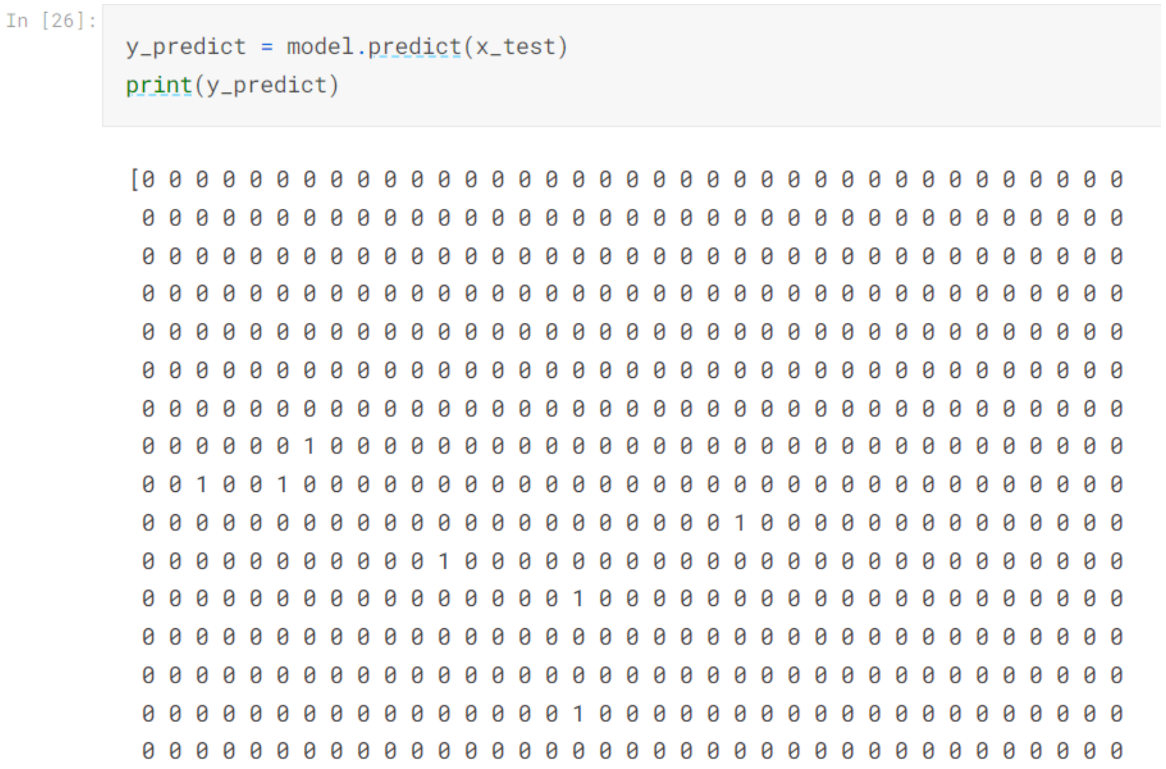


Logistic Regression Model

In [20]:

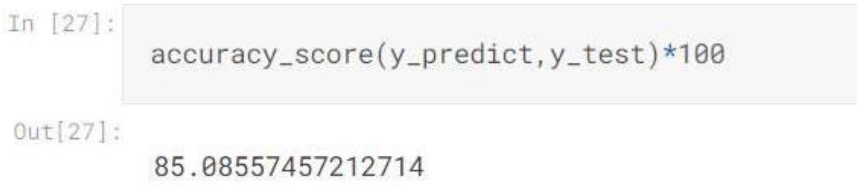
```
#rSplitting the dependent and independent variables.
x = df.drop("TenYearCHD",axis=1)
y = df['TenYearCHD']
```

and “TenYearCHD” is the dependent variable & male, currentsmoker, cigsperday, prevalantstroke & diabetes are independent variables.



```
In [26]:  y_predict = model.predict(x_test)
          print(y_predict)

[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

The diagram above is showing prediction of model, The accuracy of the model is approx **85%**



```
In [27]:  accuracy_score(y_predict,y_test)*100

Out[27]:  85.08557457212714
```

Chronic heart disease is practically primarily dependent on whether the person is a smoker or not similarly it goes parallelly with the cigs/day, prevalent strokes, gender, and current health parameters.

Logistic regression fitted above is giving 85% promise of if an individual have control on the factors which are mentioned above he/she can prevent Chronic heart disease.

Thank You !!!