Milestone 2: Project Report

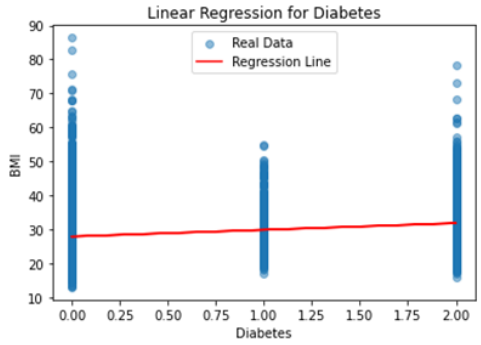
Additional Data Cleaning: In the process of additional data cleaning, labels were created to define the target variable for analysis. The dataset was meticulously examined to ensure alignment with project requirements, identifying and addressing issues such as missing data and outliers. Furthermore, specific columns were removed based on criteria like redundancy, irrelevance, and data quality, with the aim of enhancing data integrity and preventing overfitting in subsequent analysis.

Question 1: Linear Regressions.

A. Linear Regression with Diabetes, Sex, Age, COVIDPOS, Smoker, Physical_activity, Food_shortage, Average_drink, MEDCOST1, General_health, Income_category, Stroke, Walking_difficulty, HeartDiseaseorAttack, Alcohol_consumed as our independent variables and BMI as our continuous dependent variable.

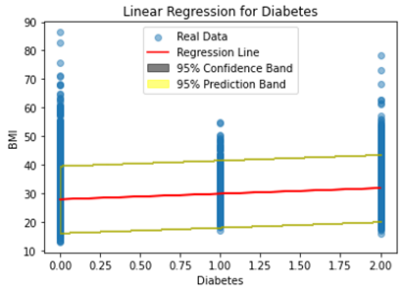
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
Diabetes	27.93	1.96	predictive	0.184	35.89

Plot




```
Residuals: 62317    -2.425669
93046    -6.625669
86085     7.574331
23841    -2.515669
84876    -2.245669
             ...
81158     4.074331
13379    -2.835669
71580    -2.875669
100018    7.494331
10206    -2.325669
```

Confidence and Prediction Bands:

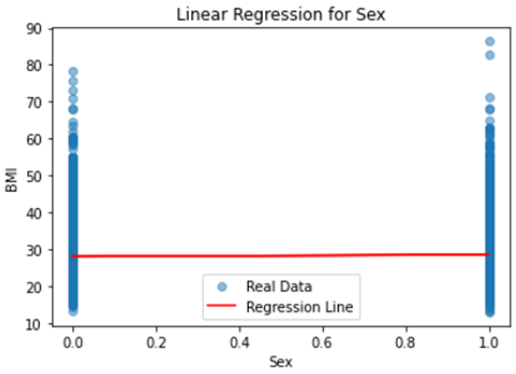


Interpretation of the Result: A correlation coefficient of 0.184 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that

there is some degree of correlation, but it's not very strong. MSE of 35.89 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

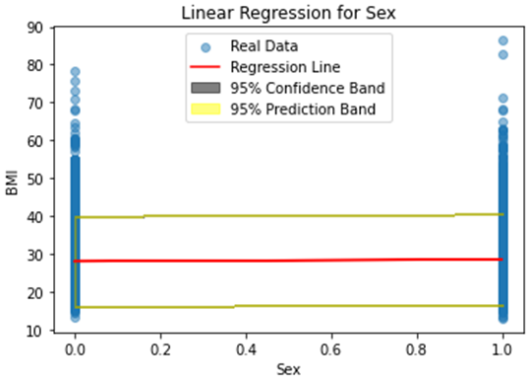
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
Sex	28.05	0.509	Weakly predictive	0.046	37.07

Plot




```
Residuals: 62317    -2.545197
93046    -6.745197
86085     7.454803
23841    -3.144459
84876    -2.365197
             ...
81158     3.954803
13379    -3.464459
71580    -2.995197
100018    7.374803
10206    -2.445197
```

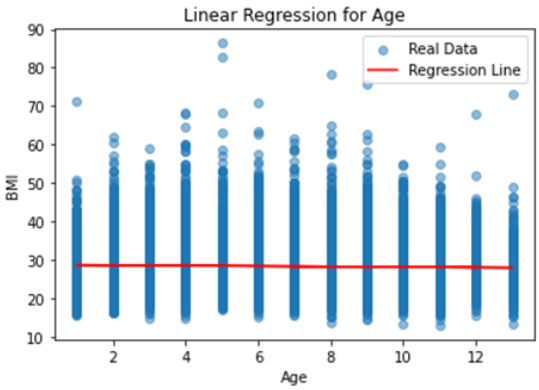
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.046 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 37.07 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

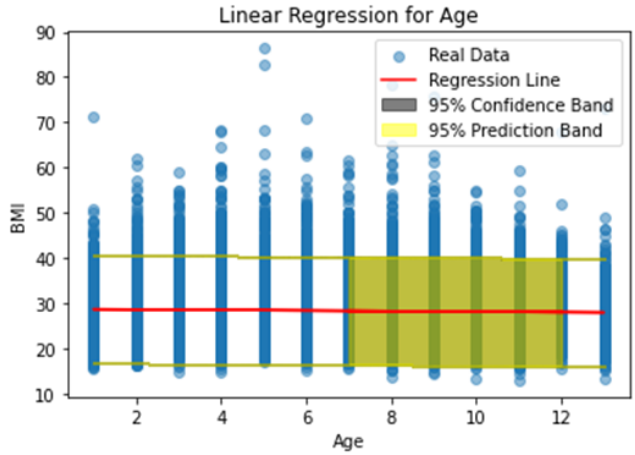
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
Age	28.76	-0.059	Weakly predictive	0.026	37.13

Plot




```
Residuals: 62317    -2.533023
93046    -6.912862
86085     7.407030
23841    -2.802862
84876    -2.712701
             ...
81158     3.787138
13379    -3.482540
71580    -3.042970
100018    6.967353
10206    -2.732755
```

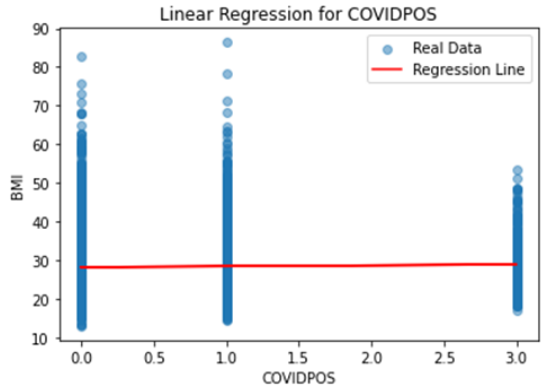
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.026 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 37.13 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

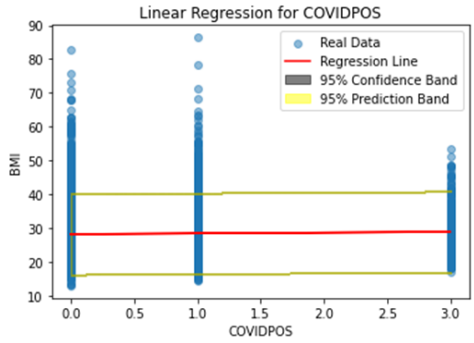
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
COVIDPOS	28.22	0.230	Weakly predictive	0.0279	37.12

Plot




```
Residuals: 62317    -2.714933
93046    -6.914933
86085     7.285067
23841    -2.804933
84876    -2.534933
             ...
81158     3.785067
13379    -3.124933
71580    -3.164933
100018    6.974202
10206    -2.614933
```

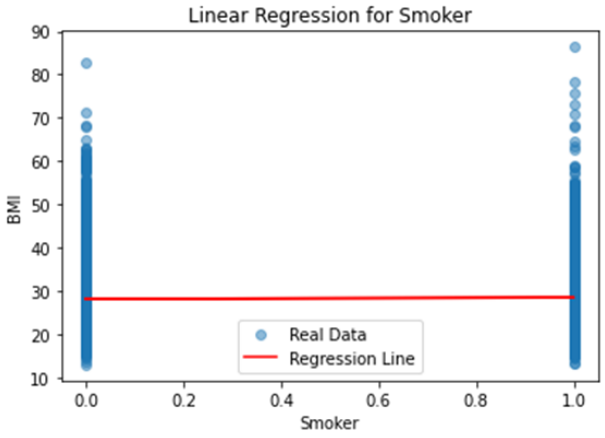
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.0279indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 37.12 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

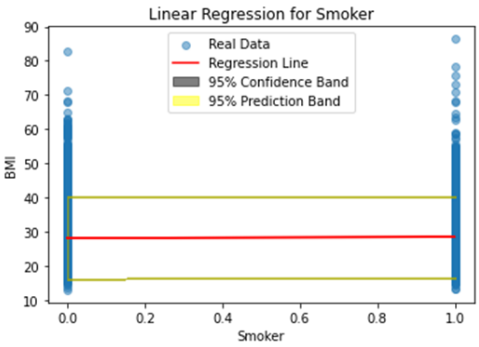
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
Smoker	28.20	0.276	Weakly predictive	0.031	37.12

Plot




```
Residuals: 62317    -2.699867
93046    -6.899867
86085     7.023801
23841    -3.066199
84876    -2.519867
             ...
81158     3.523801
13379    -3.109867
71580    -3.426199
100018    7.220133
10206    -2.599867
```

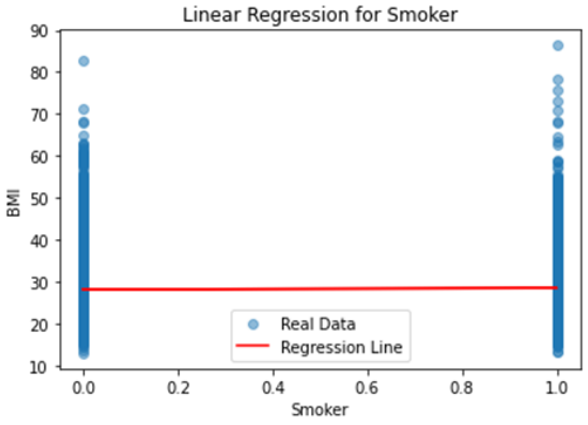
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.0231indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 37.12 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

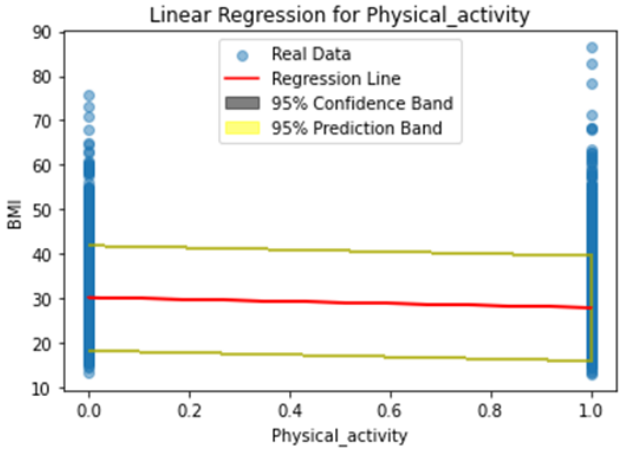
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
Physical Activity	30.23	-2.282	predictive	0.156	36.25

Plot




```
Residuals: 62317    -2.437817
93046    -6.637817
86085     7.562183
23841    -2.527817
84876    -2.257817
             ...
81158     4.062183
13379    -2.847817
71580    -2.887817
100018    5.199630
10206    -2.337817
```

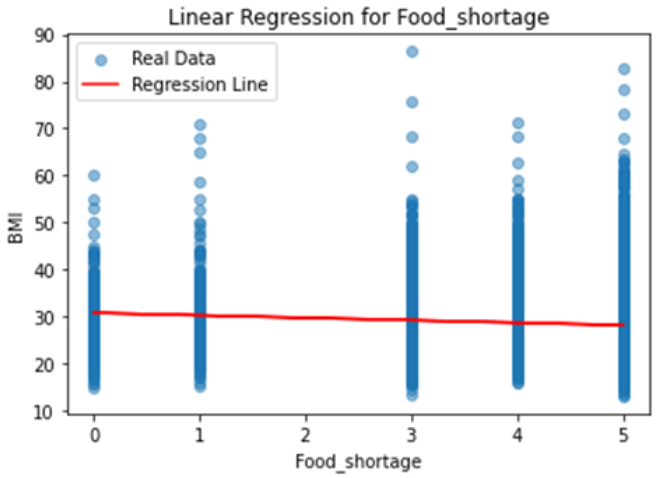
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.156 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 36.25 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

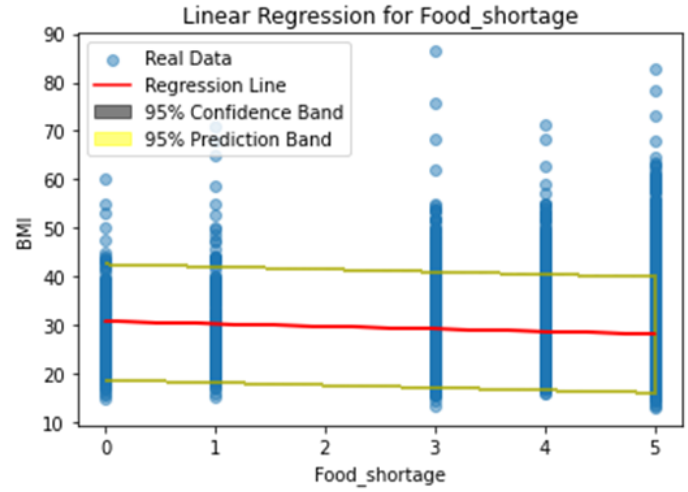
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
Food_shortage	30.78	-0.5211	Weakly predictive	0.0557	37.05

Plot




```
Residuals: 62317    -2.672264
93046    -6.872264
86085     7.327736
23841    -3.804553
84876    -2.492264
             ...
81158     3.827736
13379    -3.082264
71580    -3.122264
100018    7.247736
10206    -2.572264
```

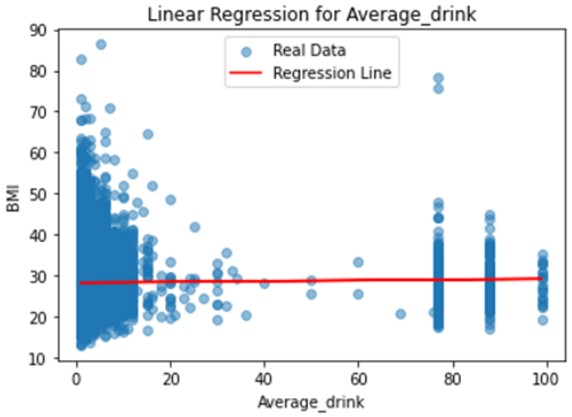
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.0557 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 37.05 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

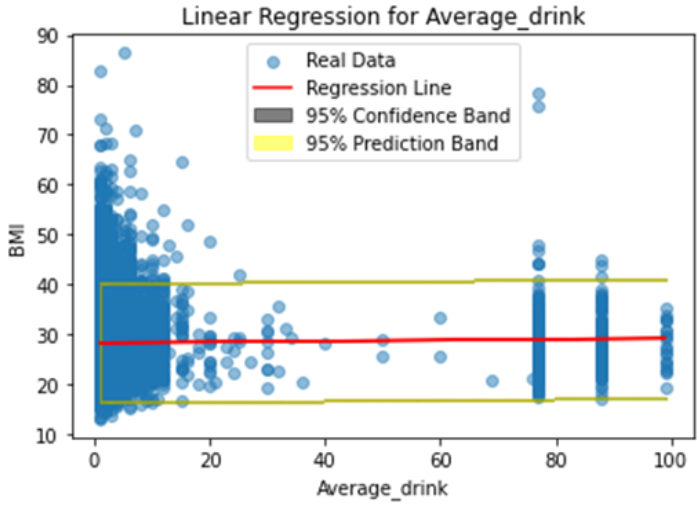
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
Average Drink	28.29	0.0092	Weakly predictive	0.018	37.14

Plot




```
Residuals: 62317    -2.793766
93046    -6.993766
86085     7.196975
23841    -2.893025
84876    -2.613766
             ...
81158     3.696975
13379    -3.203766
71580    -3.243766
100018    7.116975
10206    -2.693766
```

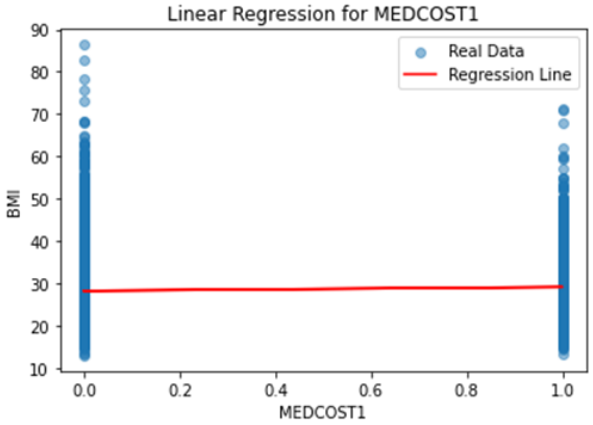
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.018 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 37.14 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

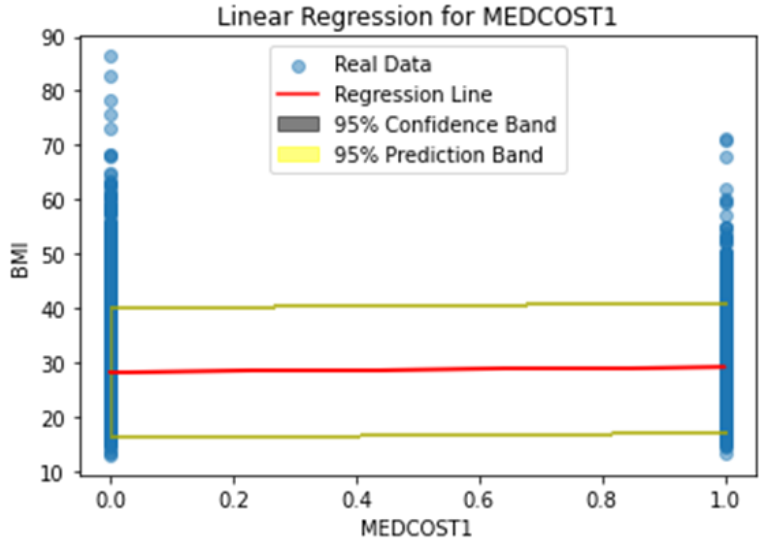
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
MEDCOST1	28.25	0.905	Weakly predictive	0.027	37.13

Plot




```
Residuals: 62317    -2.748616
93046    -6.948616
86085     7.251384
23841    -3.743908
84876    -2.568616
             ...
81158     3.751384
13379    -3.158616
71580    -3.198616
100018    7.171384
10206    -2.648616
```

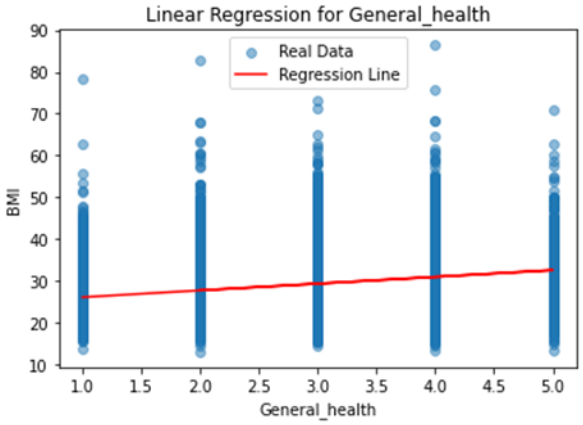
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.027 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 37.13 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

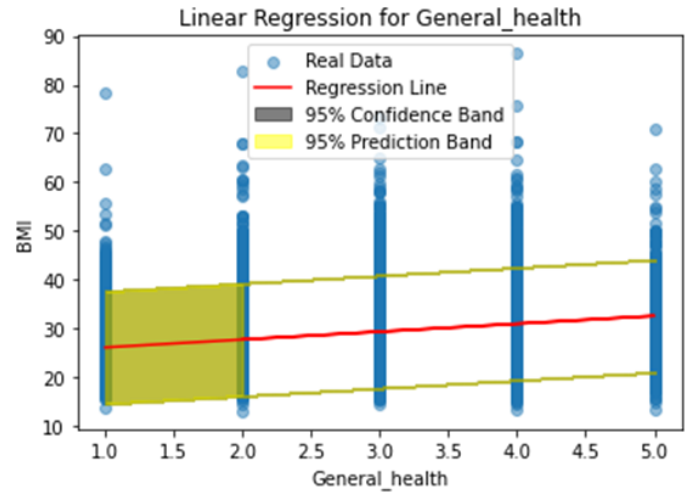
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
General health	24.49	1.6177	predictive	0.25	34.75

Plot




```
Residuals: 62317    -2.218938
93046    -9.654457
86085     6.163302
23841    -5.544457
84876    -2.038938
             ...
81158     4.281062
13379    -2.628938
71580    -4.286698
100018    4.465543
10206    -0.501179
```

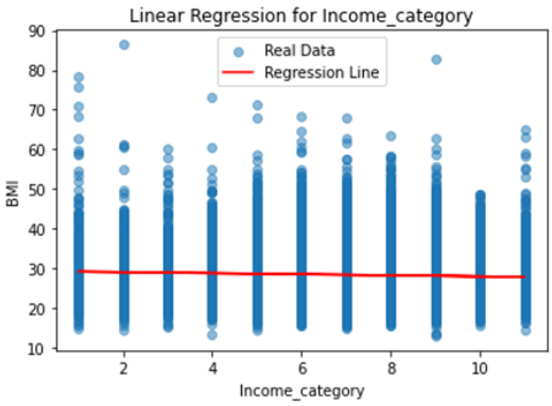
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.25 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 34.75 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

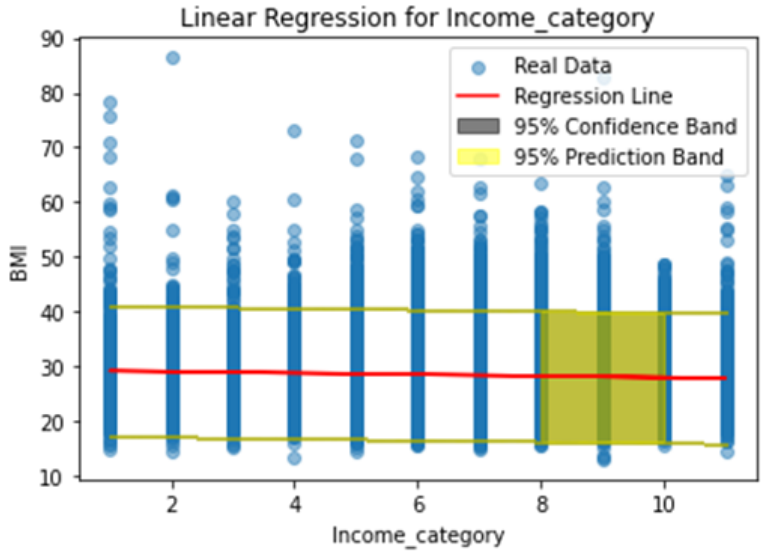
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
Income_category	29.34	-0.135	Weakly predictive	0.060	37.02

Plot:




```
Residuals: 62317    -2.476240
93046    -7.353171
86085     7.117601
23841    -3.649330
84876    -2.296240
             ...
81158     3.211443
13379    -3.292399
71580    -3.467785
100018    7.308373
10206    -2.647013
```

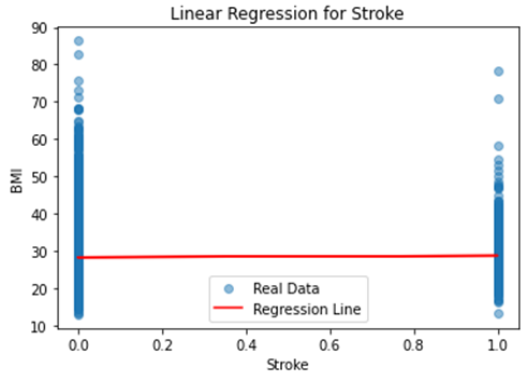
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.060 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 37.20 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

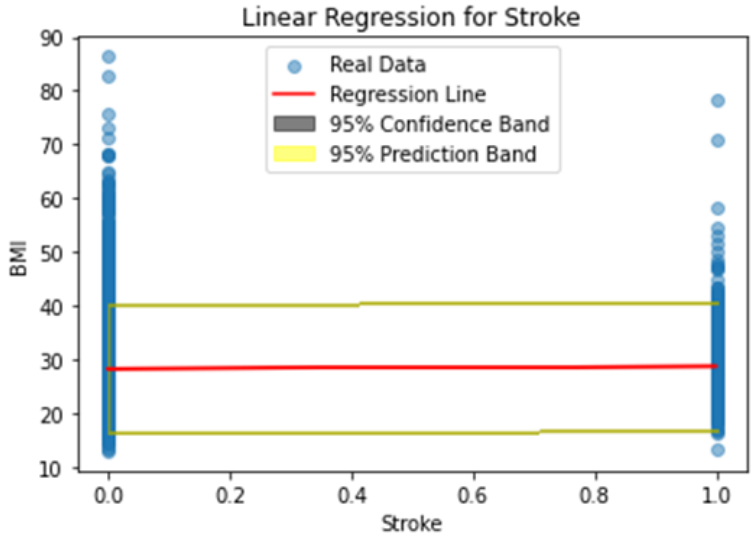
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
Stroke	28.31	0.4444	Weakly predictive	0.008	37.15

Plot




```
Residuals: 62317    -2.80285
93046    -7.00285
86085     7.19715
23841    -2.89285
84876    -2.62285
           ...
81158     3.69715
13379    -3.21285
71580    -3.25285
100018    7.11715
10206    -2.70285
```

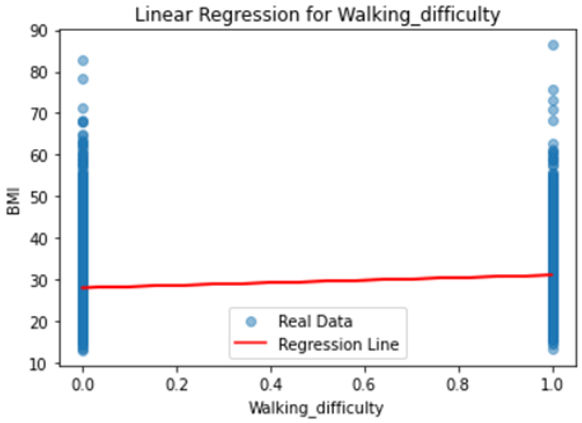
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.008 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 37.15 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

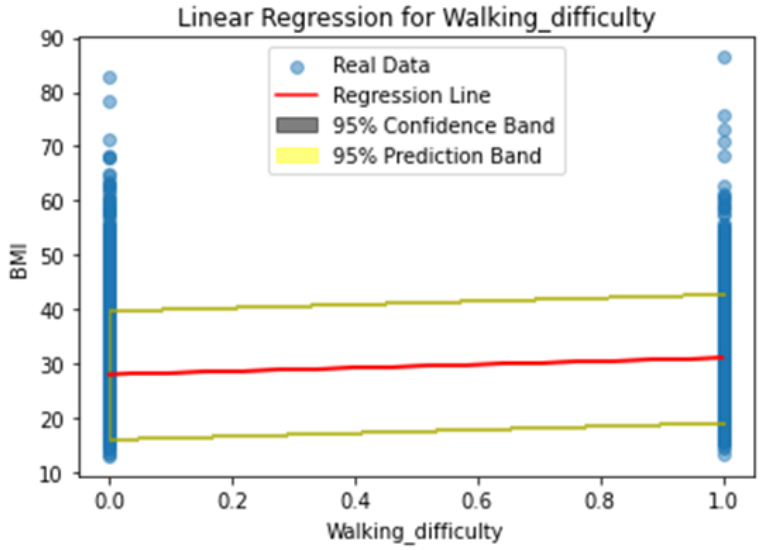
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
Walking_difficulty	28.02	3.06	predictive	0.144	36.37

Plot




```
Residuals: 62317    -2.513490
93046    -6.713490
86085     7.486510
23841    -2.603490
84876    -2.333490
             ...
81158     3.986510
13379    -2.923490
71580    -2.963490
100018    4.344443
10206    -2.413490
```

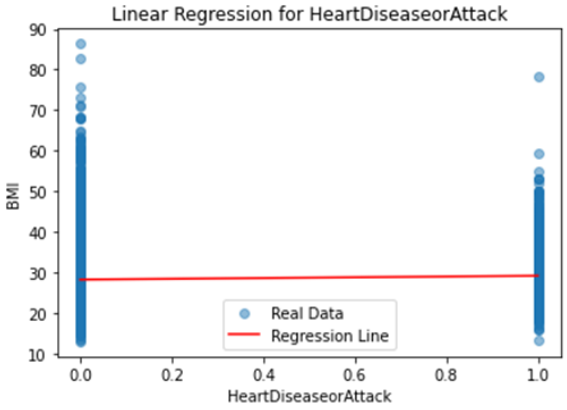
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.144 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 36.37 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

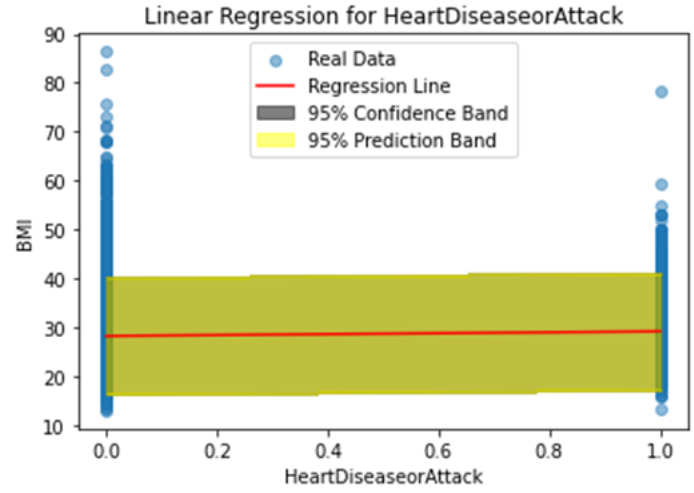
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
HeartDiseaseorattack	28.26	0.940	Weakly predictive	0.037	37.10

Plot




```
Residuals: 62317    -3.691850
93046    -6.950859
86085     7.249141
23841    -2.840859
84876    -2.570859
             ...
81158     3.749141
13379    -3.160859
71580    -3.200859
100018    7.169141
10206    -2.650859
```

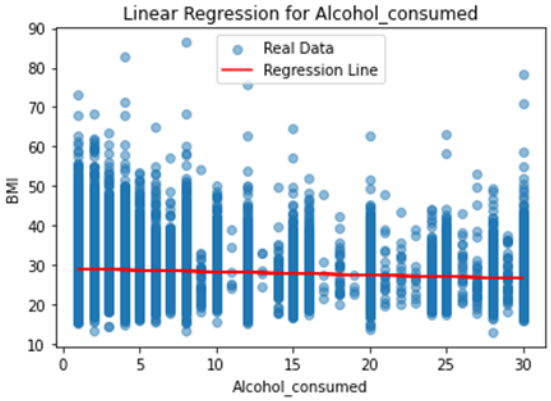
Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.037 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 37.150 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

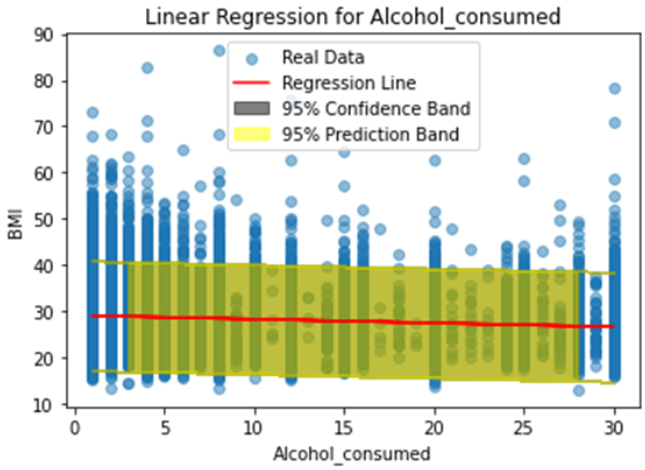
X	intercept	coefficient	Is it predictive	Corr coefficient	mse
Alcohol Consumed	29.13	-0.08	predictive	0.126	36.559

Plot




```
Residuals: 62317    -3.373236
93046    -7.242547
86085     8.032190
23841    -2.471171
84876    -1.787810
             ...
81158     3.209437
13379    -2.708498
71580    -3.988580
100018    6.464092
10206    -1.206433
```

Confidence and Prediction Bands:



Interpretation of the Result: A correlation coefficient of 0.037 indicates a weak positive linear relationship between the independent variable and the dependent variable. It suggests that there is some degree of correlation, but it's not very strong. MSE of 37.150 suggests that our model's predictions have some errors, meaning it doesn't perfectly fit the data.

The most predictive features from these are Diabetes, Physical Activity, General health, walking difficulty and alcohol consumed based on their correlation coefficient in comparison to all the other independent variables.

As the qqnorm plots for each pair of the independent and dependent variables are nonlinear we can conclude that linear regression is unsuitable for our problem . The effectiveness of our linear regression model which depends on the specific relationships between the features and BMI is low as most of the features have low correlations and high MSE indicating that they are not strongly predictive in a linear regression context.

Multivariate Linear Regression

1. Independent Variables Considered: Diabetes, Physical Activity, General health, walking difficulty, and alcohol consumed. Dependent Variable: BMI

Coefficients: [1.25615569 -1.10528884 1.21326507 -0.07322456 0.93153141]
Intercept: 26.746423569214436
Mean Squared Error (MSE): 33.38378824015716
correlation: 0.3186437628608678
R-squared (R2): 0.10150940838606604
Interpretation of the results: The correlation has improved and the MSE has become less indicating better performance.

2. Independent variables: Alcohol consumed, average drink, smoker, physical activity

Coefficients: [-0.08474693 0.00656187 0.33732934 -2.21267501]
Intercept: 30.83522111358377
correlation: 0.20267257816205891
Mean Squared Error (MSE): 35.63640478632683
R-squared (R2): 0.04088253288927257
Interpretation of the results: The correlation has improved and the MSE has become less indicating better performance but not as good as the first set of results.

3. Independent variables: Age, Sex, Income category

Coefficients: [-0.06474681 0.54991682 -0.14801778]
Intercept: 29.614991068652603
correlation: 0.0835155426768014
Mean Squared Error (MSE): 36.89854368978688
R-squared (R2): 0.006913352342390433
Interpretation of the results: The correlation has not improved and the MSE is the same indicating no improvement.

4. Independent variables: Diabetes, covid positive, general health, walking difficulty, stroke, heart disease

Coefficients: [1.41152153 0.2906196   1.34308394 1.26538354 -1.05815054 -0.69192873]
Intercept: 24.68976828290689
correlation: 0.29469473492035003
Mean Squared Error (MSE): 33.92965548497214
R-squared (R2): 0.08681794856106895
Interpretation of the results: The correlation has improved and the MSE has become less indicating better performance but not as good as the first set of results.

The most significant and **predictive feature**s(set of features) are Diabetes, Physical Activity, General health, walking difficulty, and alcohol consumed

REGULARIZATION

For regularization we used Ridge Regression model with regularization. We ran the code for both Part A and Part B and computed their corresponding results

Regularization for part A

Feature	Intercept	Coefficient	Correlation	Mean Squared Error
Diabetes	27.9357	1.9621	0.1844	35.8936
Sex	28.0552	0.5092	0.0464	37.0767
Age	28.7624	-0.0599	0.0267	37.1314
COVIDPOS	28.2249	0.2309	0.0280	37.1268
Smoker	28.2099	0.2763	0.0317	37.1218
Physical_Activity	30.2302	-2.2824	0.1565	36.2575
Food_Shortage	30.7879	-0.5211	0.0558	37.0506
Average_drink	28.2945	0.0093	0.0181	37.1442
MEDCOST1	28.2586	0.09051	0.0275	37.1341
General_Health	24.4935	1.6177	0.2546	34.7481
Incomecategory	29.3401	-0.1354	0.0608	37.0227
Stroke	28.3129	0.4444	0.0089	37.1535
WalkingDiff	28.0235	3.0617	0.1447	36.3790
HeartDisAtt	28.2609	0.9408	0.0377	37.1029

Alcohol_Cons	29.1313	-0.0827	0.1267	36.5591

Interpretation of the result: There is no change in both the correlation coefficient (R-squared) and the mean squared error (MSE) after applying regularization, which suggests that regularization might not be necessary or effective for our specific problem.

Regularization for part B

1. Independent Variables Considered: Diabetes, Physical Activity, General health, walking difficulty, and alcohol consumed. Dependent Variable: BMI
Coefficients: [1.25612261 -1.1052135 1.21327196 -0.07322468 0.93143136]
Intercept: 26.746361924107628
correlation: 0.12674577668517534
Mean Squared Error (MSE): 33.383791617621384
R-squared (R2): 0.10150931748506309

2. Independent variables: Alcohol consumed, average drink, smoker, physical activity
Coefficients: [-0.08474697 0.00656216 0.3373274 -2.21248226]
Intercept: 30.835060502273475
correlation: 0.12674577668517534
Mean Squared Error (MSE): 35.636419366785034
R-squared (R2): 0.04088214047111971

3. Independent variables: Age, Sex, Income category
Coefficients: [-0.06474673 0.54989007 -0.14801706]
Intercept: 29.614999282746
correlation: 0.12674577668517534
Mean Squared Error (MSE): 36.898544754863046
R-squared (R2): 0.006913323676952299

4. Independent variables: Diabetes, covid positive, general health, walking difficulty, stroke, heart disease
Coefficients: [1.41147085 0.29061315 1.34307808 1.26517646 -1.05765363 -0.69178849]
Intercept: 24.689791822164963
correlation: 0.12674577668517534
Mean Squared Error (MSE): 33.92965416803534
R-squared (R2): 0.0868179840050789

Interpretation of the result: There is no change in both the correlation coefficient (R-squared) and the mean squared error (MSE) after applying regularization, which suggests that regularization might not be necessary or effective for our specific problem.

Multiple Runs for Part A and Part B

For running the linear regression model on random datasets, We wrote a python code to randomly select the dataset 5 times and report its correlation coefficient and MSE. This was done for both the Parts viz A and B. The results for part A were too extensive to be reported as for each variable (14 independent variables) results were generated for each run (5 runs), they can be viewed in the python file.

Results for Part B

1.

```
Results for Run 1:
Coefficients: [ 1.27059825 -1.12652073  1.21327596 -0.07342256  0.88578973]
correlation: 0.3126672589112663
Mean Squared Error (MSE): 33.47568254361464
R-squared (R2): 0.09749679864810978

Results for Run 2:
Coefficients: [ 1.25612261 -1.1052135   1.21327196 -0.07322468  0.93143136]
correlation: 0.3186435857085497
Mean Squared Error (MSE): 33.383791617621384
R-squared (R2): 0.10150931748506309

Results for Run 3:
Coefficients: [ 1.27610109 -1.15235778  1.19535813 -0.07398075  0.82127209]
correlation: 0.32281849019301234
Mean Squared Error (MSE): 33.49551224564108
R-squared (R2): 0.1038138830170674

Results for Run 4:
Coefficients: [ 1.26313301 -1.15485892  1.21164622 -0.07321199  0.98647067]
correlation: 0.3092646959763688
Mean Squared Error (MSE): 32.39854892544028
R-squared (R2): 0.0953164461708369

Results for Run 5:
Coefficients: [ 1.2453518  -1.12043912  1.20773388 -0.07138465  0.92291617]
correlation: 0.32128764682183997
Mean Squared Error (MSE): 34.17171344799821
R-squared (R2): 0.10303893183568347
```

2.

```
Results for Run 1:
Coefficients: [-0.0859947   0.00694681  0.35827458 -2.23518292]
Correlation: 0.19222171583522413
Mean Squared Error (MSE): 35.72824638140399
R-squared (R2): 0.036767758330389144

Results for Run 2:
Coefficients: [-0.08474697  0.00656216  0.3373274  -2.21248226]
Correlation: 0.20267215864025243
Mean Squared Error (MSE): 35.636419366785034
R-squared (R2): 0.04088214047111971

Results for Run 3:
Coefficients: [-0.08629257  0.00403208  0.36766899 -2.24307047]
Correlation: 0.1909639878237928
Mean Squared Error (MSE): 36.0224534821531
R-squared (R2): 0.03620453768190368

Results for Run 4:
Coefficients: [-0.08529388  0.00762998  0.33809283 -2.27893856]
Correlation: 0.19069241395518935
Mean Squared Error (MSE): 34.51252555682531
R-squared (R2): 0.03628664529936998

Results for Run 5:
Coefficients: [-0.0839362   0.00566595  0.38573677 -2.22654559]
Correlation: 0.19800784611618047
Mean Squared Error (MSE): 36.61475026310455
R-squared (R2): 0.038912533416206374
```

3

```
Results for Run 1:
Coefficients: [-0.06234084  0.58914967 -0.15491093]
Correlation: 0.07226486407111478
Mean Squared Error (MSE): 36.906358913532834
R-squared (R2): 0.0050058866967404025

Results for Run 2:
Coefficients: [-0.06474673  0.54989007 -0.14801706]
Correlation: 0.08351547637341097
Mean Squared Error (MSE): 36.898544754863046
R-squared (R2): 0.006913323676952299

Results for Run 3:
Coefficients: [-0.05969545  0.56101649 -0.1480292 ]
Correlation: 0.08631557618897058
Mean Squared Error (MSE): 37.109341818137665
R-squared (R2): 0.007124396131095945

Results for Run 4:
Coefficients: [-0.06363051  0.58228779 -0.15435356]
Correlation: 0.07337791228555685
Mean Squared Error (MSE): 35.623007455029544
R-squared (R2): 0.00527801239910175

Results for Run 5:
Coefficients: [-0.05830225  0.57458787 -0.15059385]
Correlation: 0.08113278176552752
Mean Squared Error (MSE): 37.856283992289185
R-squared (R2): 0.006323959197177431
```

4.

```
Results for Run 1:
Coefficients: [ 1.42519059  0.31677901  1.34824187  1.23184661 -1.08846697 -0.66436081]
Correlation: 0.288002941029691
Mean Squared Error (MSE): 34.026138557916475
R-squared (R2): 0.08265652423507697

Results for Run 2:
Coefficients: [ 1.41147085  0.29061315  1.34307808  1.26517646 -1.05765363 -0.69178849]
Correlation: 0.29469472513794115
Mean Squared Error (MSE): 33.92965416803534
R-squared (R2): 0.0868179840050789

Results for Run 3:
Coefficients: [ 1.43626029  0.29288586  1.33151431  1.1696867  -1.07367703 -0.71050908]
Correlation: 0.302150350153402
Mean Squared Error (MSE): 33.978482432388816
R-squared (R2): 0.09089181832058557

Results for Run 4:
Coefficients: [ 1.42192371  0.3250012   1.34534363  1.32554115 -1.06091645 -0.64503413]
Correlation: 0.2846082286819056
Mean Squared Error (MSE): 32.926082605080246
R-squared (R2): 0.08058581594540382

Results for Run 5:
Coefficients: [ 1.39897219  0.27959358  1.34011399  1.25406262 -1.07755562 -0.63650056]
Correlation: 0.2961450568773552
Mean Squared Error (MSE): 34.763059281349676
R-squared (R2): 0.08751690683846969
```

Similarities and differences:

Observation: That there were no significant changes in the correlation coefficient and MSE across multiple runs.
Interpretation: The stability of the correlation coefficient and MSE across multiple runs indicates a robust model that is resilient to random variations in training and testing data splits. Consistent performance across different datasets suggests strong generalization capabilities, implying the model is likely to perform reliably on new, real-world data. This positive sign points to the model's stability and its potential for consistent real-world applicability. Additionally, when we use regularization techniques to exhibit stability across runs, it signifies that regularization effectively prevents overfitting and maintains consistent model performance. This further contributes to the model's reliability and its ability to provide consistent and accurate predictions.

Question 2. Logistic Regression and NB.

a. With the knowledge gathered from question 1(b), compute a logistic regression model with respect to different sets of independent features on your training dataset. Remember to categorize all your variables before running the classifier. Please report:

- What is the intercept?

In logistic regression, the baseline category (in this case, category 0, non-diabetic) serves as the reference for comparing the other categories. The intercepts for categories 1 (pre-diabetic) and 2 (diabetic) represent the differences in log-odds between those categories and the reference category when all predictor variables are at their baseline values.

For category 1, the intercept is -7.170653.
For category 2, the intercept is -7.450628.

- What are the coefficients for each of the features? Are they statistically significant?

Here's an example as to how you can interpret the coefficient for "Age3" relative to "Age1" as the reference:
For "Age3" in the Pre-diabetic category: A coefficient of 0.2495164 indicates that, compared to individuals in the reference category "Age1," those in the "Age3" group have 0.25 higher log-odds of being Pre-diabetic.
For "Age3" in the Diabetic category: A coefficient of 0.5623134 suggests that, relative to individuals in the reference category "Age1," those in the "Age3" group have 0.56 higher log-odds of being Diabetic.

Coefficients for the predictor variables:
For category 1:

1. Sex1: 0.1455622
2. Age2: 0.04369554, Age3: 0.2495164, Age4: 0.4485056, Age5: 0.588316, Age6: 1.102796, Age7: 1.209242, Age8: 1.264853, Age9: 1.432004, Age10: 1.406094, Age11: 1.556476, Age12: 1.523066, Age13: 1.485228
3. BMI: 0.06464372
4. COVIDPOS1: 0.0435261, COVIDPOS3: 0.2332161
5. Smoker1: -0.11807522
6. Physical_activity1: 0.001495306
7. Food_shortage1: 0.20604072, Food_shortage3: 0.45621038, Food_shortage4: 0.2340694, Food_shortage5: 0.1001708
8. Average_drink: 0.0002236229
9. MEDCOST11: 0.1141904
10. General_health2: 0.4326584, General_health3: 0.8448622, General_health4: 1.011853, General_health5: 0.9775213

11. Income_category2: 0.01006929, Income_category3: -0.06978511, Income_category4: -0.2017966, Income_category5: -0.09065265, Income_category6: -0.14697270, Income_category7: -0.39874309, Income_category8: -0.54929835, Income_category9: -0.61309778, Income_category10: -0.73147929, Income_category11: -0.9150418
12. Stroke1: -0.02014588
13. Walking_difficulty1: 0.1993777
14. HeartDiseaseorAttack1: 0.2215800
15. Alcohol_consumed: -0.01368659

For category 2:
1. Sex1: 0.5135587
2. Age2: 0.21720720, Age3: 0.5623134, Age4: 0.9501481, Age5: 1.368684, Age6: 1.860338, Age7: 2.110938, Age8: 2.431134, Age9: 2.549917, Age10: 2.711244, Age11: 2.857342, Age12: 2.942166, Age13: 2.735114
3. BMI: 0.07081753
4. COVIDPOS1: 0.1179766, COVIDPOS3: -0.1084739
5. Smoker1: 0.04010873
6. Physical_activity1: -0.170002056
7. Food_shortage1: -0.09827026, Food_shortage3: -0.06189959, Food_shortage4: -0.1708714, Food_shortage5: -0.3014448
8. Average_drink: 0.0010644802
9. MEDCOST11: -0.1062249
10. General_health2: 0.8197880, General_health3: 1.5401960, General_health4: 1.939174, General_health5: 2.0785582
11. Income_category2: -0.08556109, Income_category3: -0.06643024, Income_category4: 0.0354277, Income_category5: 0.02844706, Income_category6: -0.00502135, Income_category7: -0.07446621, Income_category8: -0.07239704, Income_category9: -0.06128878, Income_category10: -0.04058695, Income_category11: -0.2288525
12. Stroke1: 0.25634789
13. Walking_difficulty1: 0.1122935
14. HeartDiseaseorAttack1: 0.4417612
15. Alcohol_consumed: -0.03468563

Statistical significance
p-values for the predictors that were less than 0.05 were shortlisted-
Category 1
1. Sex1 - p-value: 0.006424617
2. Age4 - p-value: 0.03939556, Age5 - p-value: 0.006008582, Age6 - p-value: 7.831794e-08, Age7 - p-value: 2.069732e-09, Age8 - p-value: 3.244887e-10, Age9 - p-value: 3.901324e-13, Age10 - p-value: 1.437517e-12, Age11 - p-value: 6.661338e-15, Age12 - p-value: 2.73559e-13, Age13 - p-value: 3.378631e-12
3. Smoker1 - p-value: 0.02941822
4. Food_shortage3 - p-value: 0.05344686

5. General_health2 - p-value: 2.749548e-05, General_health3 - p-value: 2.220446e-16
6. Income_category7 - p-value: 0.02311896, Income_category8 - p-value: 0.002293702, Income_category9 - p-value: 0.0007167304, Income_category10 - p-value: 0.0002851345, Income_category11 - p-value: 9.199221e-06
7. Walking_difficulty1 - p-value: 0.008687915

For Category 2:
1. Sex1 - p-value: 0
2. Age3 - p-value: 0.00192008, Age4 - p-value: 2.773331e-08, Age5 - p-value: 2.220446e-16, Age6 - p-value: 0, Age7 - p-value: 0, Age8 - p-value: 0, Age9 - p-value: 0, Age10 - p-value: 0, Age11 - p-value: 0, Age12 - p-value: 0, Age13 - p-value: 0
3. Smoker1 - p-value: 0.13971985
4. Physical_activity1 - p-value: 7.823468e-08
5. Food_shortage5 - p-value: 0.006061061
6. General_health2 - p-value: 0, General_health3 - p-value: 0, General_health5 - p-value: 0
7. Income_category4 - p-value: 0.1401623, Income_category5 - p-value: 0.006061061, Income_category6 - p-value: 0, Income_category11 - p-value: 3.899406e-02
8. Walking_difficulty1 - p-value: 0.002704062
9. Alcohol_consumed - p-value: 0

These p-values provide information about the statistical significance of each predictor variable in both Category 1 (Pre-diabetic) and Category 2 (Diabetic). Smaller p-values indicate a stronger statistical significance in the relationship between the predictor variable and the respective category.

- **What are the log-odds and odd ratios of the outcome for a unit increase in each independent variable?**

Log Odds for Multinomial Logistic Regression:

Category 1:	Category 2:
- (Intercept): -7.170653	- (Intercept): -7.450628
- Sex1: 0.1455622	- Sex1: 0.5135587
- Age2: 0.04369554	- Age2: 0.21720720
- Age3: 0.2495164	- Age3: 0.5623134
- Age4: 0.4485056	- Age4: 0.9501481

- Age5: 0.588316
- Age6: 1.102796
- Age7: 1.209242
- Age8: 1.264853
- Age9: 1.432004
- Age10: 1.406094
- Age11: 1.556476
- Age12: 1.523066
- Age13: 1.485228
- BMI: 0.06464372
- COVIDPOS1: 0.0435261
- COVIDPOS3: 0.2332161
- Smoker1: -0.11807522
- Physical_activity1: 0.001495306
- Food_shortage1: 0.20604072
- Food_shortage3: 0.45621038
- Food_shortage4: 0.2340694
- Food_shortage5: 0.1001708
- Average_drink: 0.0002236229
- MEDCOST11: 0.1141904
- General_health2: 0.4326584
- General_health3: 0.8448622
- General_health4: 1.011853
- General_health5: 0.9775213
- Income_category2: 0.01006929
- Income_category3: -0.06978511
- Income_category4: -0.2017966

- Age5: 1.368684
- Age6: 1.860338
- Age7: 2.110938
- Age8: 2.431134
- Age9: 2.549917
- Age10: 2.711244
- Age11: 2.857342
- Age12: 2.942166
- Age13: 2.735114
- BMI: 0.07081753
- COVIDPOS1: 0.1179766
- COVIDPOS3: -0.1084739
- Smoker1: 0.04010873
- Physical_activity1: -0.170002056
- Food_shortage1: -0.09827026
- Food_shortage3: -0.06189959
- Food_shortage4: -0.1708714
- Food_shortage5: -0.3014448
- Average_drink: 0.0010644802
- MEDCOST11: -0.1062249
- General_health2: 0.8197880
- General_health3: 1.5401960
- General_health4: 1.939174
- General_health5: 2.0785582
- Income_category2: -0.08556109
- Income_category3: -0.06643024
- Income_category4: 0.0354277

- Income_category5: -0.09065265 - Income_category5: 0.02844706

- Income_category6: -0.14697270 - Income_category6: -0.00502135

- Income_category7: -0.39874309 - Income_category7: -0.07446621

- Income_category8: -0.54929835 - Income_category8: -0.07239704

- Income_category9: -0.61309778 - Income_category9: -0.06128878

- Income_category10: -0.73147929 - Income_category10: -0.04058695

- Income_category11: -0.9150418 - Income_category11: -0.2288525

- Stroke1: -0.02014588 - Stroke1: 0.25634789

- Walking_difficulty1: 0.1993777 - Walking_difficulty1: 0.1122935

- HeartDiseaseorAttack1: 0.2215800 - HeartDiseaseorAttack1: 0.4417612

- Alcohol_consumed: -0.01368659 - Alcohol_consumed: -0.03468563

These log odds represent the impact of each predictor variable on the likelihood of being in a specific category relative to the reference category.

Odds Ratios for Multinomial Logistic Regression:

Category 1:
- (Intercept): 0.0007688206
- Sex1: 1.156690
- Age2: 1.044664
- Age3: 1.283405
- Age4: 1.565970
- Age5: 1.800953
- Age6: 3.012579
- Age7: 3.350942
- Age8: 3.542574
- Age9: 4.187082
- Age10: 4.07999
- Age11: 4.742082
- Age12: 4.586266
- Age13: 4.415972
- BMI: 1.066779
- COVIDPOS1: 1.044487

Category 2:
- (Intercept): 0.0005810765
- Sex1: 1.671228
- Age2: 1.242602
- Age3: 1.754727
- Age4: 2.586093
- Age5: 3.930173
- Age6: 6.425911
- Age7: 8.255981
- Age8: 11.371766
- Age9: 12.806047
- Age10: 15.04798
- Age11: 17.415179
- Age12: 18.956860
- Age13: 15.411496
- BMI: 1.073385
- COVIDPOS1: 1.125218

- COVIDPOS3: 1.2626544
- Smoker1: 0.8886292
- Physical_activity1: 1.0014964
- Food_shortage1: 1.2288032
- Food_shortage3: 1.5780823
- Food_shortage4: 1.2637322
- Food_shortage5: 1.1053597
- Average_drink: 1.000224
- MEDCOST11: 1.1209655
- General_health2: 1.541350
- General_health3: 2.327657
- General_health4: 2.750692
- General_health5: 2.657860
- Income_category2: 1.0101202
- Income_category3: 0.9325942
- Income_category4: 0.8172611
- Income_category5: 0.9133349
- Income_category6: 0.8633175
- Income_category7: 0.6711631
- Income_category8: 0.5773548
- Income_category9: 0.5416703
- Income_category10: 0.4811966
- Income_category11: 0.4004999
- Stroke1: 0.9800557
- Walking_difficulty1: 1.220643
- HeartDiseaseorAttack1: 1.248047
- Alcohol_consumed: 0.9864066

- COVIDPOS3: 0.8972023
- Smoker1: 1.0409239
- Physical_activity1: 0.8436631
- Food_shortage1: 0.9064039
- Food_shortage3: 0.9399773
- Food_shortage4: 0.8429299
- Food_shortage5: 0.7397486
- Average_drink: 1.001065
- MEDCOST11: 0.8992224
- General_health2: 2.270019
- General_health3: 4.665505
- General_health4: 6.953006
- General_health5: 7.992937
- Income_category2: 0.9179971
- Income_category3: 0.9357282
- Income_category4: 1.0360627
- Income_category5: 1.0288555
- Income_category6: 0.9949912
- Income_category7: 0.9282388
- Income_category8: 0.9301615
- Income_category9: 0.9405516
- Income_category10: 0.9602257
- Income_category11: 0.7954458
- Stroke1: 1.2922022
- Walking_difficulty1: 1.118841
- HeartDiseaseorAttack1: 1.555444
- Alcohol_consumed: 0.9659090

These odds ratios represent the likelihood of being in a specific category relative to the reference category based on the predictor variables.

- **Which are the most predictive features according to the training data?**

Forthe most predictive features we considered the largest absolute coefficient value and lowest p values (less than 0.05)

The most predictive feature for Category 1 is Age group 11

	Predictor	Coefficient	P_Value
Age11	**Age11**	**1.556476**	**6.661338e-15**

The most predictive feature for category 2 is Age group 12

	Predictor	Coefficient	P_Value
Age12	Age12	2.942166	0

This indicates a strong co-relation between old age and diabetes.

- Use the trained model to predict on your testing dataset. Explain your results.

We got an Accuracy score of 89.1 % on the testing dataset.

The accuracy of 89.10% indicates that the model is relatively successful in classifying instances into the appropriate categories. It means that for approximately 89.10% of the cases in the testing dataset, the model's predictions matched the actual class labels (for the variable "Diabetes").

b Classification results using Naïve Bayes.

- Actual vs. Predicted Categories:
  - There are three actual categories: 0, 1, and 2.
  - The model makes predictions for the same three categories: 0, 1, and 2.
- Confusion Matrix:
  - The confusion matrix shows how the model's predictions compare to the actual categories.

```
> print(confusion_matrix)
      Predicted
Actual     0     1     2
     0 23933   472  3386
     1   398    16   165
     2  1676    49  1090
```

- Summary:
  - For class 0:
    - 23,933 instances were correctly predicted as class 0.
    - 472 instances were incorrectly predicted as class 1.
    - 3,386 instances were incorrectly predicted as class 2.
  - For class 1:
    - 398 instances were correctly predicted as class 1.

        - 16 instances were incorrectly predicted as class 0.
        - 165 instances were incorrectly predicted as class 2.
    - For class 2:
        - 1,676 instances were correctly predicted as class 2.
        - 49 instances were incorrectly predicted as class 0.
        - 1,090 instances were incorrectly predicted as class 1.
- Interpretation:
    - The model appears to perform relatively well for class 0, correctly classifying a large number of instances, but it has some difficulty distinguishing between classes 1 and 2.
    - Class 1 has the fewest correct predictions, indicating that the model struggles to correctly identify instances in this class.
    - Class 2 has a moderate number of correct predictions, but there are still misclassifications into other classes.

Accuracy (Naive Bayes without Laplace Smoothing): 0.8

- The accuracy of 0.8 means that the model correctly predicted the class labels for 80% of the instances in the test dataset.

- **Classification results using Naïve Bayes with the Laplace estimator. Do the results improve?**

Incorporating the Laplace estimator in our Naive Bayes classifier did not result in improved performance. Both accuracy and the confusion matrix remained consistent, indicating that the Laplace smoothing did not enhance predictive capabilities.

- Confusion Matrix: The confusion matrix shows how the model's predictions compare to the actual categories.

```
> print(confusion_matrix_laplace)
      Predicted
Actual     0     1     2
     0 23933   472  3386
     1   398    16   165
     2  1676    49  1090
```

- Actual vs. Predicted Categories:
    - There are three actual categories: 0, 1, and 2.
    - The model makes predictions for the same three categories: 0, 1, and 2.
- Summary:
    - For class 0:

        - 23,933 instances were correctly predicted as class 0.
        - 472 instances were incorrectly predicted as class 1.
        - 3,386 instances were incorrectly predicted as class 2.
    - For class 1:
        - 398 instances were correctly predicted as class 1.
        - 16 instances were incorrectly predicted as class 0.
        - 165 instances were incorrectly predicted as class 2.
    - For class 2:
        - 1,676 instances were correctly predicted as class 2.
        - 49 instances were incorrectly predicted as class 0.
        - 1,090 instances were incorrectly predicted as class 1.
- Interpretation:
    - The model appears to perform relatively well for class 0, correctly classifying a large number of instances, but it has some difficulty distinguishing between classes 1 and 2.
    - Class 1 has the fewest correct predictions, indicating that the model struggles to correctly identify instances in this class.
    - Class 2 has a moderate number of correct predictions, but there are still misclassifications into other classes.

Accuracy (Naive Bayes with Laplace Smoothing): 0.8

- The accuracy of 0.8 means that the model correctly predicted the class labels for 80% of the instances in the test dataset.

Question 3. Decision Trees and Random Forests.

a. Split your dataset into training and testing sets assuming that samples are not randomly ordered (Hint: generate random numbers). Show that the distribution after the split is similar to the original.

```
Original Distribution:
0.0    0.890879
2.0    0.090237
1.0    0.018884
Name: Diabetes, dtype: float64
Training Set Distribution:
0.0    0.891052
2.0    0.089948
1.0    0.019000
Name: Diabetes, dtype: float64
Testing Set Distribution:
0.0    0.890188
```

As we can see from the above screenshot the distribution of Non-diabetic (0), pre-diabetic (1) and Diabetic (2) is roughly the same across original, training and testing data.

b . Train a decision tree and interpret the main resulting if-then rules (together with their corresponding plots). Test the trained tree with your testing dataset and compare the confusion matrices obtained during training and testing: compute percentages of correctly and incorrectly classified samples and compare results in training and testing.

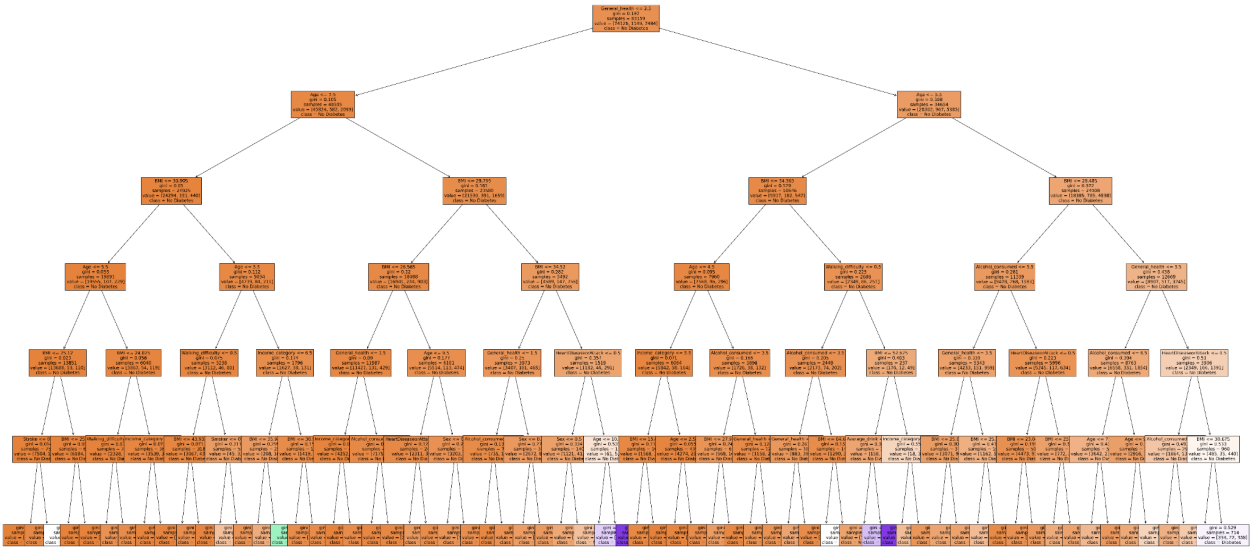


Fig. Decision Tree

The above decision tree starts off by considering general health as the root node. The decision tree decided this feature to be the root node based on the gini score which is a function that determines how well a decision tree was split.

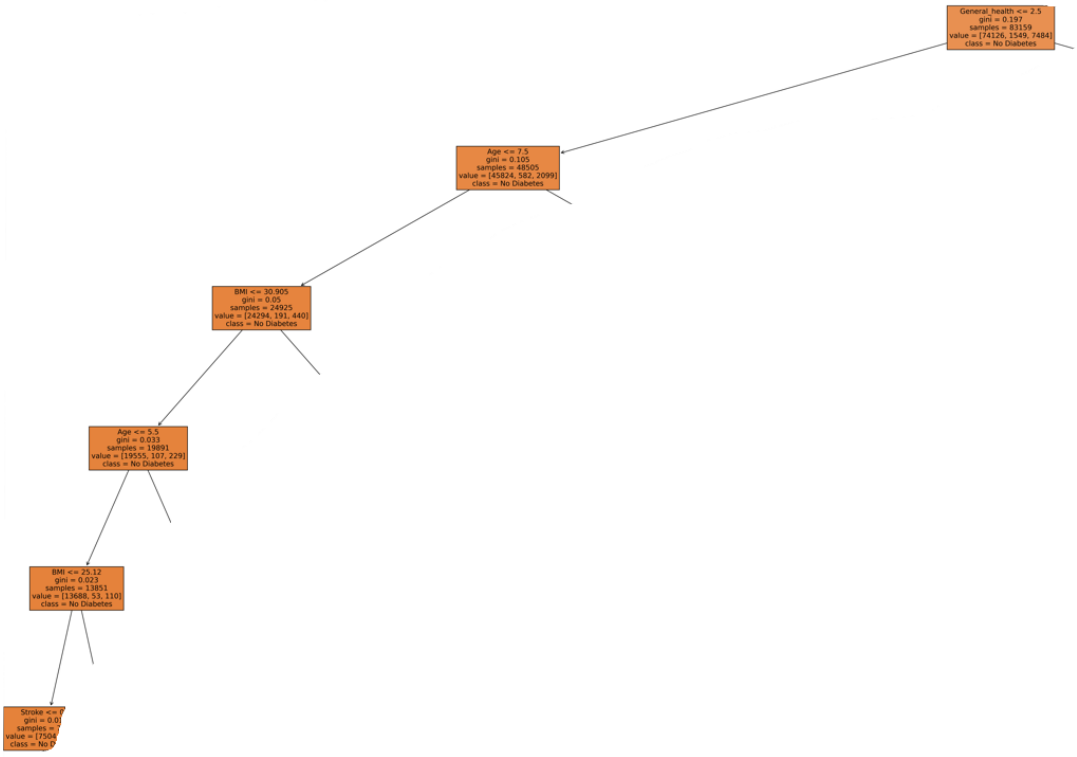


Fig. Left nodes of the decision tree

If General_health is less than or equal to 2.50, the tree proceeds to the left. As shown in the above figure.

Second node (left branch of the first node):
Feature tested: Age
If Age is less than or equal to 7.50, the tree proceeds to the left.

Third node (left branch of the second node):
Feature tested: BMI
If BMI is less than or equal to 30.90, the tree proceeds to the left.

Fourth node (left branch of the third node):
Feature tested: Age
If Age is less than or equal to 5.50, the tree proceeds to the left.

Fifth node (left branch of the fourth node):
Feature tested: BMI
If BMI is less than or equal to 25.12, the tree proceeds to the left.

Sixth node (left branch of the fifth node):
Feature tested: Stroke
If Stroke is less than or equal to 0.50, the predicted class is 0 (non-diabetic).

We experimented with various max_depth values and discovered that the highest testing accuracy of 89 percent was achieved when we reduced the depth of the decision tree to 6. This adjustment was made to mitigate overfitting, which was occurring with our original model, when the tree was too deep.

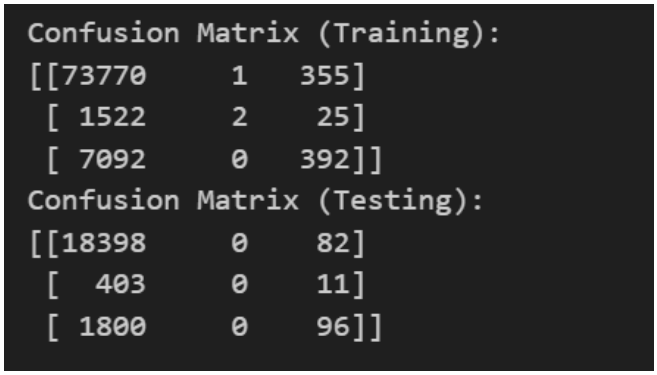


Fig. Confusion matrices

```
Accuracy (Training): 89.18%
Accuracy (Testing): 88.96%
Classification Report (Training):
              precision    recall  f1-score   support

 No Diabetes       0.90      1.00      0.94     74126
Pre-Diabetes       0.67      0.00      0.00      1549
    Diabetes       0.51      0.05      0.09      7484

    accuracy                           0.89     83159
   macro avg       0.69      0.35      0.35     83159
weighted avg       0.86      0.89      0.85     83159

Classification Report (Testing):
              precision    recall  f1-score   support

 No Diabetes       0.89      1.00      0.94     18480
Pre-Diabetes       0.00      0.00      0.00       414
    Diabetes       0.51      0.05      0.09      1896

    accuracy                           0.89     20790
   macro avg       0.47      0.35      0.34     20790
weighted avg       0.84      0.89      0.85     20790
```

Fig. Accuracy Training vs Testing

Precision measures the accuracy of positive predictions made by a model. When our model predicts that a person does not have Diabetes it is correct 89 percent of the time.

Recall measures the model's ability to identify all relevant instances in the dataset. In our case it means that the model correctly identified 94 percent of the people as non diabetic (missing 6 percent).

However the precision and recall values for identifying people with diabetes is quite low. 51 percent and 9 percent respectively.

c. Apply boosting with different numbers of trees and analyze the impact on the prediction results: what is the impact of the number of trees in the accuracy of the classifier?

```
Number of Trees | Accuracy
50 | 0.8898
100 | 0.8899
150 | 0.8901
200 | 0.8901

Classification Reports:
Number of Trees: 50
              precision    recall  f1-score   support

 No Diabetes       0.89      0.99      0.94     18480
Pre-Diabetes       0.00      0.00      0.00       414
    Diabetes       0.50      0.07      0.13      1896

    accuracy                           0.89     20790
   macro avg       0.46      0.36      0.36     20790
weighted avg       0.84      0.89      0.85     20790

---------------------------------------
Number of Trees: 100
              precision    recall  f1-score   support

 No Diabetes       0.90      0.99      0.94     18480
Pre-Diabetes       0.00      0.00      0.00       414
    Diabetes       0.51      0.07      0.13      1896
...
   macro avg       0.47      0.36      0.36     20790
weighted avg       0.84      0.89      0.85     20790
```

```
Number of Trees: 150
              precision    recall  f1-score   support

 No Diabetes       0.90      0.99      0.94     18480
Pre-Diabetes       0.00      0.00      0.00       414
    Diabetes       0.51      0.08      0.13      1896

    accuracy                           0.89     20790
   macro avg       0.47      0.36      0.36     20790
weighted avg       0.84      0.89      0.85     20790

---------------------------------------
Number of Trees: 200
              precision    recall  f1-score   support

 No Diabetes       0.90      0.99      0.94     18480
Pre-Diabetes       0.00      0.00      0.00       414
    Diabetes       0.51      0.08      0.13      1896

    accuracy                           0.89     20790
   macro avg       0.47      0.36      0.36     20790
weighted avg       0.84      0.89      0.85     20790
```

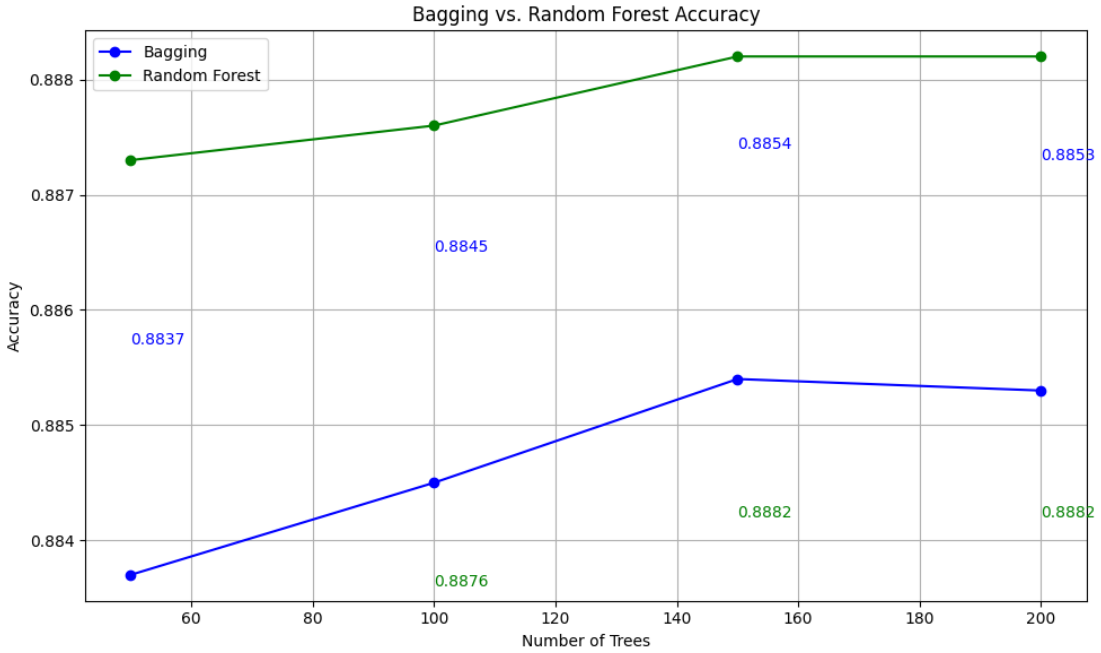
We can see that as the number of trees increases the accuracy of the model also increases slightly (from 88.9 percent to 89.01 percent).

We can also observe an increase in the precision and recall values for the Diabetes class (from 50% precision to 51% and 7% recall to 8%). This happens because adding more trees allows

the boosting algorithm to focus on and correct the mistakes made by the previous trees thereby improving the model.

d. Do the same analysis with bagging and random forests: train with bagging and then train with random forests using at least four different numbers of trees. Compare prediction results over the testing sets. Which are the most important features in each random forest?

The testing accuracy increases for both bagging and random forests when we increase the number of trees. Random forests consistently outperformed bagging.



Number of Trees: 50

Random Forest Feature Importance:

BMI: 0.2925

Age: 0.1268

Alcohol_consumed: 0.1257

Income_category: 0.1187

Number of Trees: 100

Random Forest Feature Importance:

BMI: 0.2925

Age: 0.1266

Alcohol_consumed: 0.1264

Income_category: 0.1190

Average_drink: 0.0760

Number of Trees: 150

Random Forest Feature Importance:

BMI: 0.2931

Age: 0.1280

Alcohol_consumed: 0.1257

Income_category: 0.1184

Average_drink: 0.0763

Number of Trees: 200

Random Forest Feature Importance:

BMI: 0.2932

Age: 0.1278

Alcohol_consumed: 0.1253

Income_category: 0.1183

Average_drink: 0.0763

Question 4. Comparative Analysis.

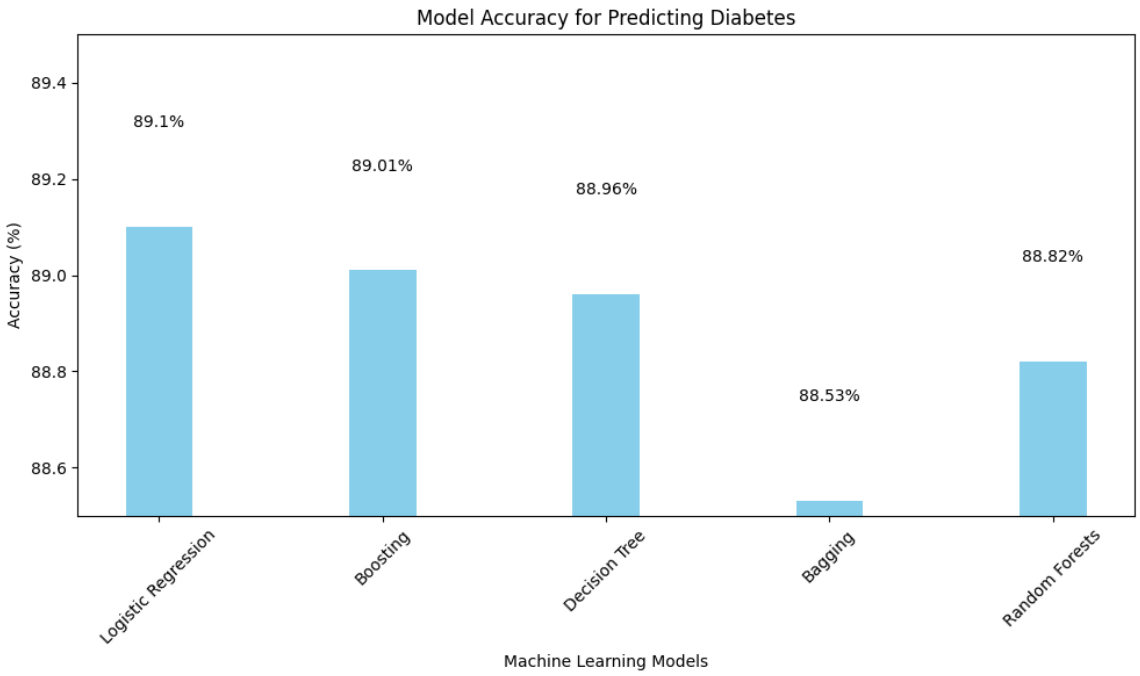
Write a summary of all classifiers, their predictive quality and which one would you use to answer your research question(s).

Research Question 2- Predicting diabetes, a classification problem

Logistic regression accuracy - 89.1%

Boosting accuracy - 89.01% (ensemble of 200 trees)
Decision Tree accuracy - 88.96%

Bagging accuracy - 88.53% (ensemble of 200 trees)
Random forests accuracy - 88.82% (ensemble of 200 trees)



These results suggest that both Logistic Regression and Boosting outperform the other models in terms of accuracy. Logistic Regression achieved the highest accuracy at 89.1%, closely followed by Boosting at 89.01%. Decision Tree, Bagging, and Random Forests also performed well but with slightly lower accuracy scores.

The similar accuracy levels of Boosting and Logistic Regression suggest that both could be suitable for predicting diabetes. Boosting might offer better performance when dealing with complex relationships in the data, but it can be computationally expensive. Logistic Regression, on the other hand, is interpretable and relatively computationally efficient.

Contribution:

Deliverables	Dhiraj	Sakshi	Shashank

Q1(Linear Regression) Code, Report, Presentation		100%	
Q2(Logistic Regression) Code, Report, Presentation	100%		
Q3(Decision Tree) Code, Report, Presentation			100%