# Prediction of Smoking Trend in Gender.

By Collin Tully, Sahrish Afzal, Eric Cacdac,Rucha Soni

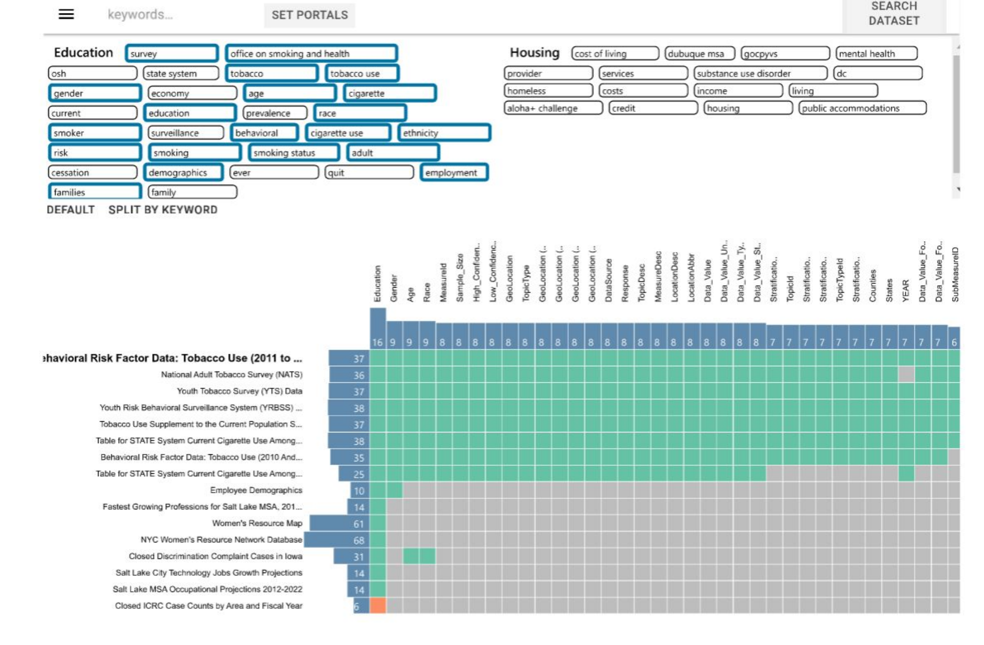
# Objective of the Analysis

Find the smoking trend in gender by using age , race, and echolocation( Alaska and Alabama) features

# The Dataset

Used data set:

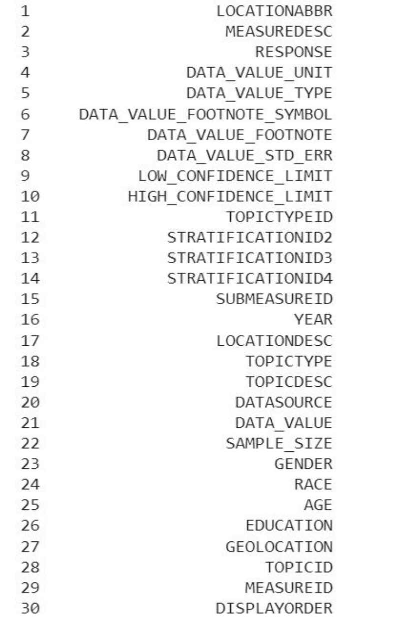
- WE have used  the Behavioral Risk Factor Data by using shown portals
- That gave us data with 30 features



| | |
|---|---|
| ehavioral Risk Factor Data: Tobacco Use (2011 to ... | 37 |
| National Adult Tobacco Survey (NATS) | 36 |
| Youth Tobacco Survey (YTS) Data | 37 |
| Youth Risk Behavioral Surveillance System (YRBSS) ... | 38 |
| Tobacco Use Supplement to the Current Population S... | 37 |
| Table for STATE System Current Cigarette Use Among... | 38 |
| Behavioral Risk Factor Data: Tobacco Use (2010 And... | 35 |
| Table for STATE System Current Cigarette Use Among... | 25 |
| Employee Demographics | 10 |
| Fastest Growing Professions for Salt Lake MSA, 201... | 14 |
| Women's Resource Map | 61 |
| NYC Women's Resource Network Database | 68 |
| Closed Discrimination Complaint Cases in Iowa | 31 |
| Salt Lake City Technology Jobs Growth Projections | 14 |
| Salt Lake MSA Occupational Projections 2012-2022 | 14 |
| Closed ICRC Case Counts by Area and Fiscal Year | 6 |

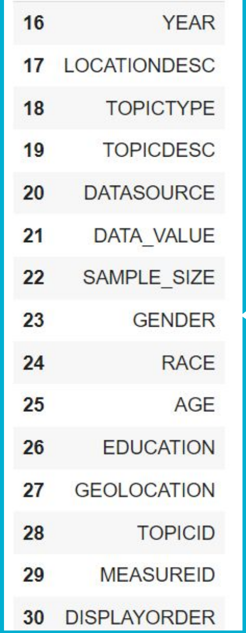
# Features of the Dataset

## Original Data



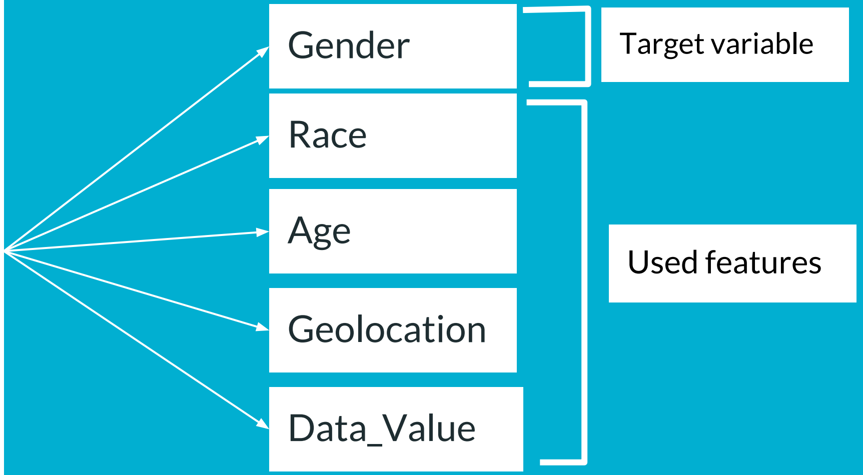
```
1                    LOCATIONABBR
2                    MEASUREDESC
3                       RESPONSE
4                DATA_VALUE_UNIT
5                DATA_VALUE_TYPE
6     DATA_VALUE_FOOTNOTE_SYMBOL
7            DATA_VALUE_FOOTNOTE
8             DATA_VALUE_STD_ERR
9           LOW_CONFIDENCE_LIMIT
10         HIGH_CONFIDENCE_LIMIT
11                  TOPICTYPEID
12             STRATIFICATIONID2
13             STRATIFICATIONID3
14             STRATIFICATIONID4
15                 SUBMEASUREID
16                         YEAR
17                 LOCATIONDESC
18                    TOPICTYPE
19                    TOPICDESC
20                   DATASOURCE
21                   DATA_VALUE
22                  SAMPLE_SIZE
23                       GENDER
24                         RACE
25                          AGE
26                    EDUCATION
27                  GEOLOCATION
28                      TOPICID
29                    MEASUREID
30                 DISPLAYORDER
```

## Present Data



| | |
|---|---|
| 16 | YEAR |
| 17 | LOCATIONDESC |
| 18 | TOPICTYPE |
| 19 | TOPICDESC |
| 20 | DATASOURCE |
| 21 | DATA_VALUE |
| 22 | SAMPLE_SIZE |
| 23 | GENDER |
| 24 | RACE |
| 25 | AGE |
| 26 | EDUCATION |
| 27 | GEOLOCATION |
| 28 | TOPICID |
| 29 | MEASUREID |
| 30 | DISPLAYORDER |

## Focus Data



Gender

Race

Age

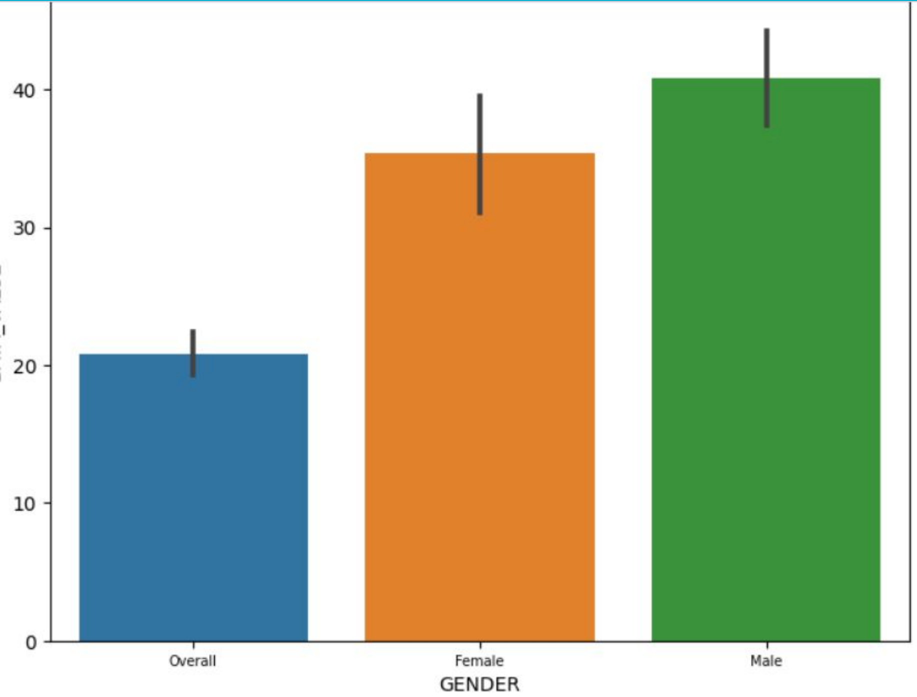
Geolocation

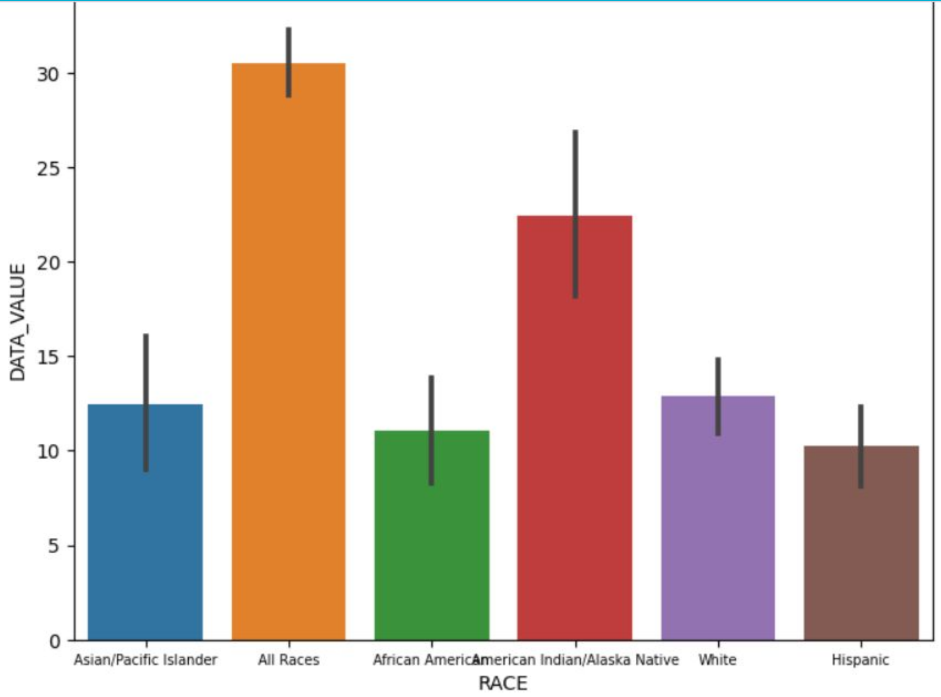
Data_Value

Target variable

Used features

# Graph: Feature Vs Target



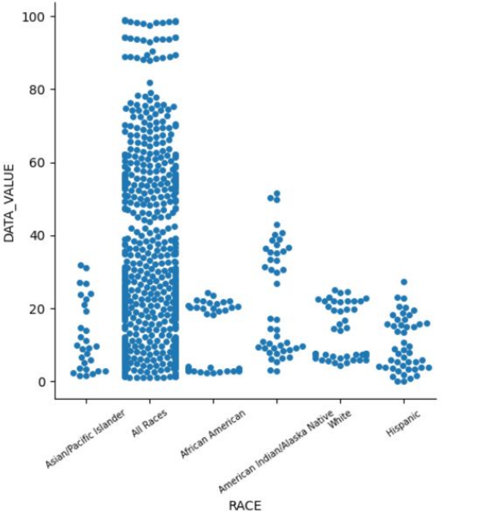


# First Model: Linear Regression

Purpose: To determine if there is a relationship between the features and the target variable.

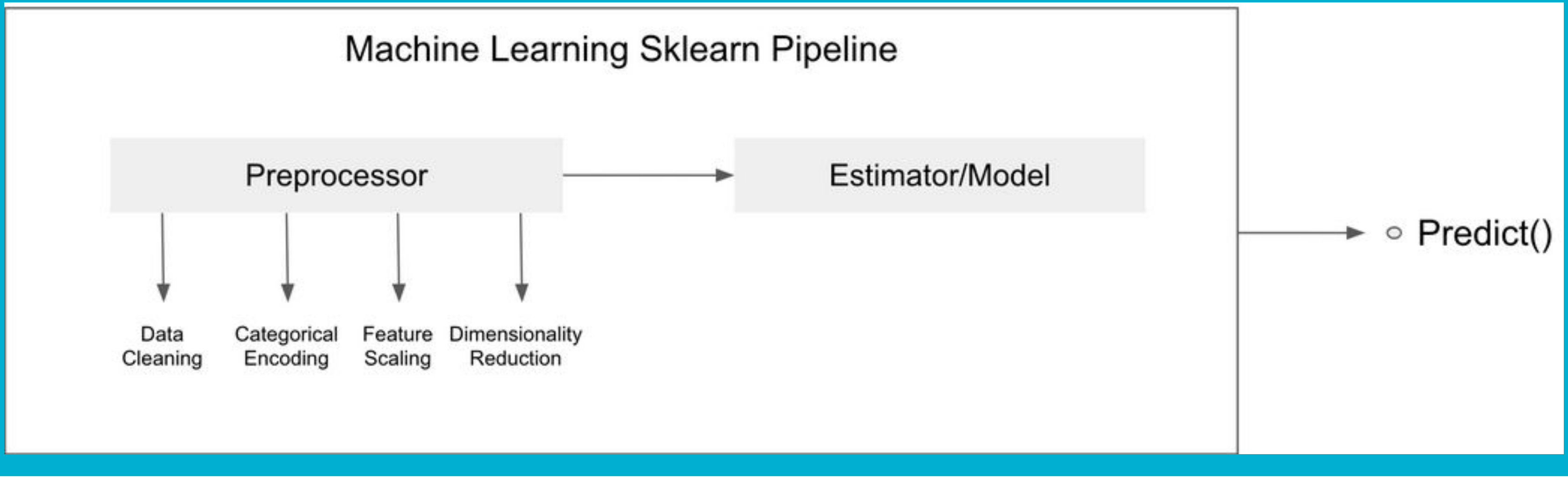
Features Used:  Race, Age, Location, Probability of Smoking.

Target Variable: Sex



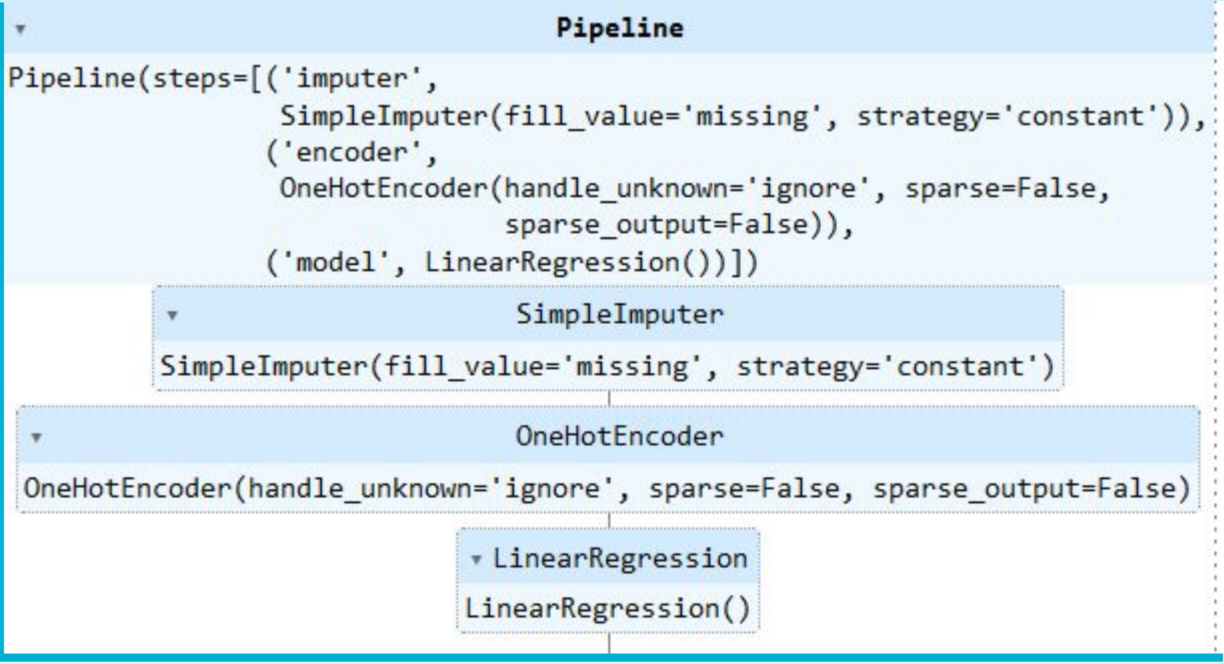
# First Mode: Linear Regression Encoding

- Binary Encoding
- Encoding using Sklearn Pipeline



# Evaluation of First Model: Linear Regression

R2 = 0.728



▾ **Pipeline**

```
Pipeline(steps=[('imputer',
                 SimpleImputer(fill_value='missing', strategy='constant')),
                ('encoder',
                 OneHotEncoder(handle_unknown='ignore', sparse=False,
                               sparse_output=False)),
                ('model', LinearRegression())])
```

▾ SimpleImputer

```
SimpleImputer(fill_value='missing', strategy='constant')
```

▾ OneHotEncoder

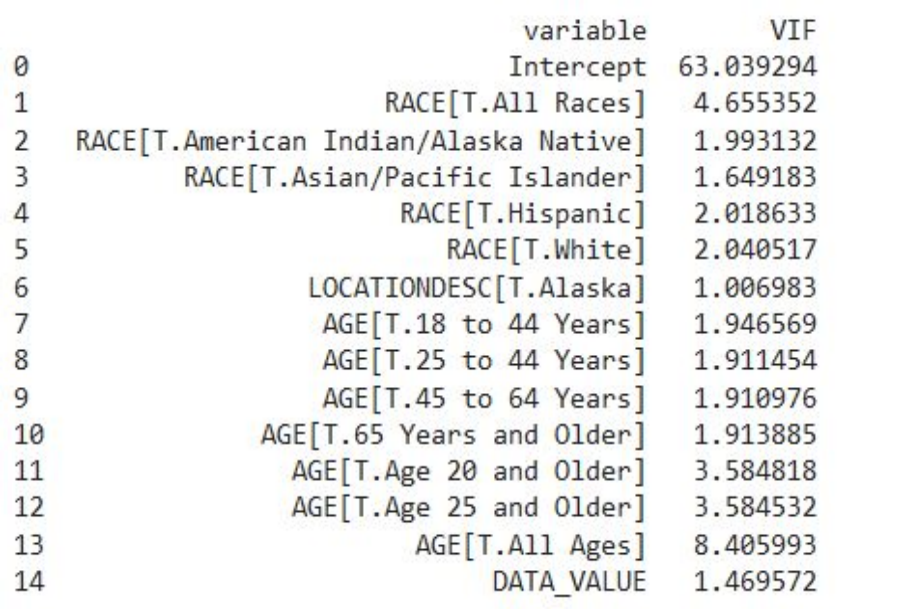
```
OneHotEncoder(handle_unknown='ignore', sparse=False, sparse_output=False)
```

▾ LinearRegression

```
LinearRegression()
```

# Issues With the First Model: Linear Regression

- Encoding of categorical target variable. Some of the data was mixed with male and female.
- Issue with collinearity
- VIF: 1 lack of collinearity,
- VIF: >5 correlation between predictor
- Variance Inflation Factor



```
                                     variable        VIF
0                                   Intercept  63.039294
1                         RACE[T.All Races]     4.655352
2   RACE[T.American Indian/Alaska Native]      1.993132
3            RACE[T.Asian/Pacific Islander]     1.649183
4                          RACE[T.Hispanic]     2.018633
5                             RACE[T.White]     2.040517
6                    LOCATIONDESC[T.Alaska]     1.006983
7                    AGE[T.18 to 44 Years]      1.946569
8                    AGE[T.25 to 44 Years]      1.911454
9                    AGE[T.45 to 64 Years]      1.910976
10               AGE[T.65 Years and Older]      1.913885
11                 AGE[T.Age 20 and Older]      3.584818
12                 AGE[T.Age 25 and Older]      3.584532
13                          AGE[T.All Ages]     8.405993
14                               DATA_VALUE     1.469572
```

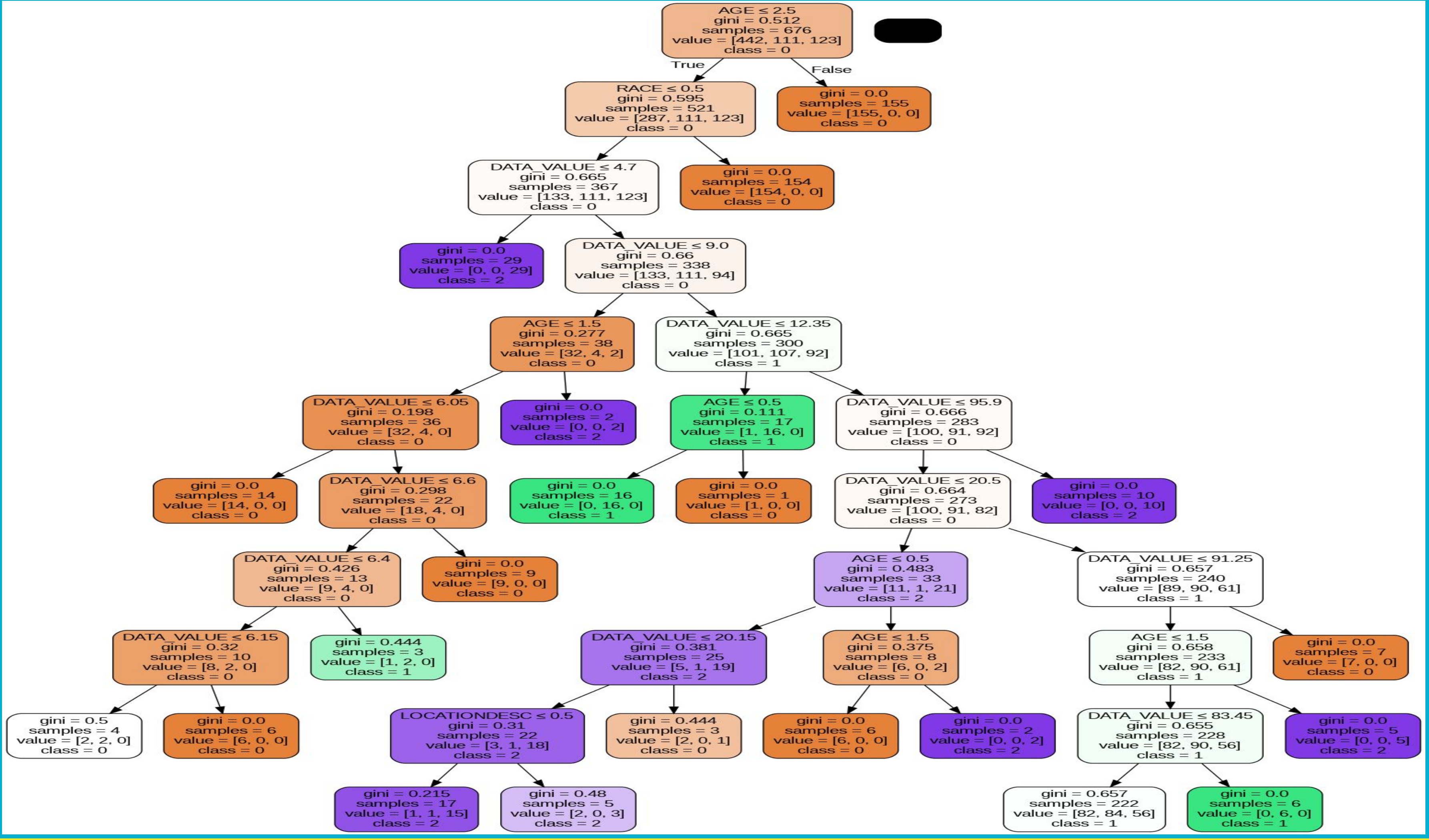
# Second Model: Decision Tree

Objective: Predict the target variable (gender) based on the independent variables (features: race, age, location, data value)

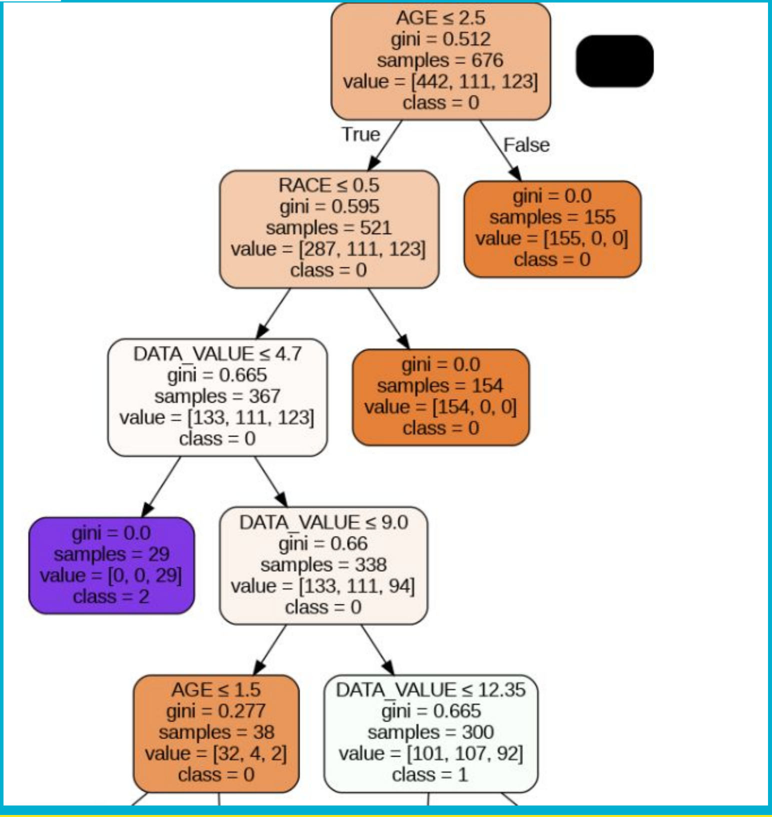
Assigned all feature values (non-numerical) to a numerical value:

- Gender: Male (0), Female (1), Overall (2)
- Race: All Races (0), Asian/Pacific Islander (1), American Indian/Alaskan Native(2), African American (3), White (4), Hispanic (5)
- Age: All Ages (0), 18 to 24 (1), 18 to 44 (2), 20+ (3), 25+ (4), 25 to 44 (5), 45 to 64 (6), 65+ (7)
- Location: Alabama (0), Alaska (1)

[redacted]


AGE ≤ 2.5
gini = 0.512
samples = 676
value = [442, 111, 123]
class = 0
True
False
RACE ≤ 0.5
gini = 0.595
samples = 521
value = [287, 111, 123]
class = 0
gini = 0.0
samples = 155
value = [155, 0, 0]
class = 0
DATA_VALUE ≤ 4.7
gini = 0.665
samples = 367
value = [133, 111, 123]
class = 0
gini = 0.0
samples = 154
value = [154, 0, 0]
class = 0
gini = 0.0
samples = 29
value = [0, 0, 29]
class = 2
DATA_VALUE ≤ 9.0
gini = 0.66
samples = 338
value = [133, 111, 94]
class = 0
AGE ≤ 1.5
gini = 0.277
samples = 38
value = [32, 4, 2]
class = 0
DATA_VALUE ≤ 12.35
gini = 0.665
samples = 300
value = [101, 107, 92]
class = 1
DATA_VALUE ≤ 6.05
gini = 0.198
samples = 36
value = [32, 4, 0]
class = 0
gini = 0.0
samples = 2
value = [0, 0, 2]
class = 2
AGE ≤ 0.5
gini = 0.111
samples = 17
value = [1, 16, 0]
class = 1
DATA_VALUE ≤ 95.9
gini = 0.666
samples = 283
value = [100, 91, 92]
class = 0
gini = 0.0
samples = 14
value = [14, 0, 0]
class = 0
DATA_VALUE ≤ 6.6
gini = 0.298
samples = 22
value = [18, 4, 0]
class = 0
gini = 0.0
samples = 16
value = [0, 16, 0]
class = 1
gini = 0.0
samples = 1
value = [1, 0, 0]
class = 0
DATA_VALUE ≤ 20.5
gini = 0.664
samples = 273
value = [100, 91, 82]
class = 0
gini = 0.0
samples = 10
value = [0, 0, 10]
class = 2
DATA_VALUE ≤ 6.4
gini = 0.426
samples = 13
value = [9, 4, 0]
class = 0
gini = 0.0
samples = 9
value = [9, 0, 0]
class = 0
AGE ≤ 0.5
gini = 0.483
samples = 33
value = [11, 1, 21]
class = 2
DATA_VALUE ≤ 91.25
gini = 0.657
samples = 240
value = [89, 90, 61]
class = 1
DATA_VALUE ≤ 6.15
gini = 0.32
samples = 10
value = [8, 2, 0]
class = 0
gini = 0.444
samples = 3
value = [1, 2, 0]
class = 1
DATA_VALUE ≤ 20.15
gini = 0.381
samples = 25
value = [5, 1, 19]
class = 2
AGE ≤ 1.5
gini = 0.375
samples = 8
value = [6, 0, 2]
class = 0
AGE ≤ 1.5
gini = 0.658
samples = 233
value = [82, 90, 61]
class = 1
gini = 0.0
samples = 7
value = [7, 0, 0]
class = 0
gini = 0.5
samples = 4
value = [2, 2, 0]
class = 0
gini = 0.0
samples = 6
value = [6, 0, 0]
class = 0
LOCATIONDESC ≤ 0.5
gini = 0.31
samples = 22
value = [3, 1, 18]
class = 2
gini = 0.444
samples = 3
value = [2, 0, 1]
class = 0
gini = 0.0
samples = 6
value = [6, 0, 0]
class = 0
gini = 0.0
samples = 2
value = [0, 0, 2]
class = 2
DATA_VALUE ≤ 83.45
gini = 0.655
samples = 228
value = [82, 90, 56]
class = 1
gini = 0.0
samples = 5
value = [0, 0, 5]
class = 2
gini = 0.215
samples = 17
value = [1, 1, 15]
class = 2
gini = 0.48
samples = 5
value = [2, 0, 3]
class = 2
gini = 0.657
samples = 222
value = [82, 84, 56]
class = 1
gini = 0.0
samples = 6
value = [0, 6, 0]
class = 1

# Evaluation of Second Model: Decision Tree



- Accuracy score: 0.742268
- Each node is assigned a class (gender) based on whether they fit the criteria of the feature
- Gini is the impurity score (rate of samples that do not match the assigned class)
- R^2 score: 0.482

# Comparisons Between the Models

**Decision Tree R^2 = 0.482    Linear Regression R^2=0.728**

## Advantages of Decision Tree

- Decision Tree can capture complex relationships in the data
- It does not require much data preprocessing for feature engineering
- Decision tree can handle both categorical and numerical data

## Disadvantages of Decision Tree

- Prone to overfitting, especially with complex trees
- Decision tree is computationally expensive with large datasets
- It can be unstable and sensitive to small changes in the data

## Advantages of Linear Regression

- Linear Regression can provide insight into the relationship between the independent and dependent variables
- It is Computationally efficient and can handle large datasets
- Linear Regression model is simple and easy to understand model

## Disadvantages of Linear Regression

- Assumes a linear relationship between independent and dependent variables
- Can be affected by multicollinearity among independent variables
- This model cannot capture complex relationships in the data

# Possible Flaws With Data Itself

- Encoding with target variable with data that was unspecific. Some of the surveys marked male and female.
- Age bounds of some surveys overlapped.
- Difference in R2 Models