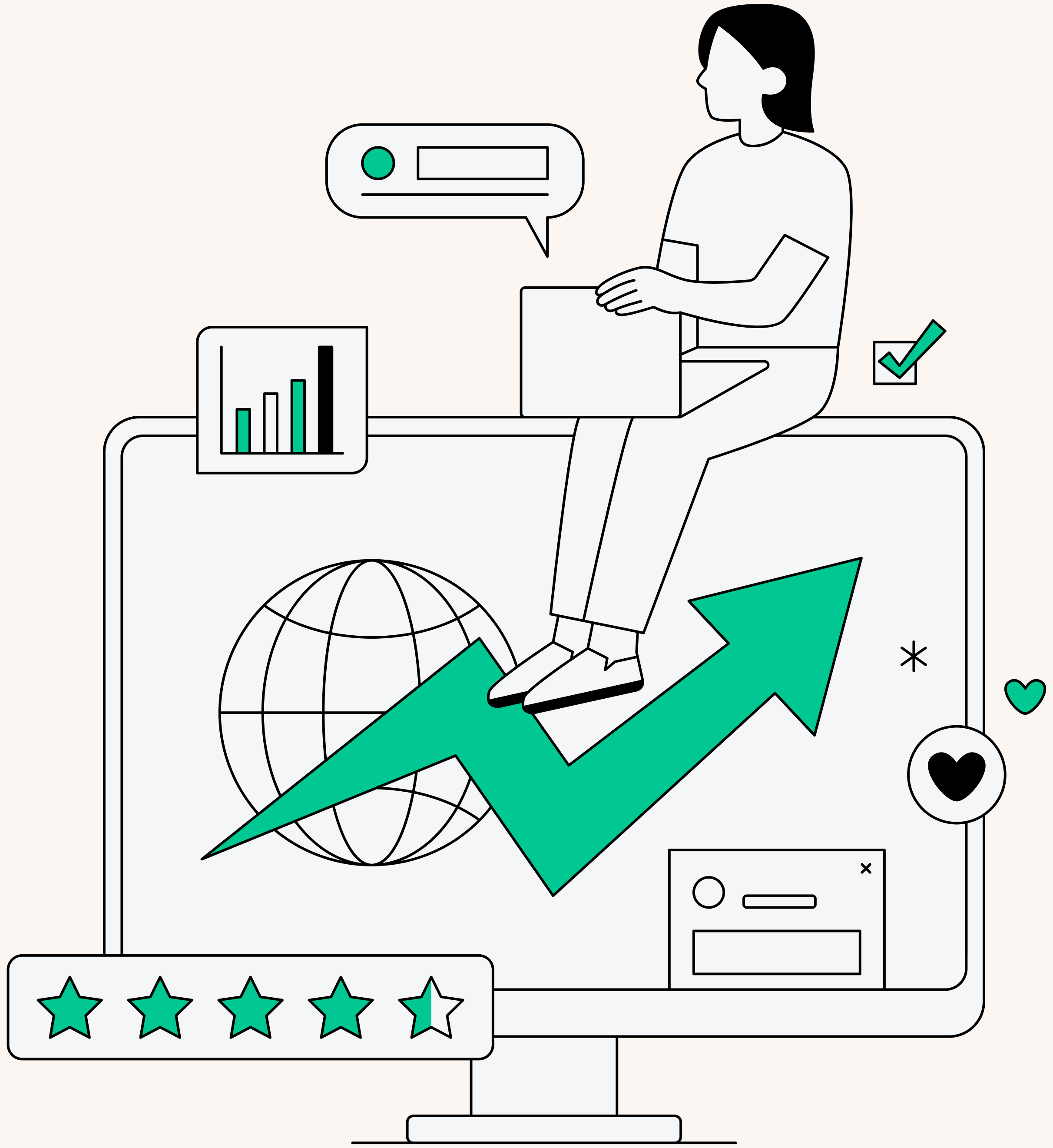
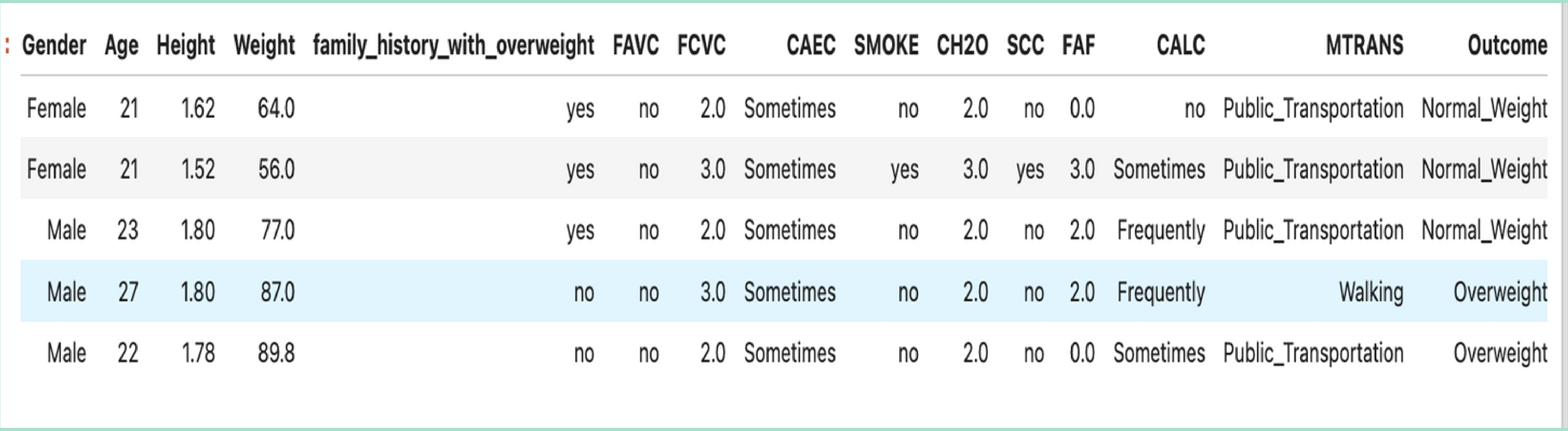
SCM 516 TEAM PROJECT

Presented by

Vinit Vijaykumar Adke,
Dhirraj Suresh Kumar,
Yash Sanjaykumar Pardeshi,
Jaswant Giridharagopalan



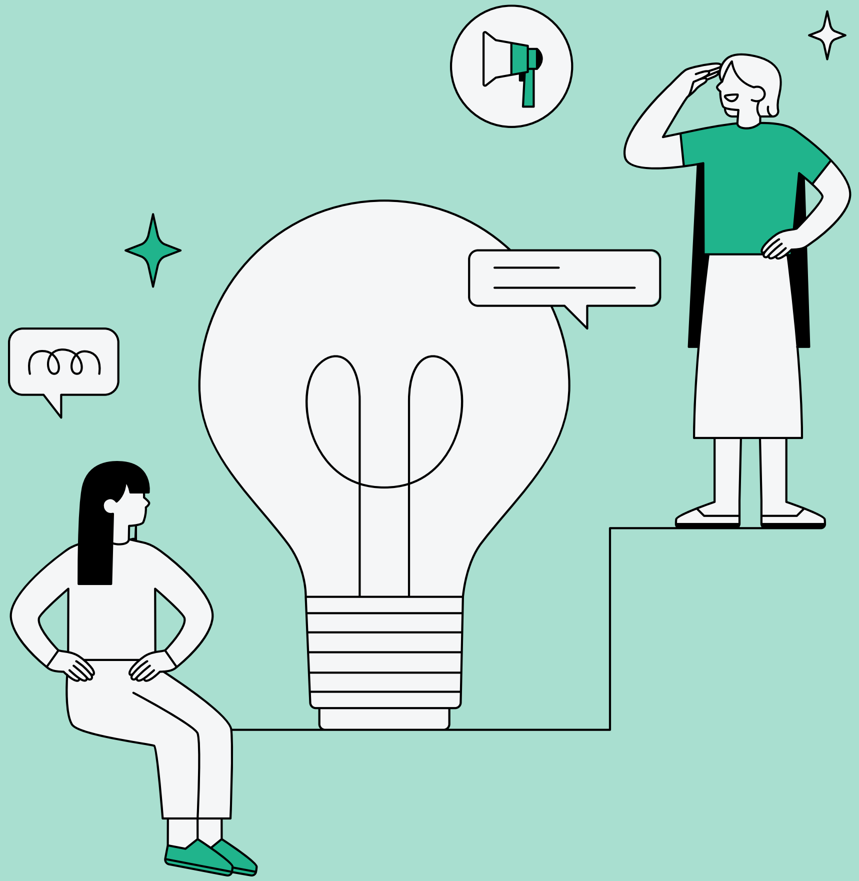
CLASSIFICATION DATASET



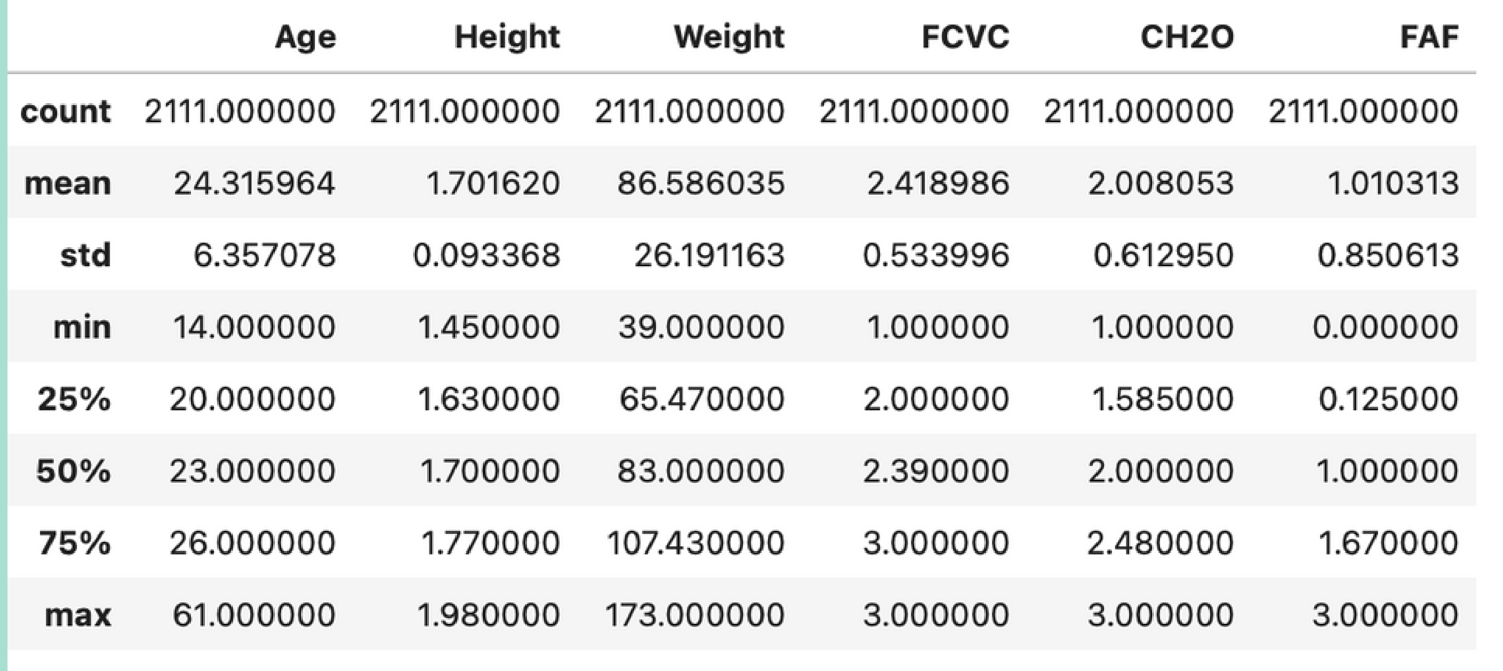
Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	CAEC	SMOKE	CH2O	SCC	FAF	CALC	MTRANS	Outcome
Female	21	1.62	64.0	yes	no	2.0	Sometimes	no	2.0	no	0.0	no	Public_Transportation	Normal_Weight
Female	21	1.52	56.0	yes	no	3.0	Sometimes	yes	3.0	yes	3.0	Sometimes	Public_Transportation	Normal_Weight
Male	23	1.80	77.0	yes	no	2.0	Sometimes	no	2.0	no	2.0	Frequently	Public_Transportation	Normal_Weight
Male	27	1.80	87.0	no	no	3.0	Sometimes	no	2.0	no	2.0	Frequently	Walking	Overweight
Male	22	1.78	89.8	no	no	2.0	Sometimes	no	2.0	no	0.0	Sometimes	Public_Transportation	Overweight

Here are common descriptions for the variables in the dataset.

FAVC - If the person frequently consumes high-calorie foods (yes/no)
FCVC - Frequency of Vegetable consumption (Scale 1-3)
CAEC - Frequency of Consuming foods between meals(Never,Sometimes, Frequently, Always)
CH2O- Daily Water Intake (Scale 1-3)
SCC - If the person monitors their calorie intake(Yes, No)
FAF - Physical activity frequency(Scale 0 - 4)
CALC - Frequency of alcohol consumption (Never,Sometimes, Frequently, Always)
MTRANS - Mode of Transportation(Bike, Motorbike, Public_Transport, Automobile, Walking)
Outcome - Person falls in which category(Insufficient_Weight, Normal_Weight, Obesity, Overweight)



STATISTICAL DESCRIPTION



	Age	Height	Weight	FCVC	CH2O	FAF
count	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000
mean	24.315964	1.701620	86.586035	2.418986	2.008053	1.010313
std	6.357078	0.093368	26.191163	0.533996	0.612950	0.850613
min	14.000000	1.450000	39.000000	1.000000	1.000000	0.000000
25%	20.000000	1.630000	65.470000	2.000000	1.585000	0.125000
50%	23.000000	1.700000	83.000000	2.390000	2.000000	1.000000
75%	26.000000	1.770000	107.430000	3.000000	2.480000	1.670000
max	61.000000	1.980000	173.000000	3.000000	3.000000	3.000000

Age - Mean (24 years) - The average age is relatively young, indicating a younger demographic in the study. Min - Max Range - Most individuals are in the 20-26 age range.

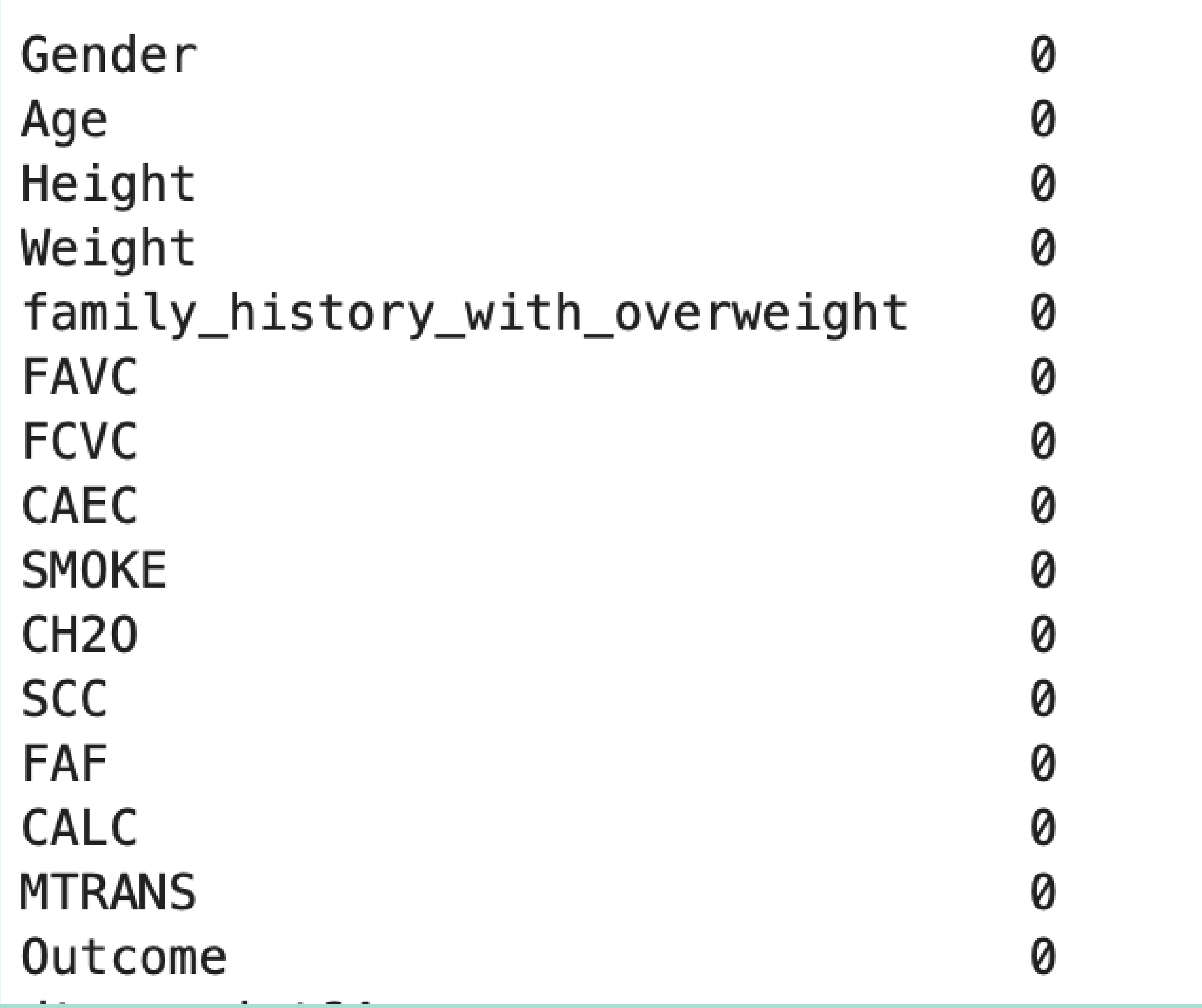
Weight - Mean (86kgs) - Suggesting overweight or obese individual sample.

FCVC - Mean (2.42) - On average, participants consume vegetables moderately.

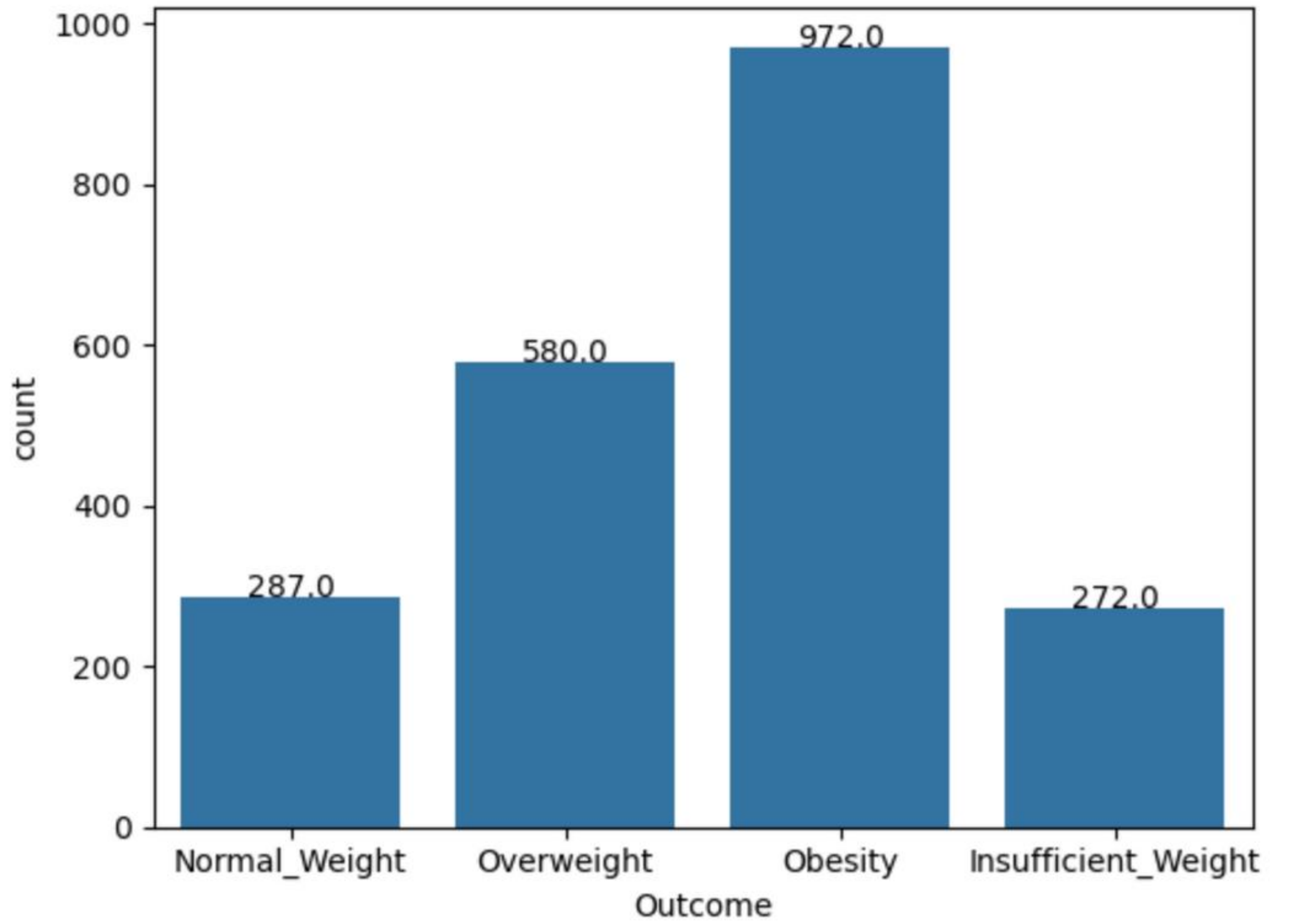
CH2O - Mean (2.01) - This suggests an average daily water intake of about 2 liters.

FAF - Mean (1.01) - Participants engage in low physical activity.

NULL VALUES & UNBALANCED DATASET

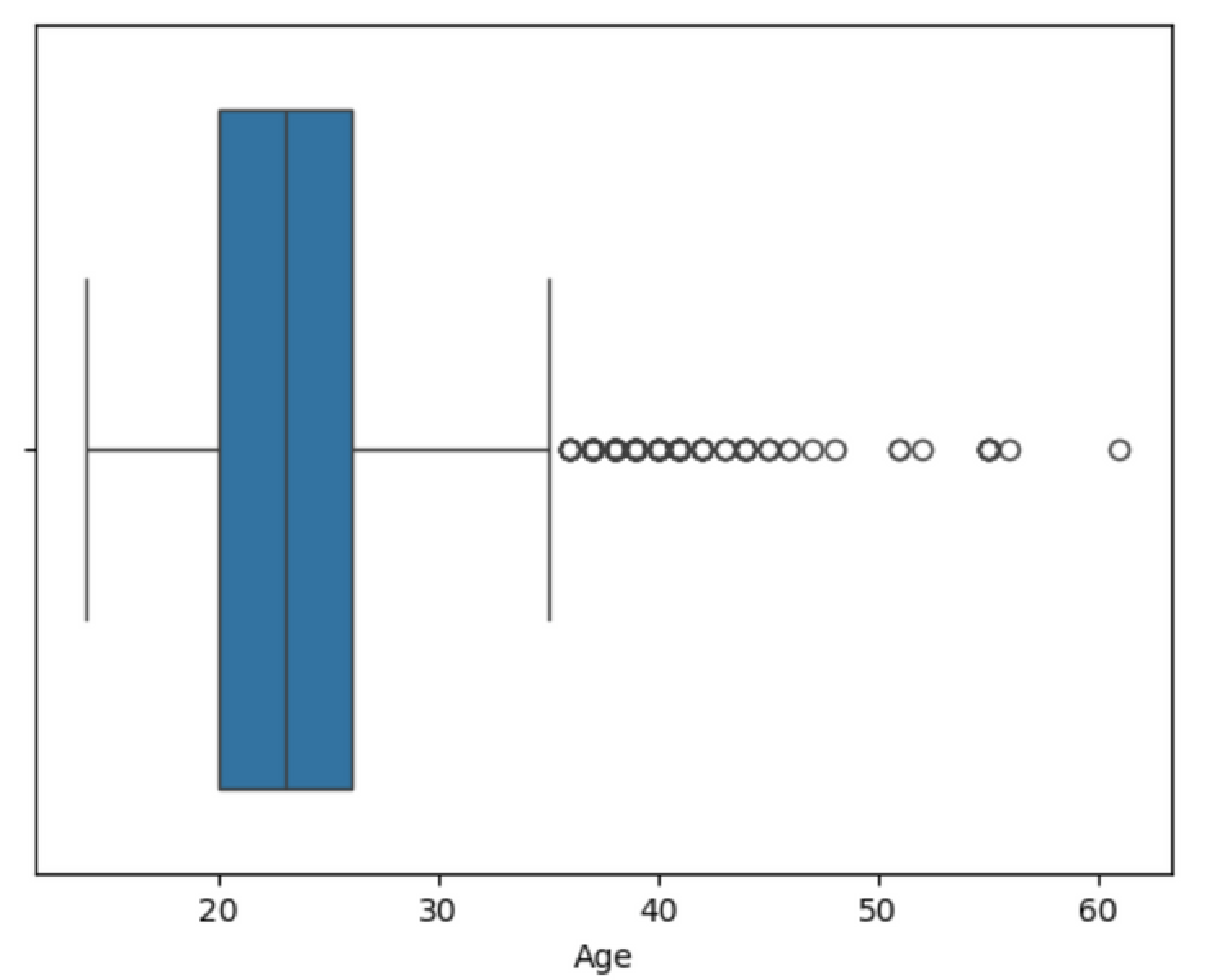


```
Gender                            0
Age                               0
Height                            0
Weight                            0
family_history_with_overweight    0
FAVC                              0
FCVC                              0
CAEC                              0
SMOKE                             0
CH2O                              0
SCC                               0
FAF                               0
CALC                              0
MTRANS                            0
Outcome                           0
```

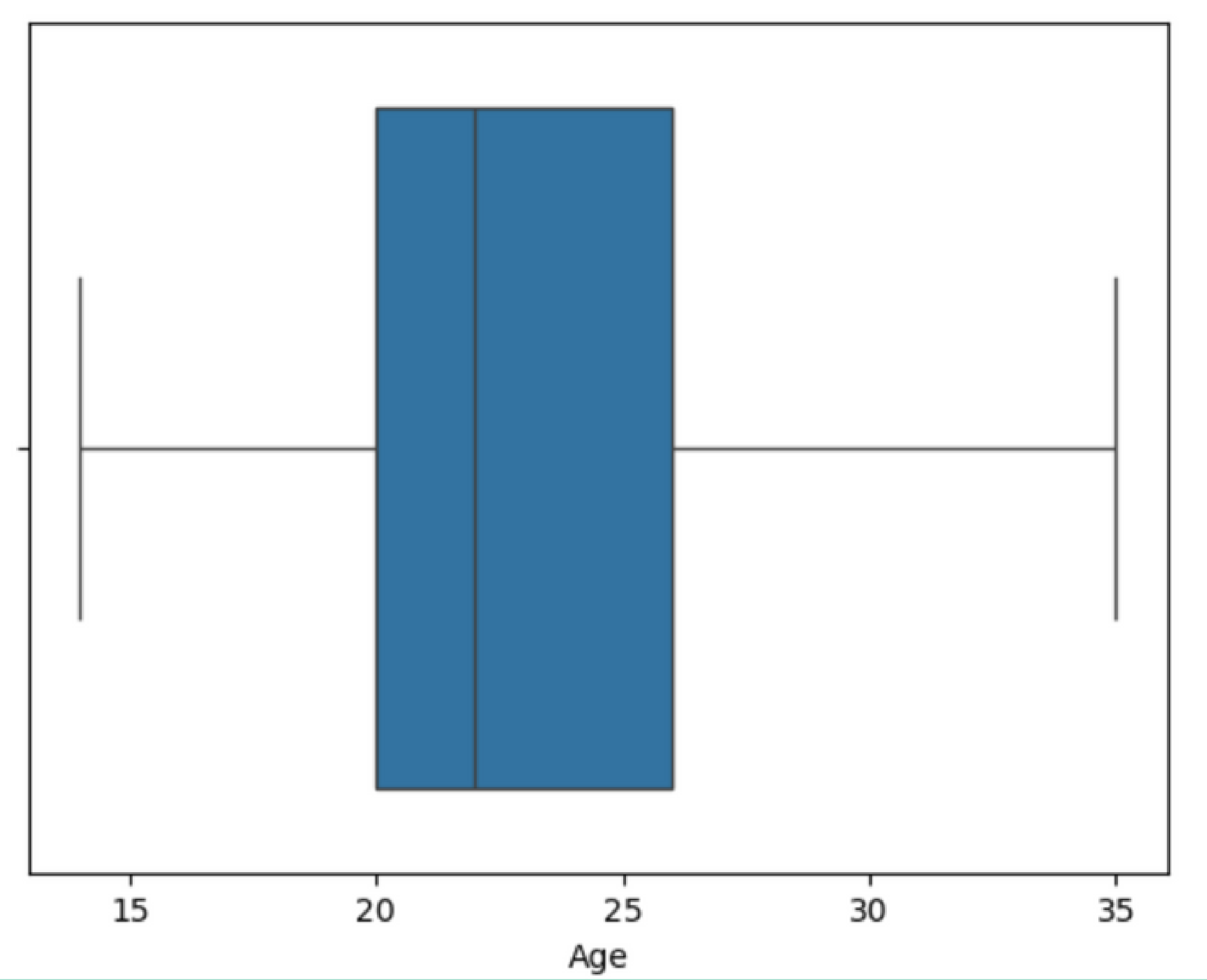


OUTLIER REMOVAL FOR AGE

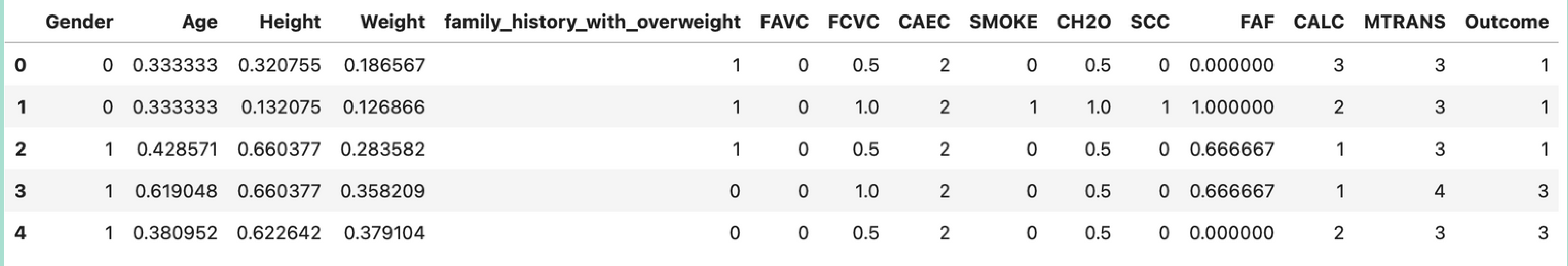
Age with outlier



Age after Outlier Removal



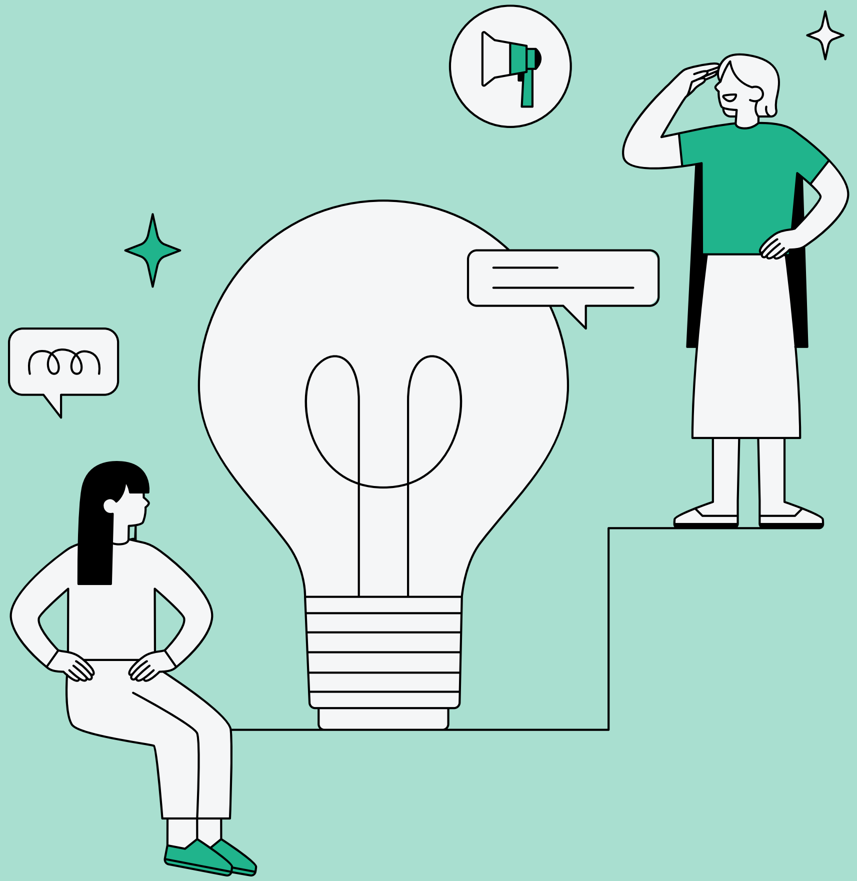
ENCODED & STANDARDIZED DATASET



	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	CAEC	SMOKE	CH2O	SCC	FAF	CALC	MTRANS	Outcome
0	0	0.333333	0.320755	0.186567	1	0	0.5	2	0	0.5	0	0.000000	3	3	1
1	0	0.333333	0.132075	0.126866	1	0	1.0	2	1	1.0	1	1.000000	2	3	1
2	1	0.428571	0.660377	0.283582	1	0	0.5	2	0	0.5	0	0.666667	1	3	1
3	1	0.619048	0.660377	0.358209	0	0	1.0	2	0	0.5	0	0.666667	1	4	3
4	1	0.380952	0.622642	0.379104	0	0	0.5	2	0	0.5	0	0.000000	2	3	3

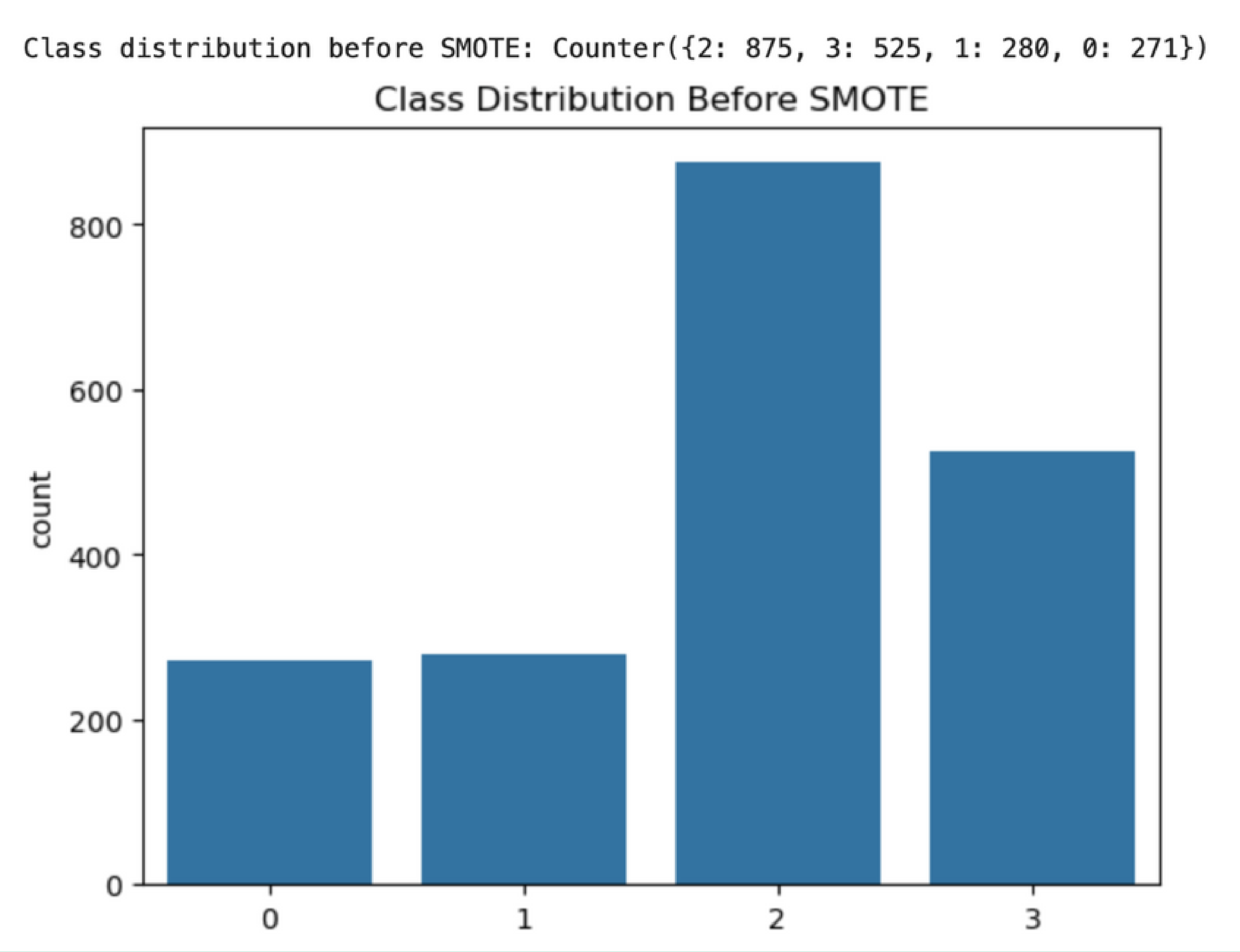
Encoded Dataset

Gender - (Female - 0, Male - 1)
family_history_with_overweight - (Yes - 1, No - 0)
FAVC - (Yes - 1, No - 0)
CAEC - (Never - 0, Sometimes - 1, Frequently - 2, Always - 3)
SMOKE - (Yes - 1, No - 0)
SCC - (Yes - 1, No - 0)
CALC - (Never - 0, Sometimes - 1, Frequently - 2, Always - 3)
MTRANS - (Bike - 0, Motorbike - 1, Public_Transport - 2, Automobile - 3, Walking - 4)



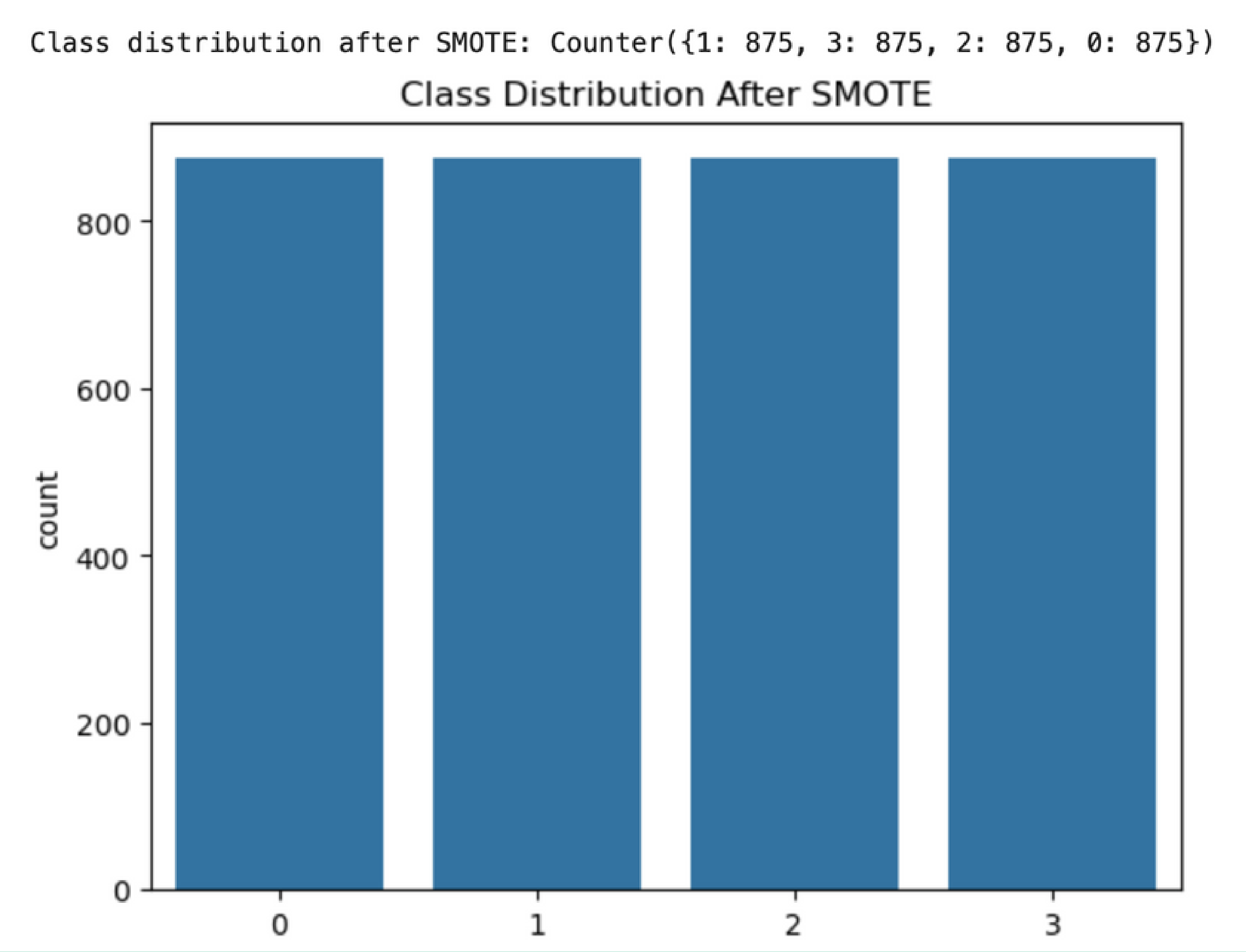
SMOTE Sampling

Before Sampling



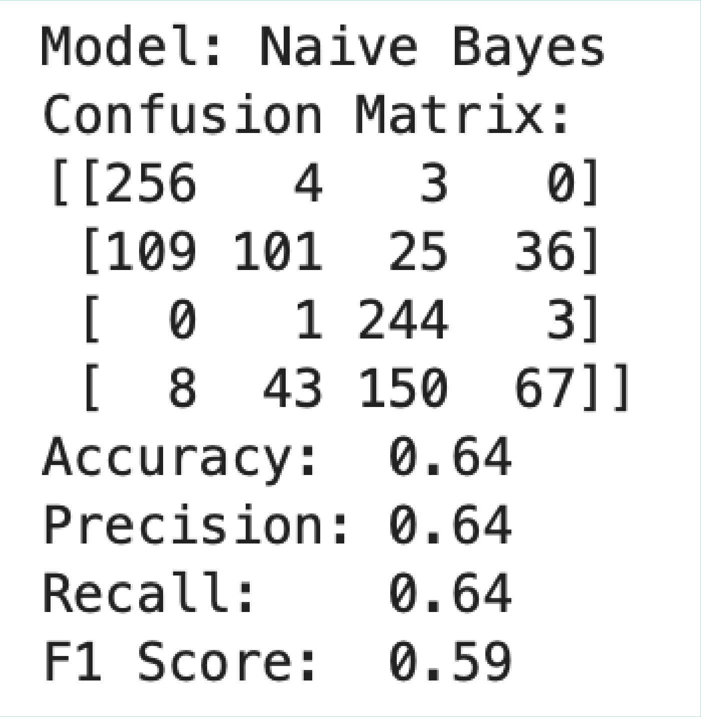
```
Class distribution before SMOTE: Counter({2: 875, 3: 525, 1: 280, 0: 271})
```

After Sampling

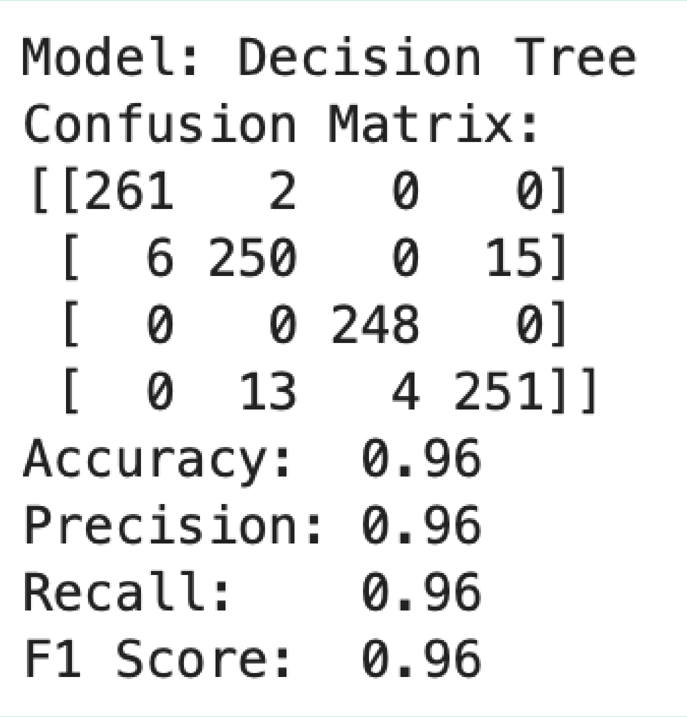


```
Class distribution after SMOTE: Counter({1: 875, 3: 875, 2: 875, 0: 875})
```

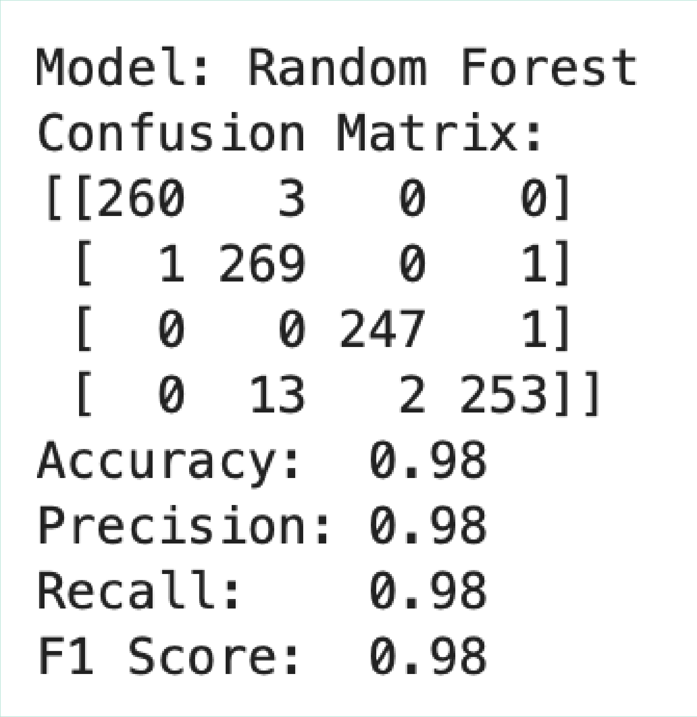
CONFUSION MATRIX - RESULTS



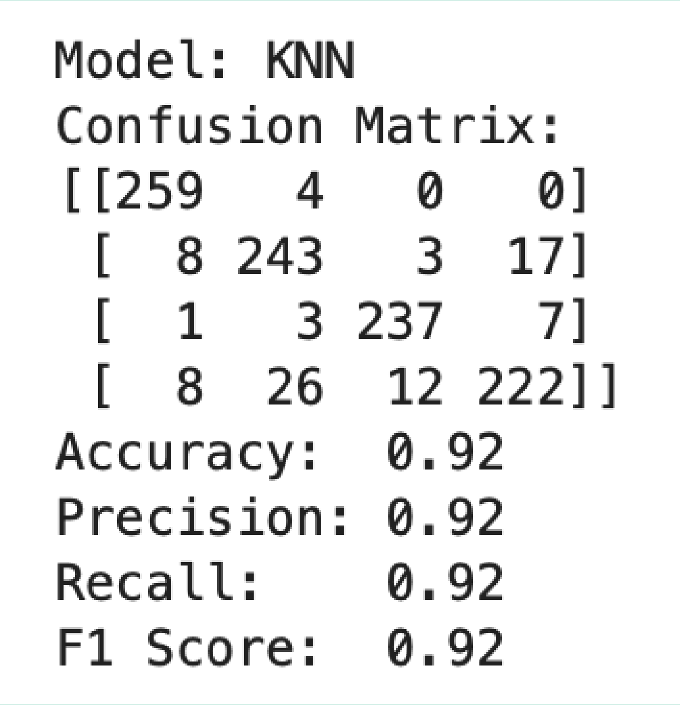
```
Model: Naive Bayes
Confusion Matrix:
[[256   4   3   0]
 [109 101  25  36]
 [  0   1 244   3]
 [  8  43 150  67]]
Accuracy:  0.64
Precision: 0.64
Recall:    0.64
F1 Score:  0.59
```



```
Model: Decision Tree
Confusion Matrix:
[[261   2   0   0]
 [  6 250   0  15]
 [  0   0 248   0]
 [  0  13   4 251]]
Accuracy:  0.96
Precision: 0.96
Recall:    0.96
F1 Score:  0.96
```

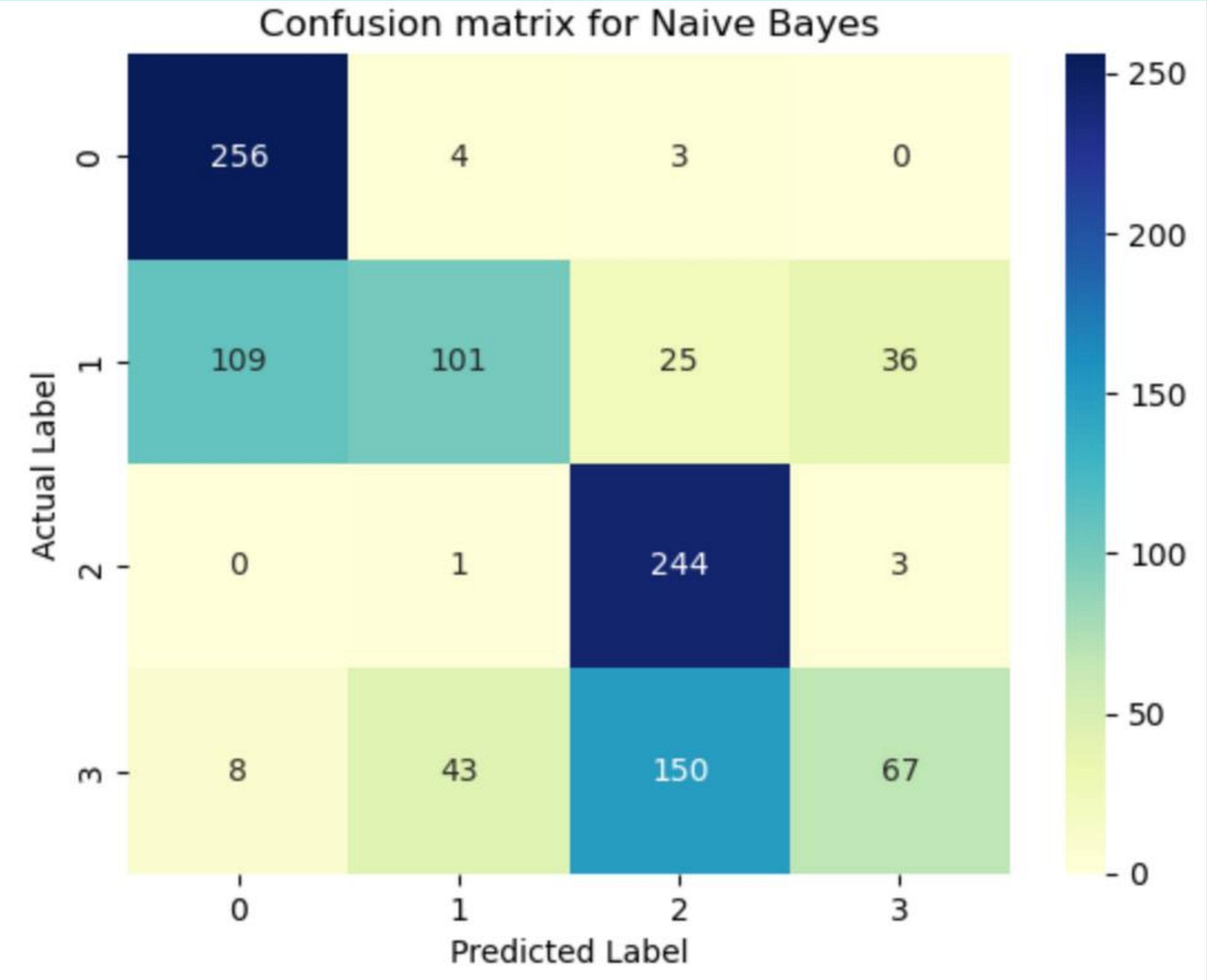


```
Model: Random Forest
Confusion Matrix:
[[260   3   0   0]
 [  1 269   0   1]
 [  0   0 247   1]
 [  0  13   2 253]]
Accuracy:  0.98
Precision: 0.98
Recall:    0.98
F1 Score:  0.98
```



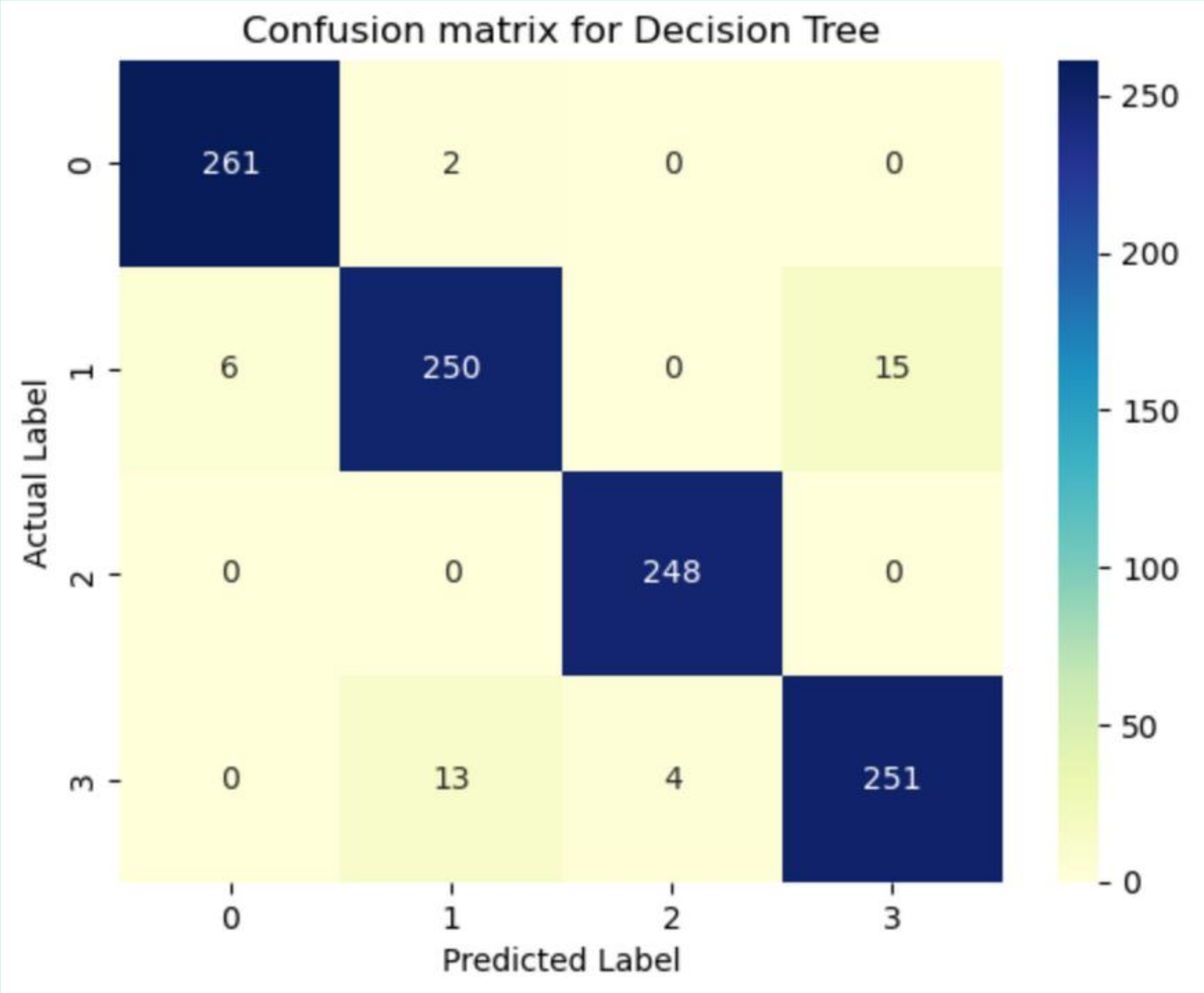
```
Model: KNN
Confusion Matrix:
[[259   4   0   0]
 [  8 243   3  17]
 [  1   3 237   7]
 [  8  26  12 222]]
Accuracy:  0.92
Precision: 0.92
Recall:    0.92
F1 Score:  0.92
```

CONFUSION MATRIX - PLOTS



Confusion matrix for Naive Bayes

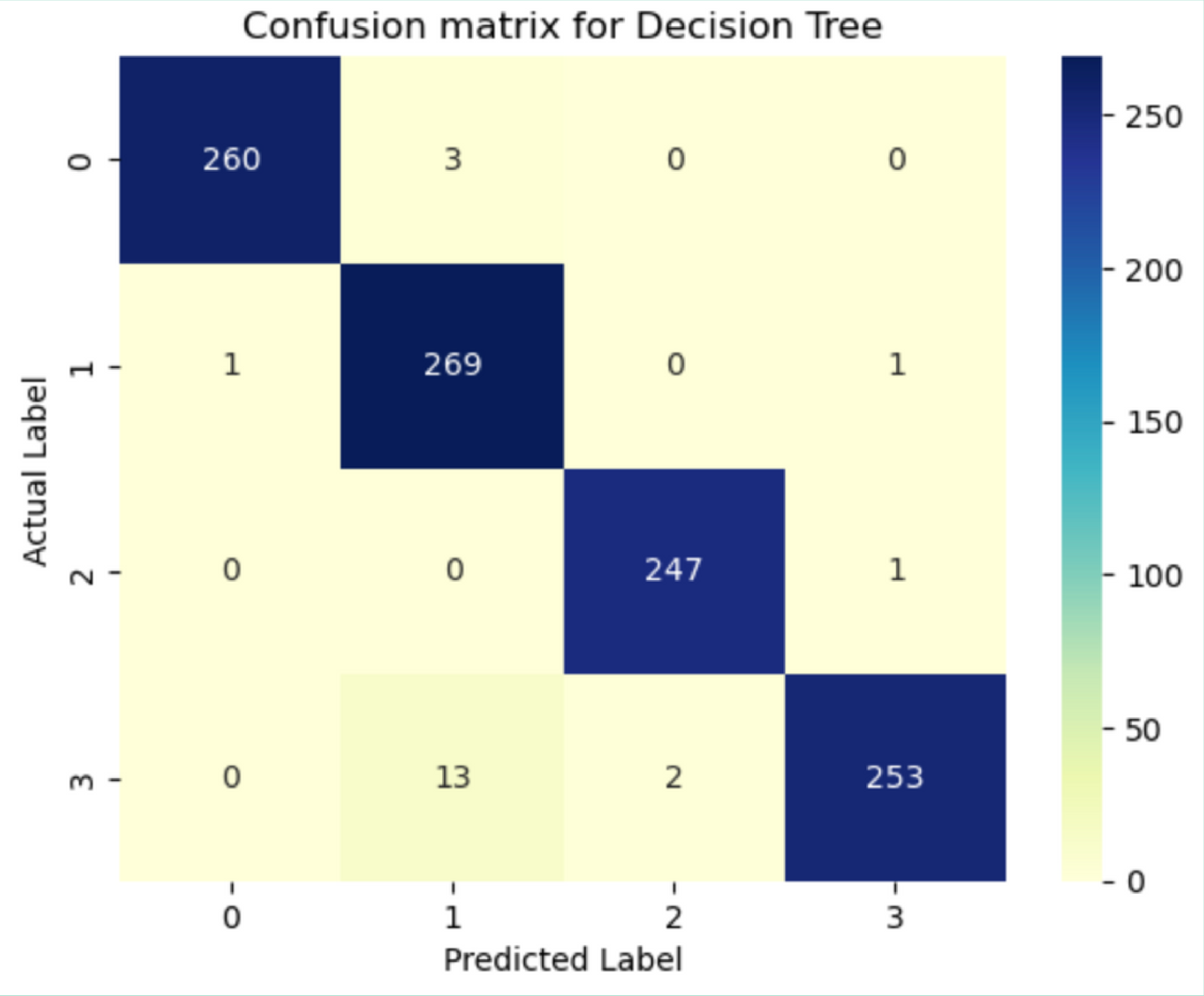
Actual Label \ Predicted Label	0	1	2	3
0	256	4	3	0
1	109	101	25	36
2	0	1	244	3
3	8	43	150	67

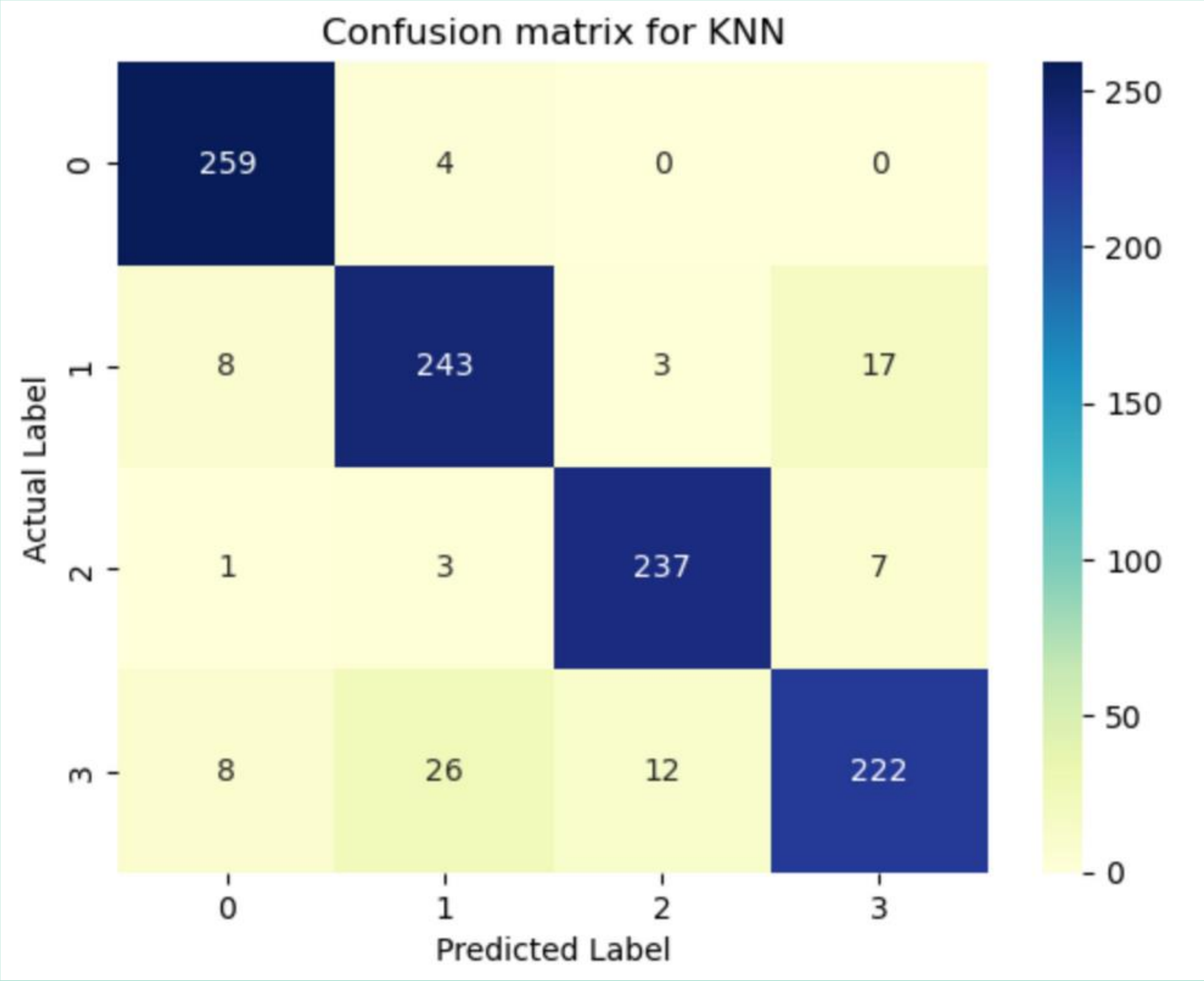


Confusion matrix for Decision Tree

Actual Label \ Predicted Label	0	1	2	3
0	261	2	0	0
1	6	250	0	15
2	0	0	248	0
3	0	13	4	251

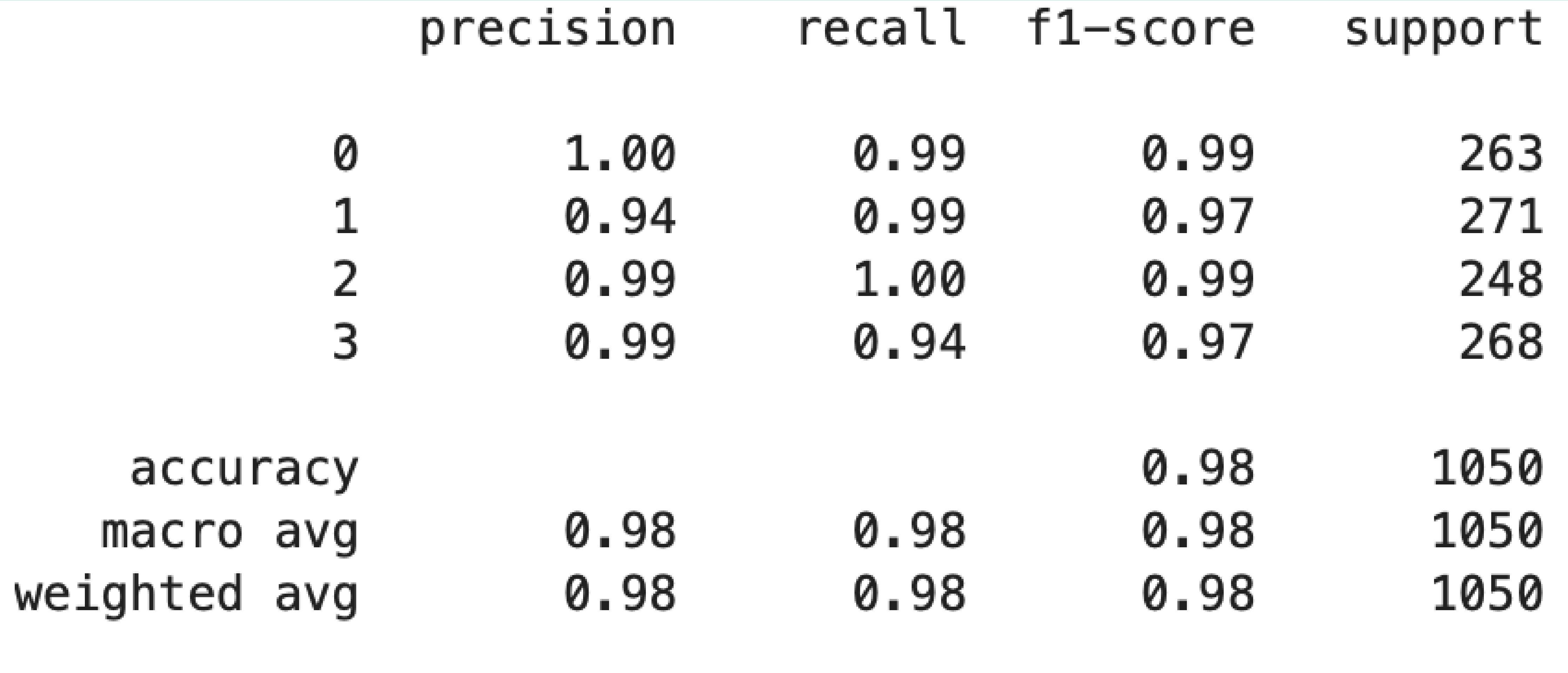
CONFUSION MATRIX - PLOTS


Confusion matrix for Decision Tree


Confusion matrix for KNN

BEST PERFORMANCE - RANDOM FOREST

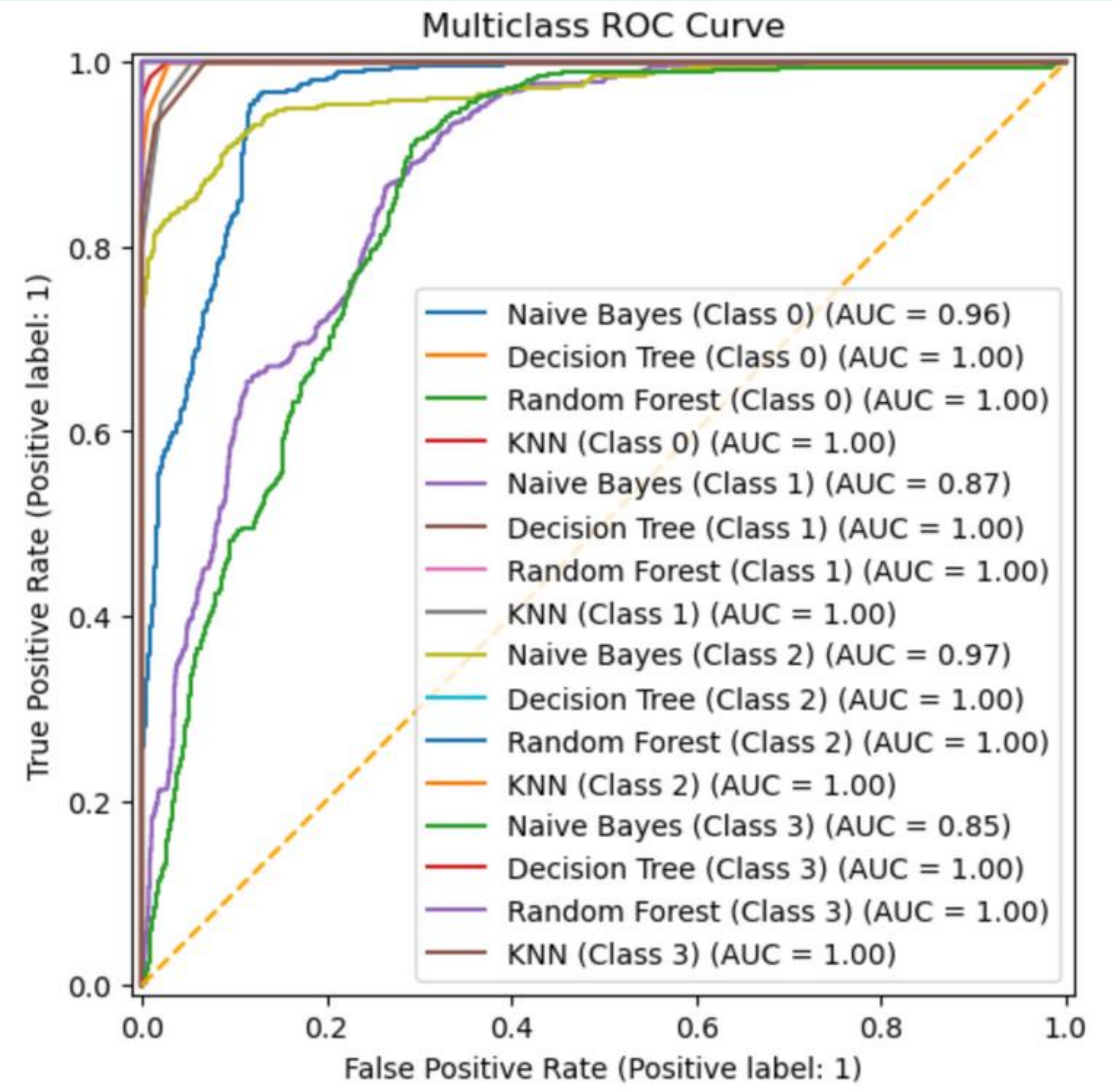
- Class 0: Perfect precision (1.00) and nearly perfect recall (0.99), with 263 samples
- Class 1: Very high precision (0.94) and nearly perfect recall (0.99), with 271 samples
- Class 2: Nearly perfect precision (0.99) and perfect recall (1.00), with 248 samples
- Class 3: Nearly perfect precision (0.99) and high recall (0.94), with 268 samples
- F1-scores are excellent across all classes (0.97-0.99), indicating a strong balance between precision and recall.



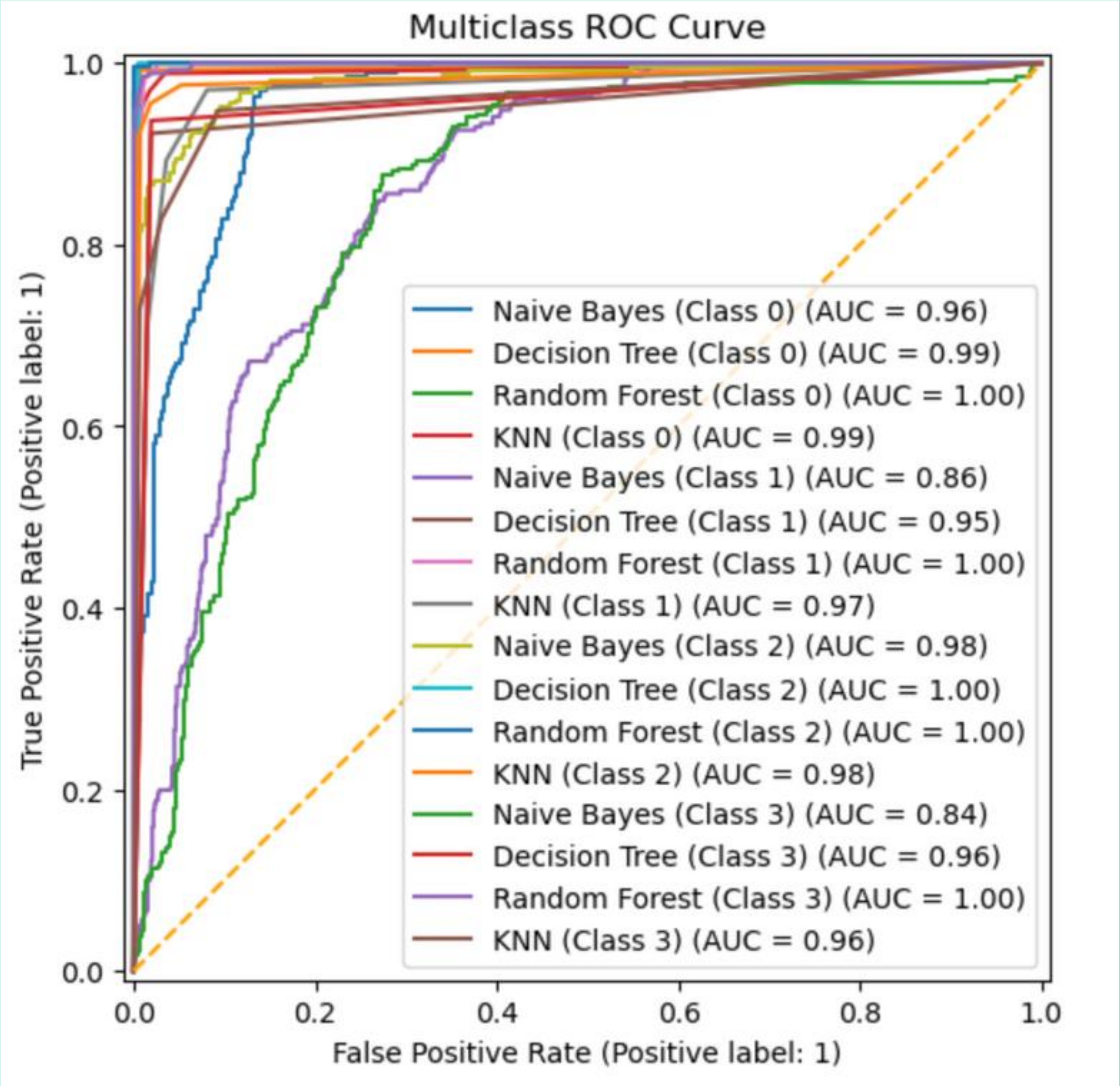
```
              precision    recall  f1-score   support

           0       1.00      0.99      0.99       263
           1       0.94      0.99      0.97       271
           2       0.99      1.00      0.99       248
           3       0.99      0.94      0.97       268

    accuracy                           0.98      1050
   macro avg       0.98      0.98      0.98      1050
weighted avg       0.98      0.98      0.98      1050
```

ROC CURVE

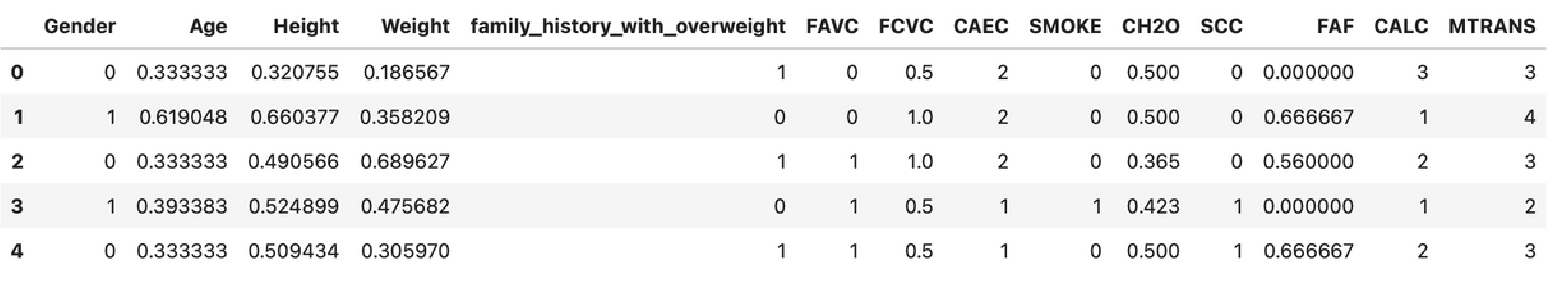
ROC Curve for Training Dataset



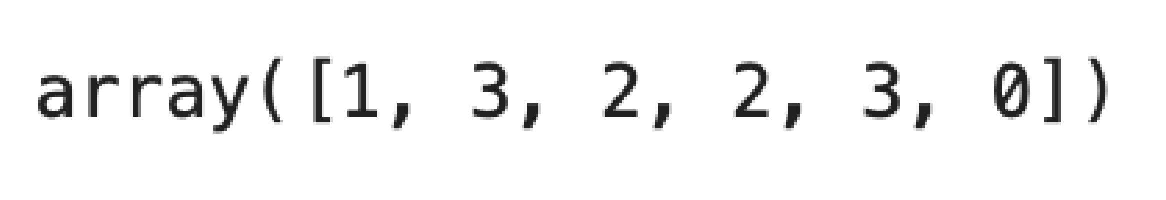
ROC Curve for Testing Dataset



PREDICTION DATASET



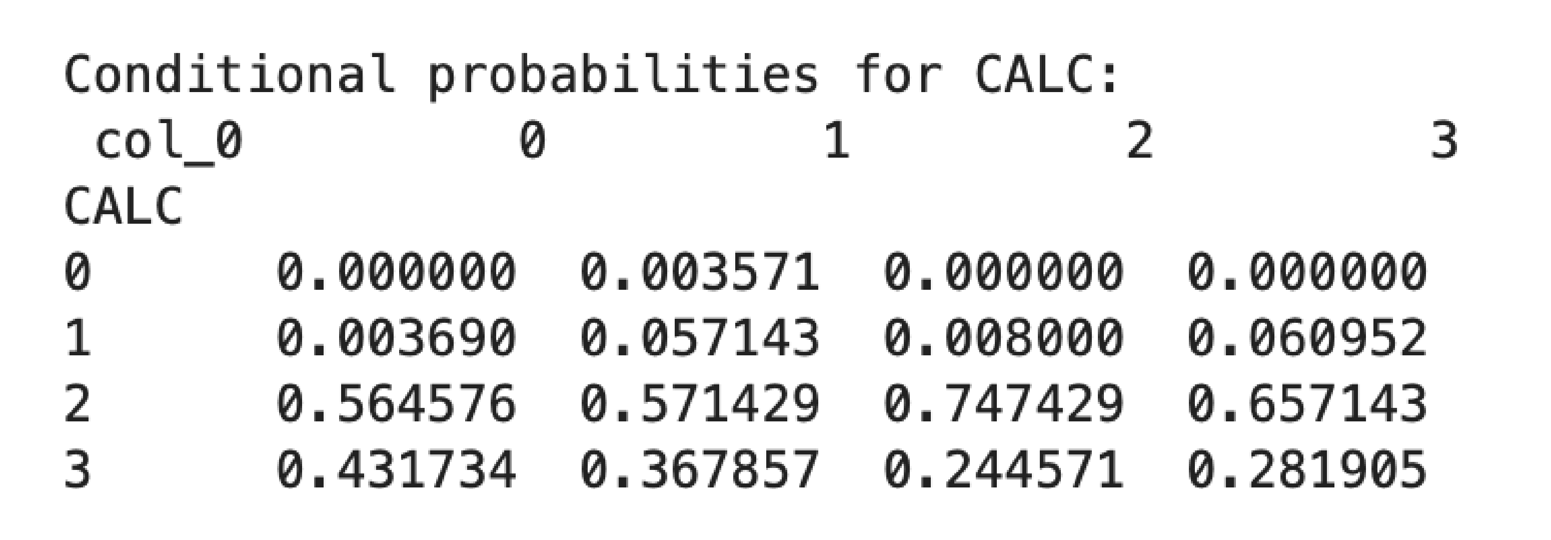
	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	CAEC	SMOKE	CH2O	SCC	FAF	CALC	MTRANS
0	0	0.333333	0.320755	0.186567	1	0	0.5	2	0	0.500	0	0.000000	3	3
1	1	0.619048	0.660377	0.358209	0	0	1.0	2	0	0.500	0	0.666667	1	4
2	0	0.333333	0.490566	0.689627	1	1	1.0	2	0	0.365	0	0.560000	2	3
3	1	0.393383	0.524899	0.475682	0	1	0.5	1	1	0.423	1	0.000000	1	2
4	0	0.333333	0.509434	0.305970	1	1	0.5	1	0	0.500	1	0.666667	2	3



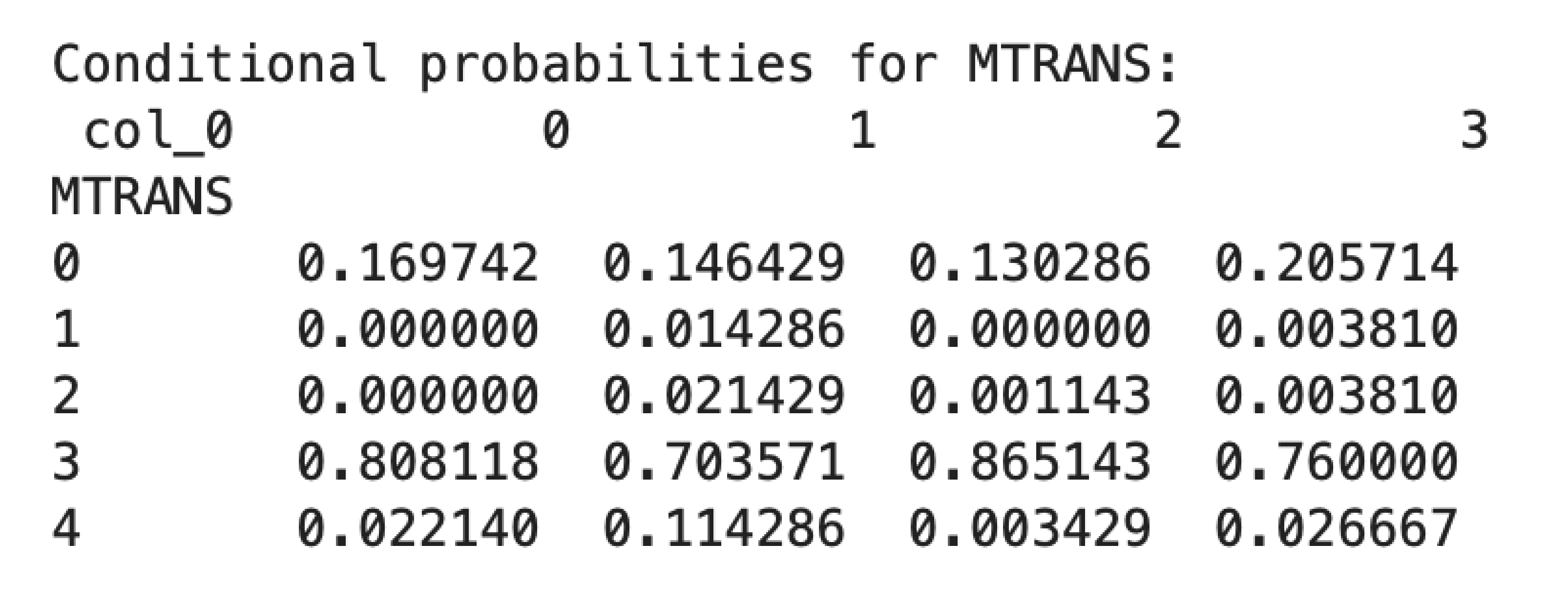
```
array([1, 3, 2, 2, 3, 0])
```

The model predicts correct outcomes for the given dataset.

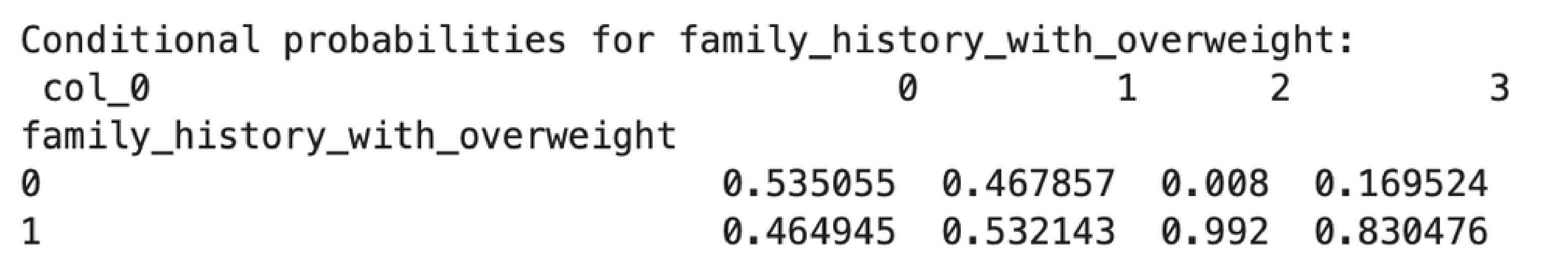
CONDITIONAL PROBABILITY



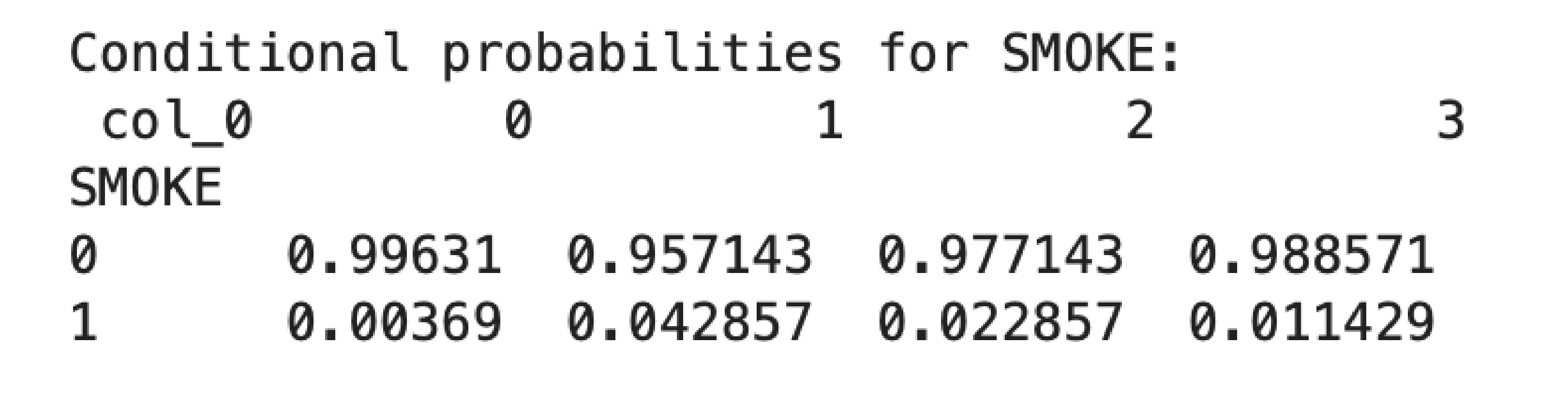
```
Conditional probabilities for CALC:
 col_0         0         1         2         3
CALC
0       0.000000  0.003571  0.000000  0.000000
1       0.003690  0.057143  0.008000  0.060952
2       0.564576  0.571429  0.747429  0.657143
3       0.431734  0.367857  0.244571  0.281905
```



```
Conditional probabilities for MTRANS:
 col_0          0         1         2         3
MTRANS
0        0.169742  0.146429  0.130286  0.205714
1        0.000000  0.014286  0.000000  0.003810
2        0.000000  0.021429  0.001143  0.003810
3        0.808118  0.703571  0.865143  0.760000
4        0.022140  0.114286  0.003429  0.026667
```

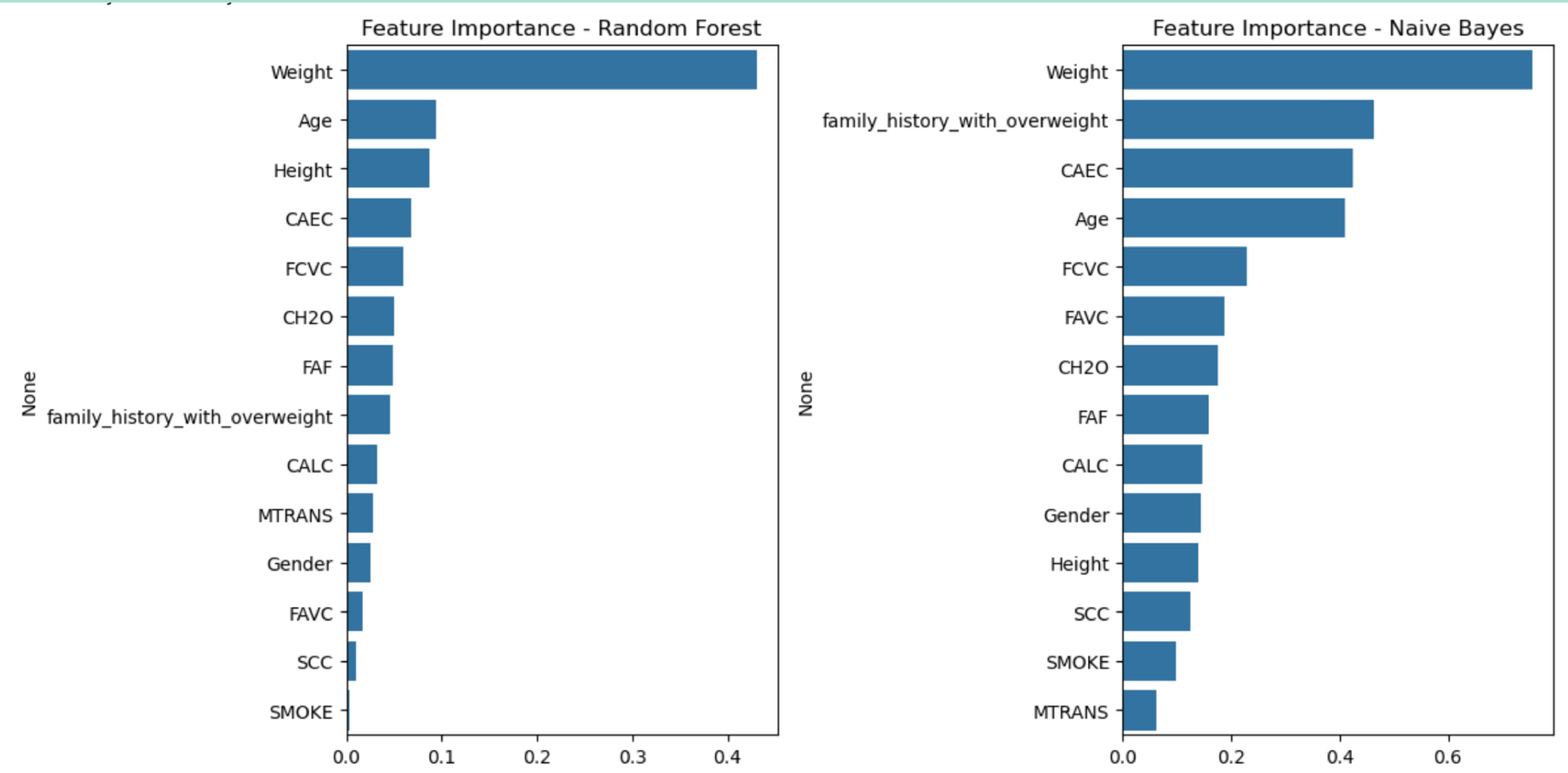


```
Conditional probabilities for family_history_with_overweight:
 col_0                                 0         1      2         3
family_history_with_overweight
0                               0.535055  0.467857  0.008  0.169524
1                               0.464945  0.532143  0.992  0.830476
```

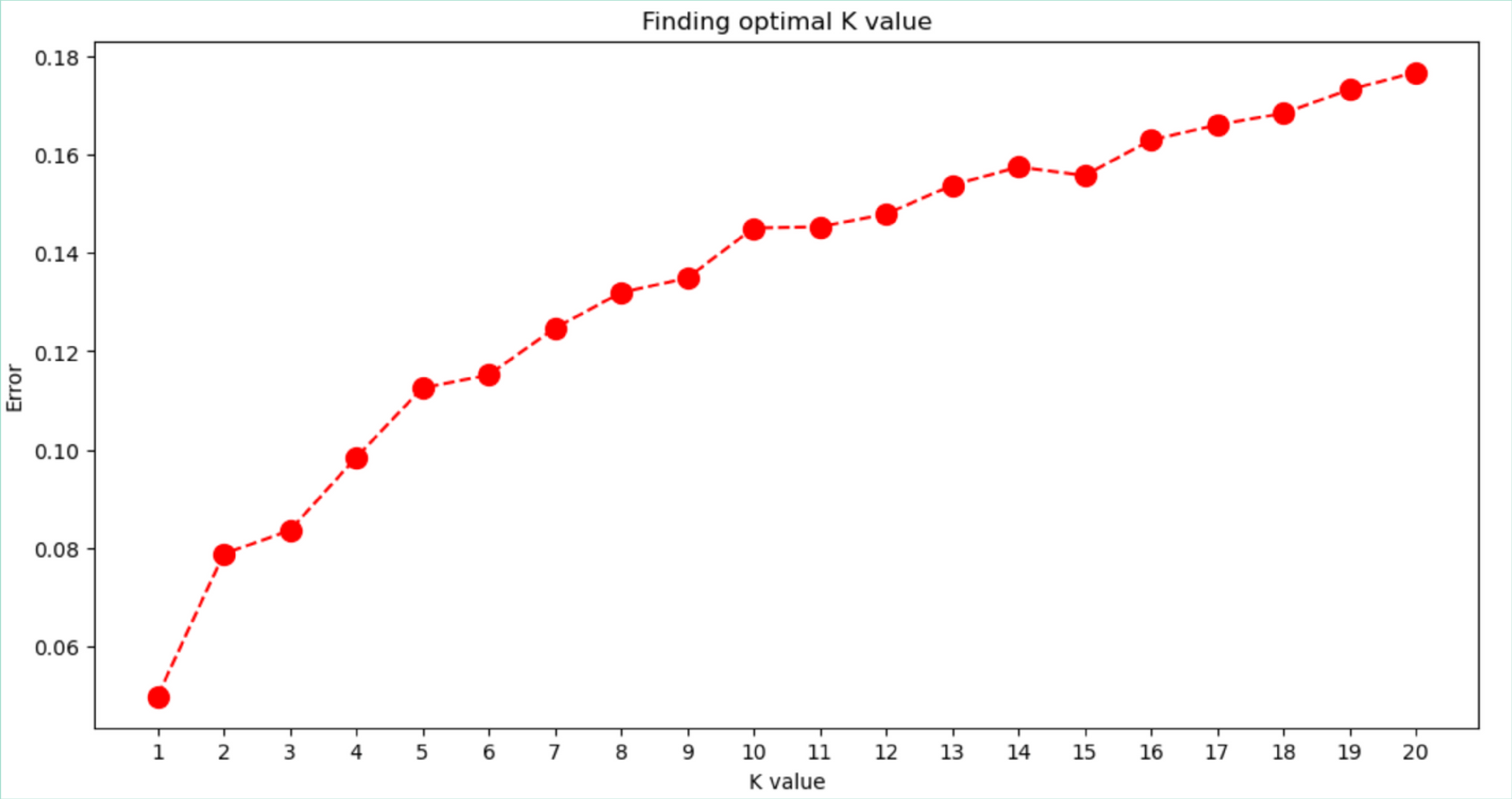


```
Conditional probabilities for SMOKE:
 col_0         0         1         2         3
SMOKE
0       0.99631  0.957143  0.977143  0.988571
1       0.00369  0.042857  0.022857  0.011429
```

FEATURE IMPORTANCE

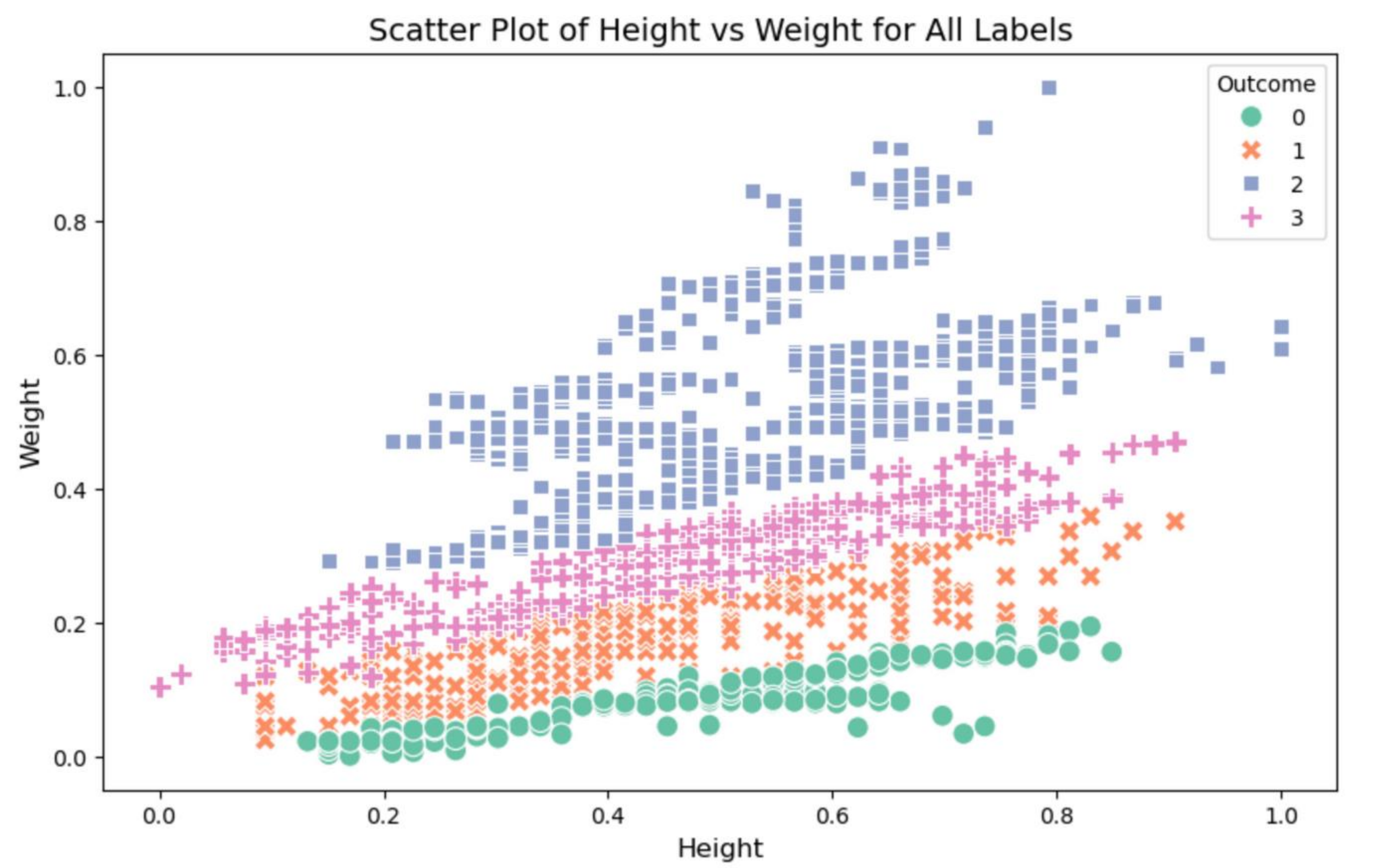


OPTIMAL K VALUE



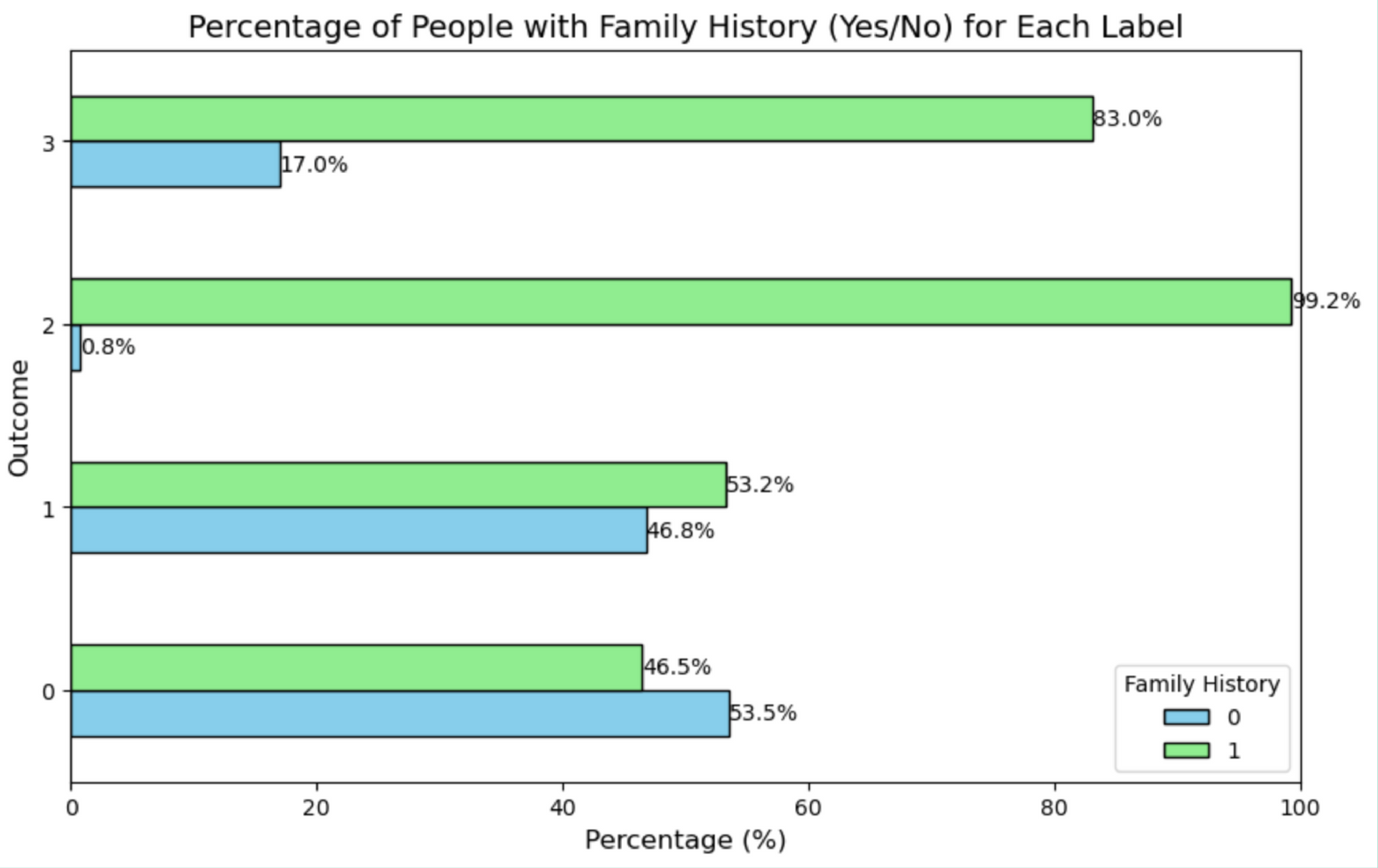
- Error Trend: The graph shows that as the value of k increases from 1 to 20, the error consistently increases, indicating that larger K values may lead to less optimal clustering performance.
- Optimal K Value: 1
-

VISUALIZATIONS - SCATTER PLOT



- The categories form distinct bands across the plot, showing a clear stratification of weight categories regardless of height.
- There's a positive correlation between height and weight within each category.

VISUALIZATIONS - BAR CHART



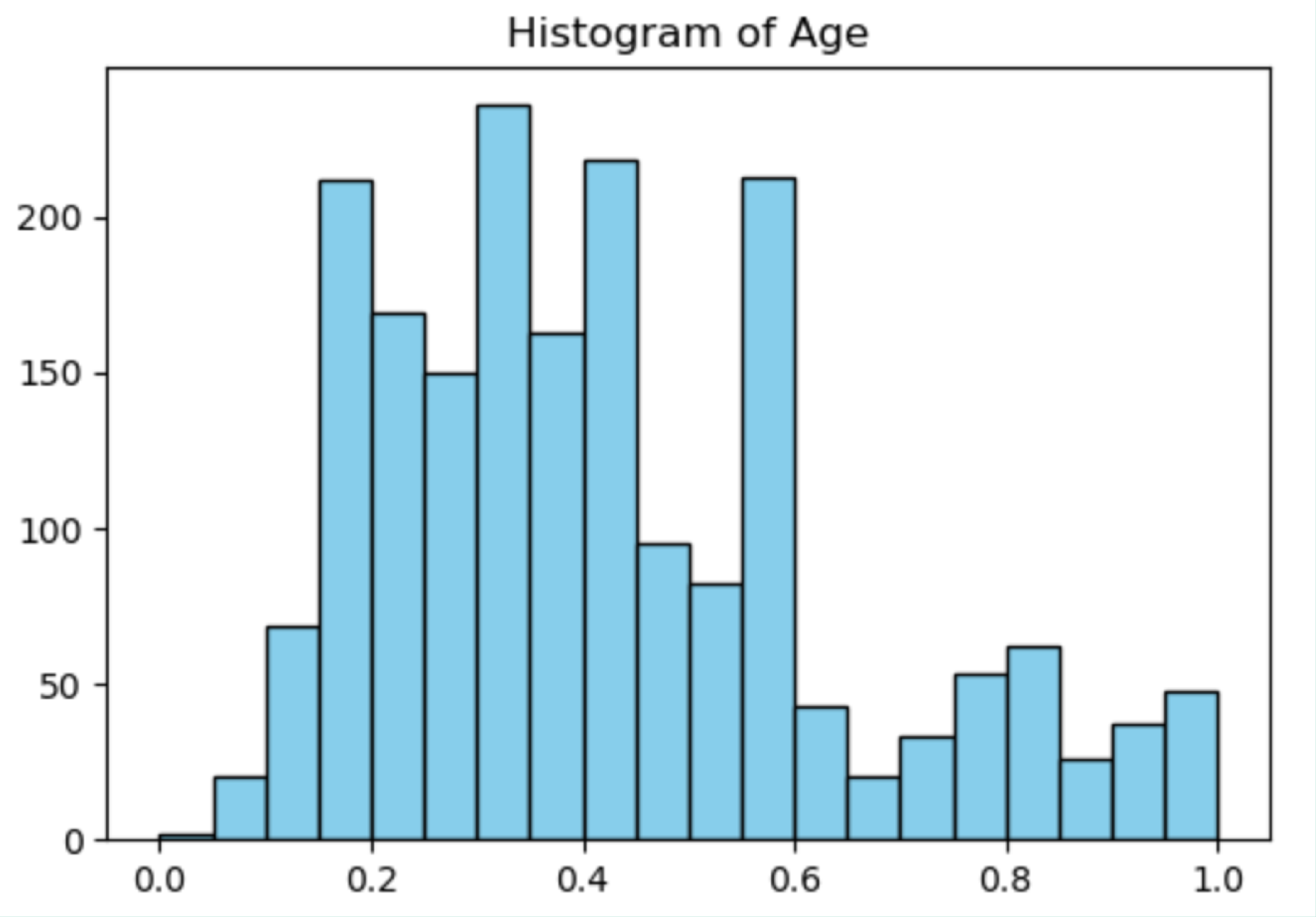
- Obesity individuals (Category 2) shows strongest relationship with family history with 99.2% of obese individuals having a family history.

- Overweight individuals (Category 3) also show a strong genetic link, with 83% having a family history of weight issues.

- Obesity & Overweight supports significant hereditary components.

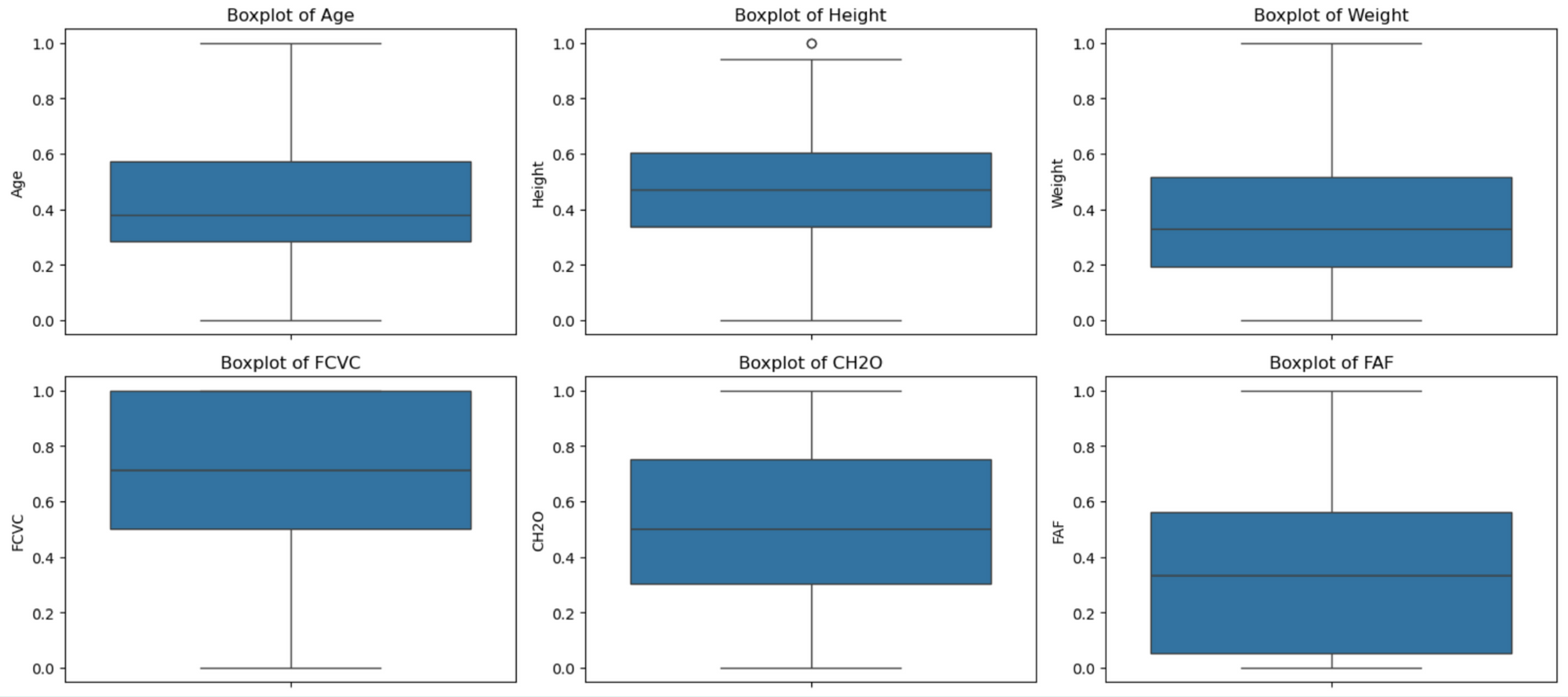
-

VISUALIZATIONS - HISTOGRAM



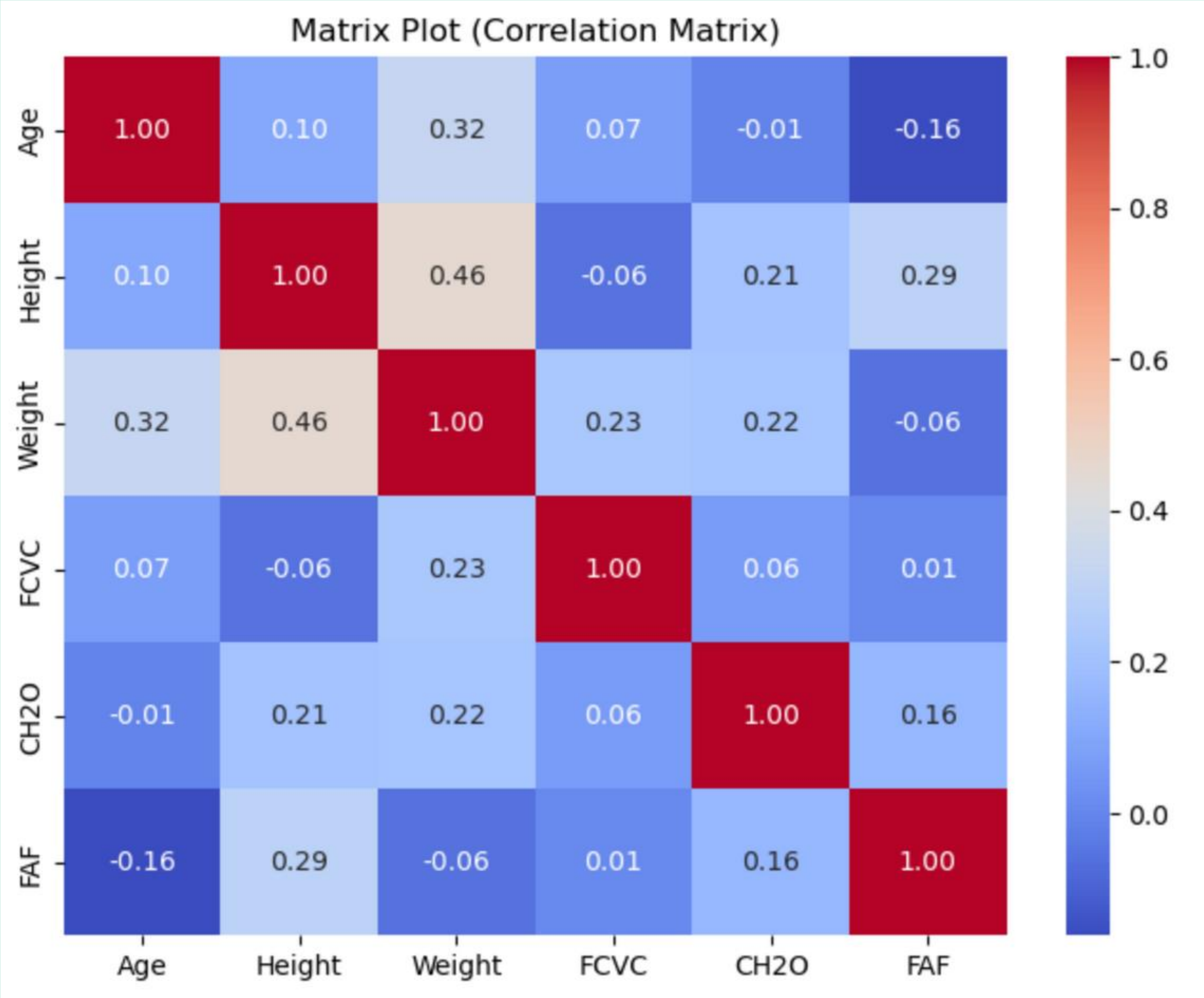
- The majority are in the range 0.15-0.6, suggesting the dataset primarily consists of young to middle-aged individuals.

VISUALIZATIONS - BOX PLOT



- Age: The age distribution is relatively uniform.
- Height: The height box plot, suggests some individuals are significantly taller than the majority.
- Weight: The weight distribution has a wider interquartile range, indicating variability in weight among individuals.
- FCVC: The FCVC values are concentrated with minimal spread, suggesting consistent food consumption patterns across the dataset.
- CH2O: The CH2O box plot shows balanced distribution and no significant outliers, reflecting stable water consumption levels.
- FAF (Physical Activity Frequency): The FAF distribution shows a median around 0.4, with a moderate spread, indicating variability in physical activity levels among individuals.

VISUALIZATIONS – CO-RELATION MATRIX

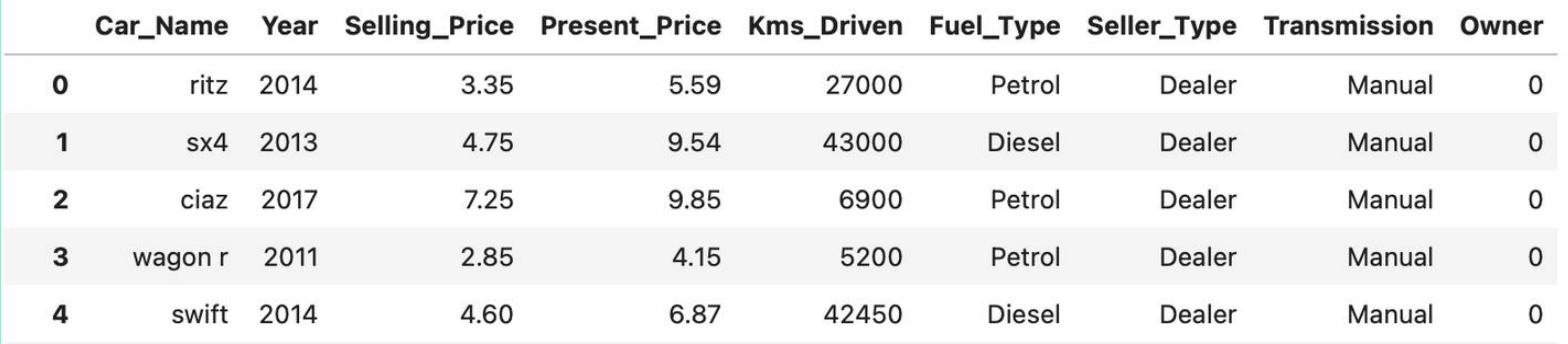


Matrix Plot (Correlation Matrix)

	Age	Height	Weight	FCVC	CH2O	FAF
Age	1.00	0.10	0.32	0.07	-0.01	-0.16
Height	0.10	1.00	0.46	-0.06	0.21	0.29
Weight	0.32	0.46	1.00	0.23	0.22	-0.06
FCVC	0.07	-0.06	0.23	1.00	0.06	0.01
CH2O	-0.01	0.21	0.22	0.06	1.00	0.16
FAF	-0.16	0.29	-0.06	0.01	0.16	1.00

- Height and Weight Correlation: There is a moderate positive correlation (0.46) between height and weight, indicating that taller individuals tend to weigh more.
- Weak Correlations: The other variables show weak correlations with each other, particularly between age and most other factors, suggesting limited relationships among age, food consumption variability (FCVC), water consumption (CH2O), and physical activity frequency (FAF).

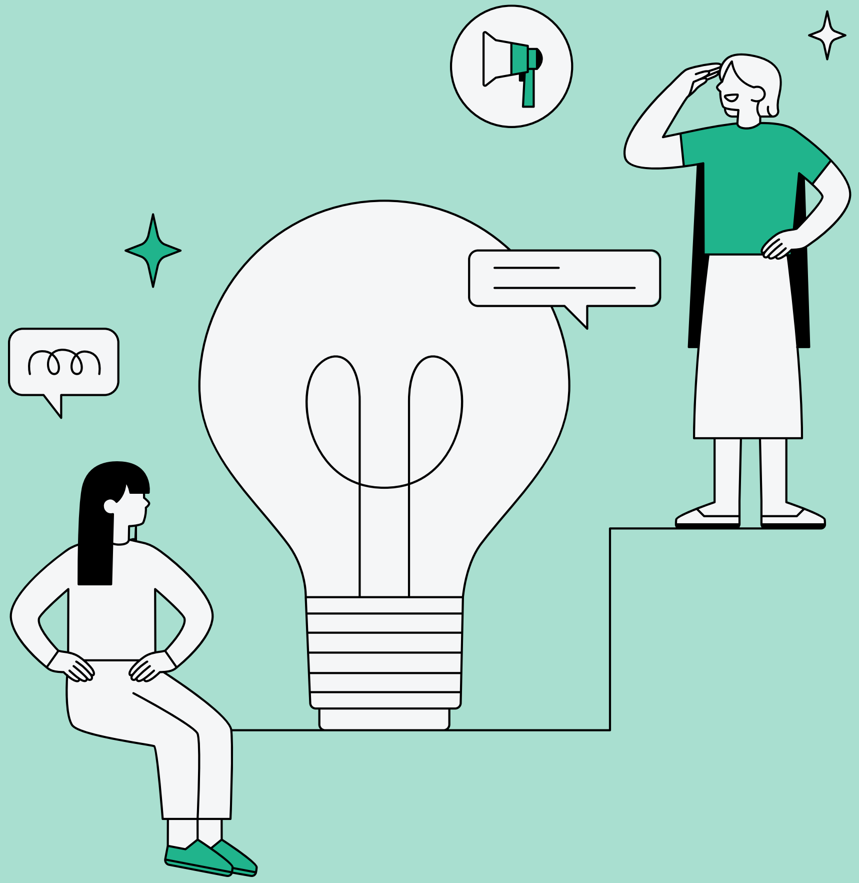
REGRESSION DATASET



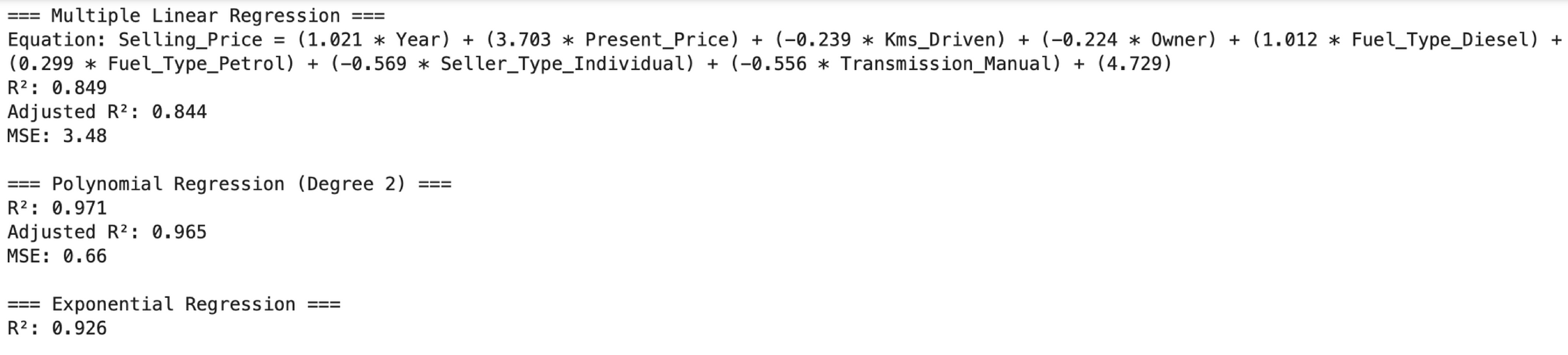
	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

Here are common descriptions for the variables in the dataset.

Car_Name - Name of the car
Year - Year of Manufacturing
Selling_Price - Selling Price
Present_Price - Present Price
Kms_Driven - Kms driven
Fuel_Type - Fuel Type (Petrol, Diesel, CNG)
Seller_Type - Seller (Individual, Dealer)
Transmission - Manual, Automatic
Owner - No of owners (0, 1, 3)



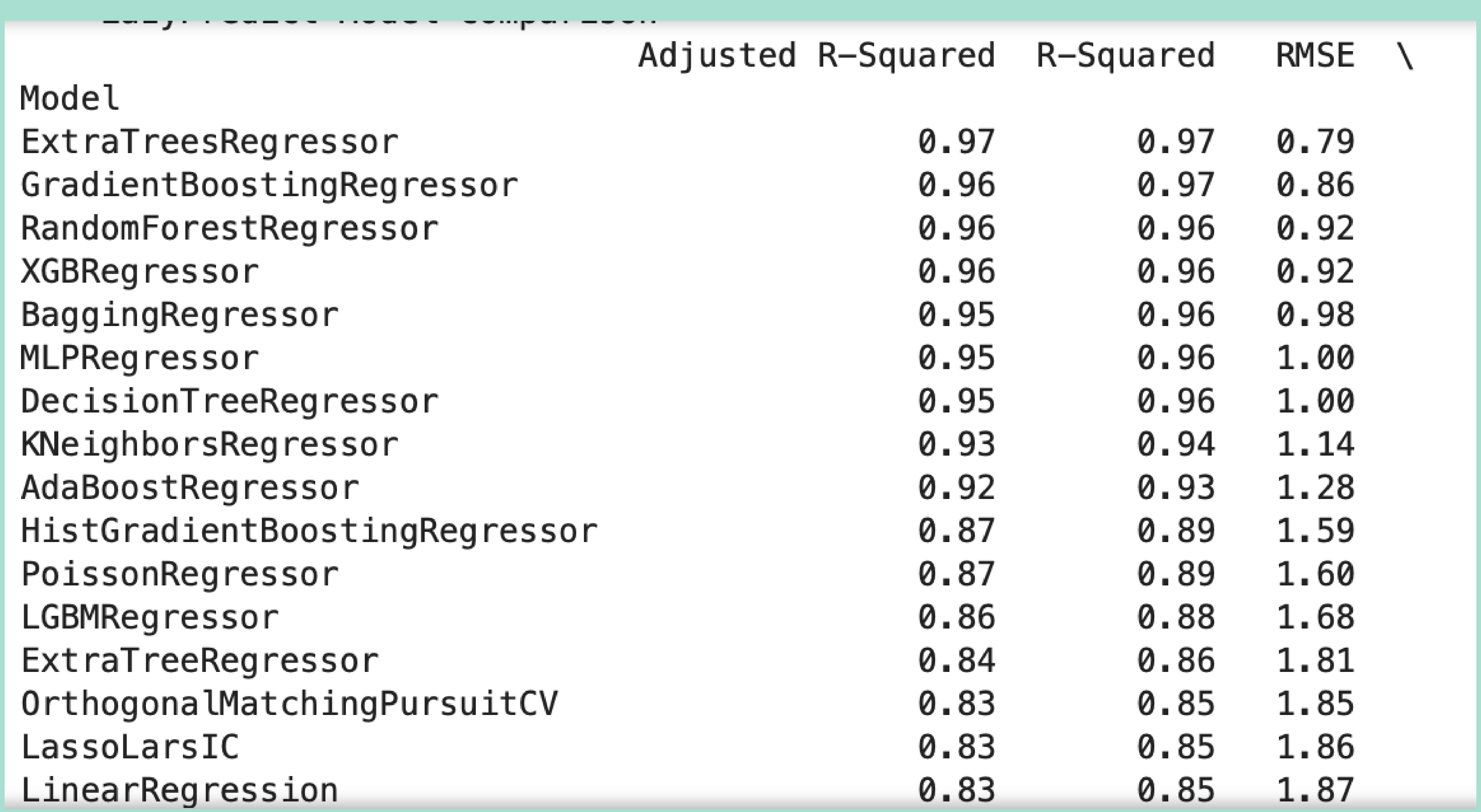
REGRESSION - R2 & Adj R2 (Linear, Polynomial, Exponential)



```
=== Multiple Linear Regression ===
Equation: Selling_Price = (1.021 * Year) + (3.703 * Present_Price) + (-0.239 * Kms_Driven) + (-0.224 * Owner) + (1.012 * Fuel_Type_Diesel) +
(0.299 * Fuel_Type_Petrol) + (-0.569 * Seller_Type_Individual) + (-0.556 * Transmission_Manual) + (4.729)
R²: 0.849
Adjusted R²: 0.844
MSE: 3.48

=== Polynomial Regression (Degree 2) ===
R²: 0.971
Adjusted R²: 0.965
MSE: 0.66

=== Exponential Regression ===
R²: 0.926
```

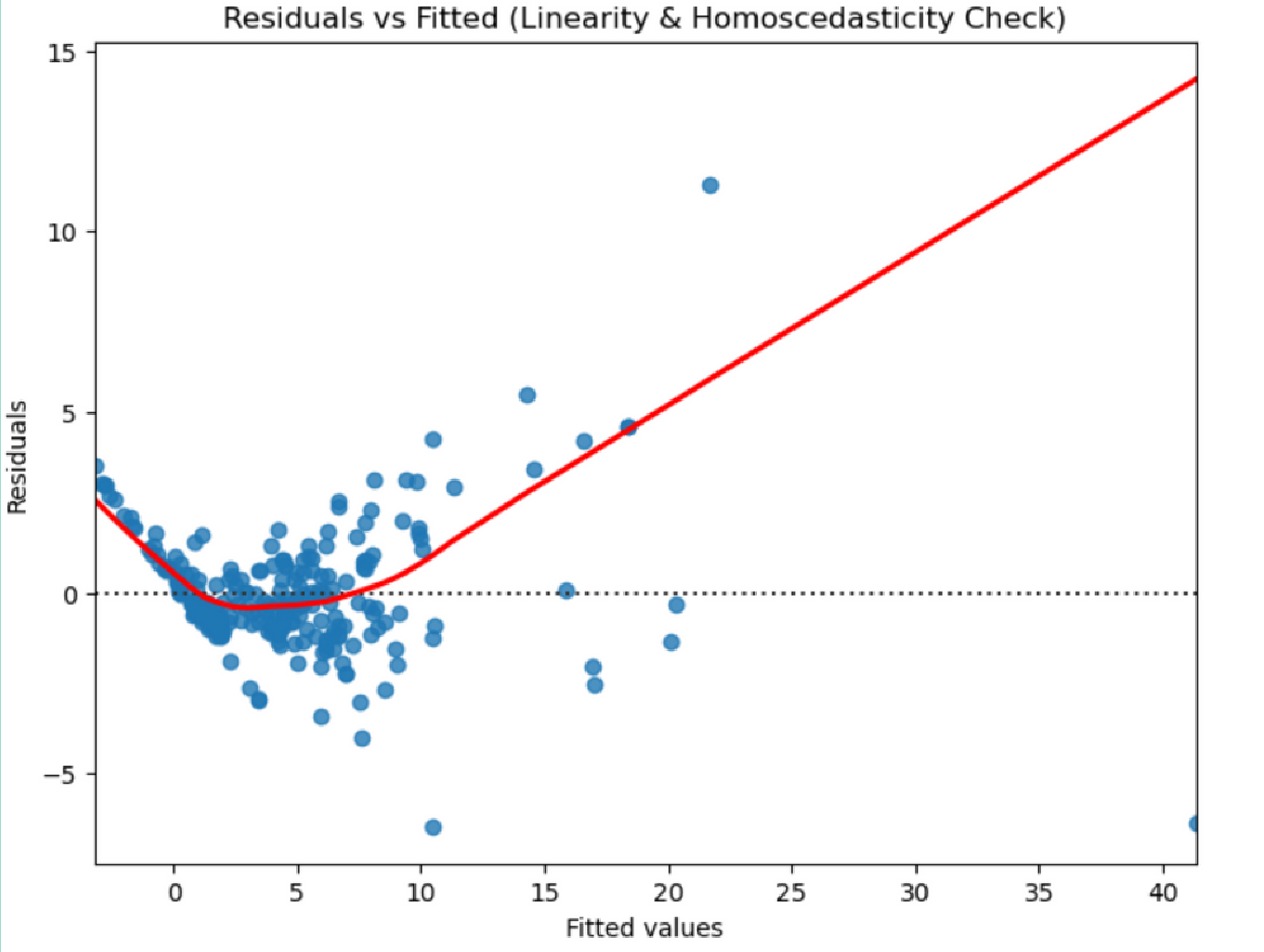
- Model Performance: The polynomial regression model (degree 2) outperforms the multiple linear and exponential regression models, with an R^2 of 0.971 and a low MSE of 0.66, indicating it effectively captures the non-linear relationship between predictors and selling price.
- Impact of Variables: In the multiple linear regression, "Present_Price" has the highest positive influence on selling price, while "Kms_Driven" and "Owner" negatively affect it. This highlights the importance of these factors in car valuations.

LAZY PREDICT



```
    LazyPredict Model Comparison
                               Adjusted R-Squared  R-Squared  RMSE  \
Model
ExtraTreesRegressor                          0.97       0.97  0.79
GradientBoostingRegressor                    0.96       0.97  0.86
RandomForestRegressor                        0.96       0.96  0.92
XGBRegressor                                 0.96       0.96  0.92
BaggingRegressor                             0.95       0.96  0.98
MLPRegressor                                 0.95       0.96  1.00
DecisionTreeRegressor                        0.95       0.96  1.00
KNeighborsRegressor                          0.93       0.94  1.14
AdaBoostRegressor                            0.92       0.93  1.28
HistGradientBoostingRegressor                0.87       0.89  1.59
PoissonRegressor                             0.87       0.89  1.60
LGBMRegressor                                0.86       0.88  1.68
ExtraTreeRegressor                           0.84       0.86  1.81
OrthogonalMatchingPursuitCV                  0.83       0.85  1.85
LassoLarsIC                                  0.83       0.85  1.86
LinearRegression                             0.83       0.85  1.87
```

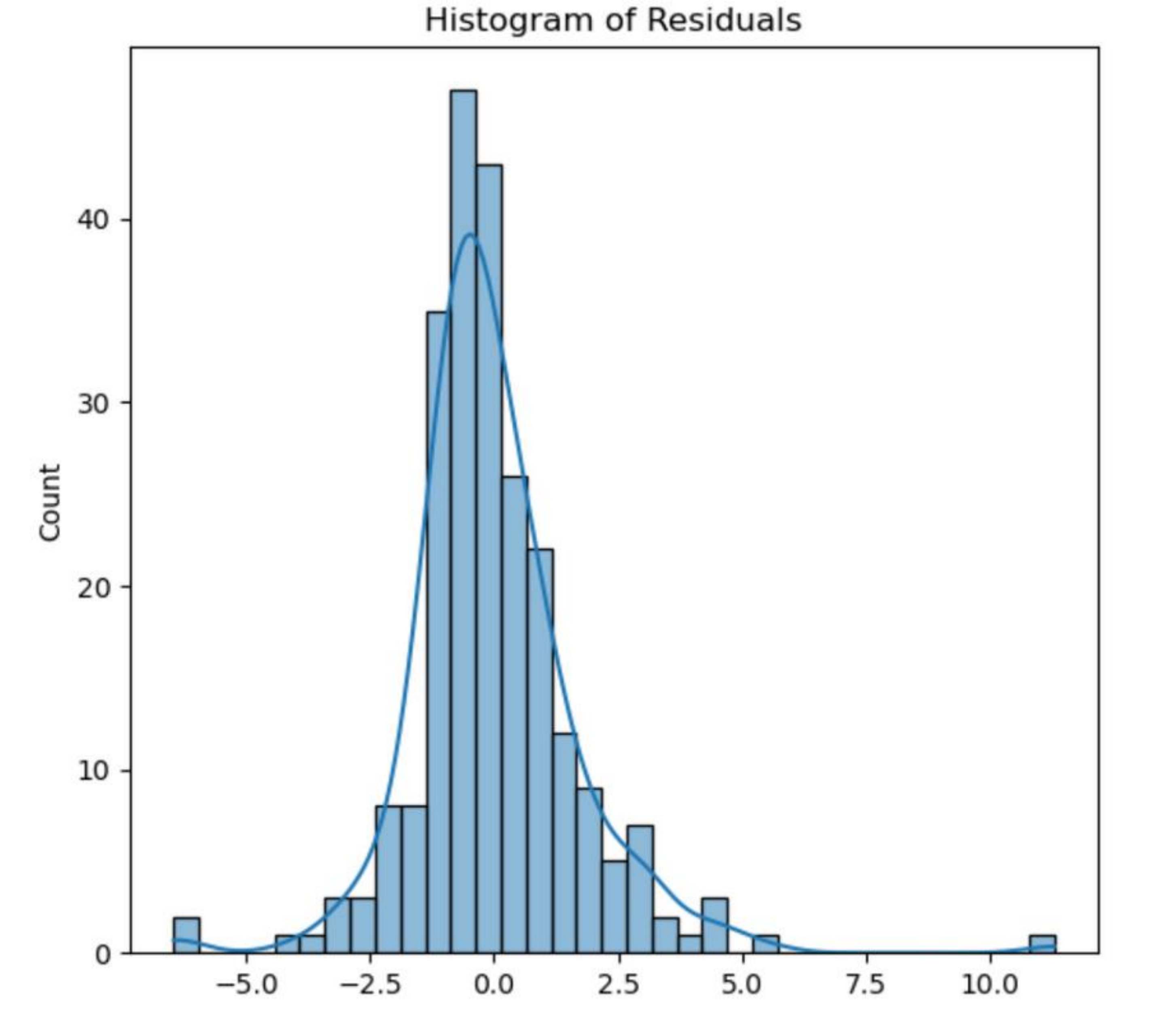
RESIDUALS Vs FITTED

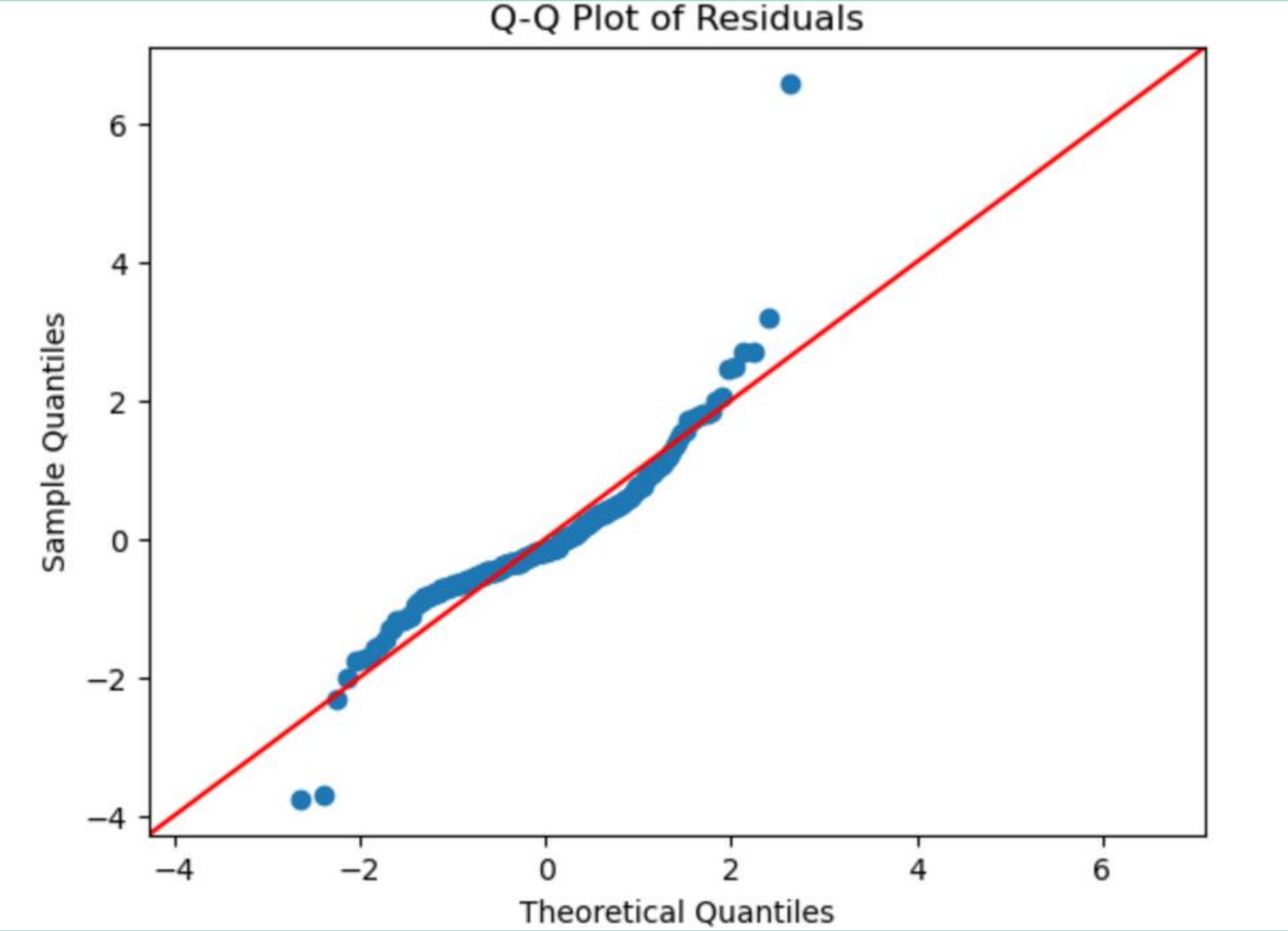


- Non-Linearity Indication: The residuals display a systematic pattern rather than random scatter, suggesting that a linear model may not adequately fit the data and that a polynomial regression could be more suitable.

- Heteroscedasticity: The increasing spread of residuals with higher fitted values indicates heteroscedasticity, meaning the variance of the residuals is not constant. This suggests varying predictive accuracy across different ranges of the data.

RESIDUALS HISTOGRAM & Q-Q PLOT

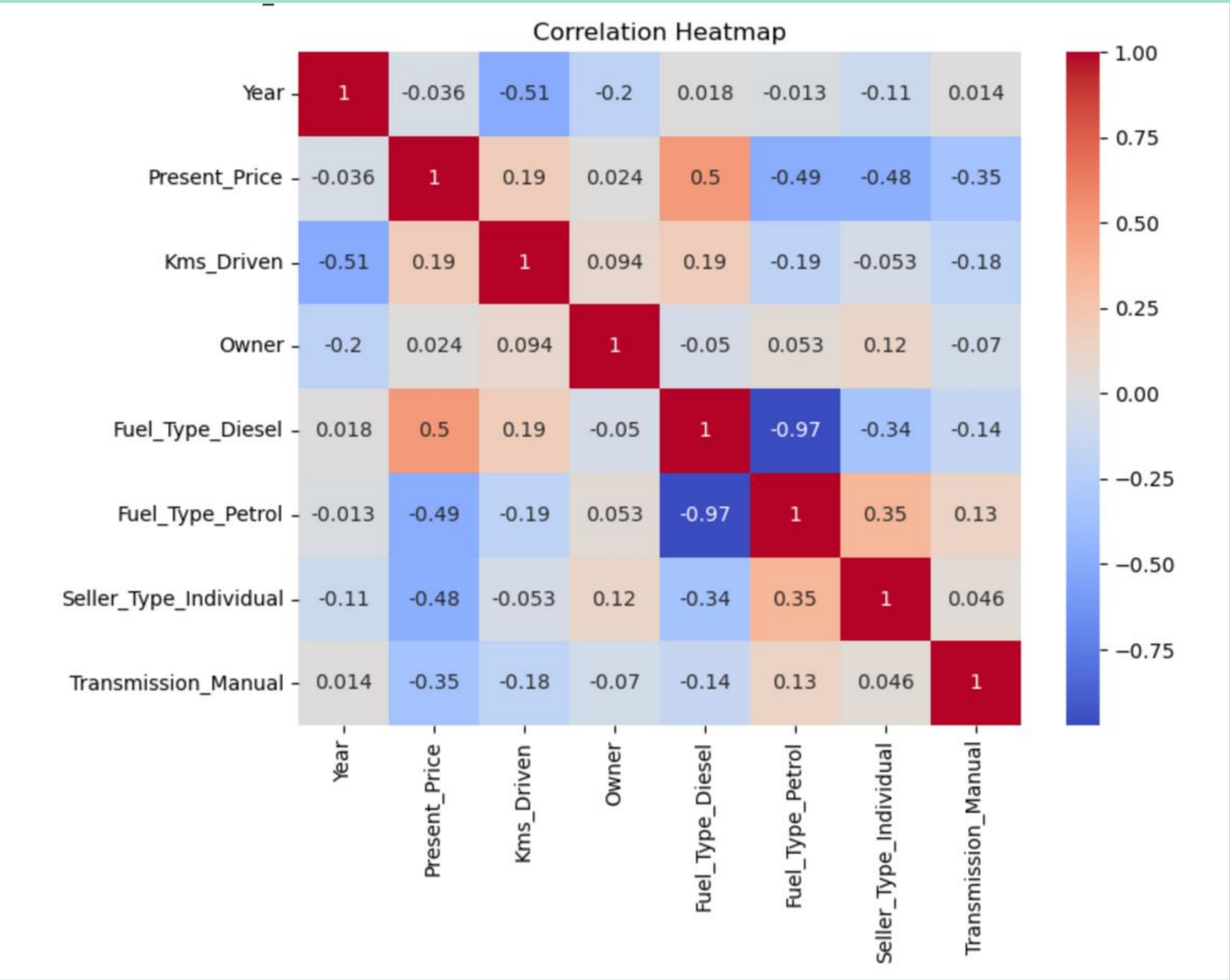




- The histogram shows that the residuals are approximately normally distributed.

- Most of the points in the centre follow the red diagonal line.
- The model captures general trends.
- There are some infuential outliers affecting the model.

CORRELATION HEATMAP



	Year	Present_Price	Kms_Driven	Owner	Fuel_Type_Diesel	Fuel_Type_Petrol	Seller_Type_Individual	Transmission_Manual
Year	1	-0.036	-0.51	-0.2	0.018	-0.013	-0.11	0.014
Present_Price	-0.036	1	0.19	0.024	0.5	-0.49	-0.48	-0.35
Kms_Driven	-0.51	0.19	1	0.094	0.19	-0.19	-0.053	-0.18
Owner	-0.2	0.024	0.094	1	-0.05	0.053	0.12	-0.07
Fuel_Type_Diesel	0.018	0.5	0.19	-0.05	1	-0.97	-0.34	-0.14
Fuel_Type_Petrol	-0.013	-0.49	-0.19	0.053	-0.97	1	0.35	0.13
Seller_Type_Individual	-0.11	-0.48	-0.053	0.12	-0.34	0.35	1	0.046
Transmission_Manual	0.014	-0.35	-0.18	-0.07	-0.14	0.13	0.046	1

- **Strongest negative correlation (-0.97) is between Fuel_Type_Diesel and Fuel_Type_Petrol, which is expected since these are mutually exclusive categories (a car cannot be both diesel and petrol).**

- **There's a moderate negative correlation (-0.51) between Year and Kms_Driven, indicating newer cars tend to have lower mileage.**

- **Present_Price shows a moderate positive correlation (0.5) with Fuel_Type_Diesel, suggesting diesel cars are generally priced higher in this dataset.**

Thank you very much!