Group 9 Final Project Report

Peter Lu, Shan Santhakumar, Tristan Dull, Eden Fraczkiewicz
CS108, Winter 2024

Project Problem:

The goal of our project is to build a model to predict whether or not a person is diabetic or prediabetic and then identify any biases in the model that can compromise protected classes. Our specific dataset includes the protected classes of age and sex. We will also look to identify bias in other categories that may also be unfair to penalize a person for, such as education and income as well. If we do encounter bias in our model, we plan to make attempts to correct this bias.

Data Description:

Our data came clean with no missing values. Our dataset came from the UCI Machine Learning Repository and contained 253,680 rows and 22 features. There were 14 binary variables and 7 categorical variables, and 1 response variable.

Binary variables:

HighBP - 0 = no high, BP 1 = high BP
HighChol - 0 = no high cholesterol, 1 = high cholesterol
CholCheck - 0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years
Smoker - Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]. 0 = no, 1 = yes
Stroke - (Ever told) you had a stroke. 0 = no, 1 = yes
HeartDiseaseorAttack - coronary heart disease (CHD) or myocardial infarction (MI). 0 = no, 1 = yes
PhysActivity - physical activity in past 30 days - not including job. 0 = no, 1 = yes
Fruits - Consume Fruit 1 or more times per day, 0 = no, 1 = yes
Veggies - Consume Vegetables 1 or more times per day. 0 = no, 1 = yes
HvyAlcoholConsump - Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week). 0 = no, 1 = yes
AnyHealthcare - Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no, 1 = yes
NoDocbcCost - Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no, 1 = yes
DiffWall - Do you have serious difficulty walking or climbing stairs? 0 = no, 1 = yes
Sex - 0 = female, 1 = male

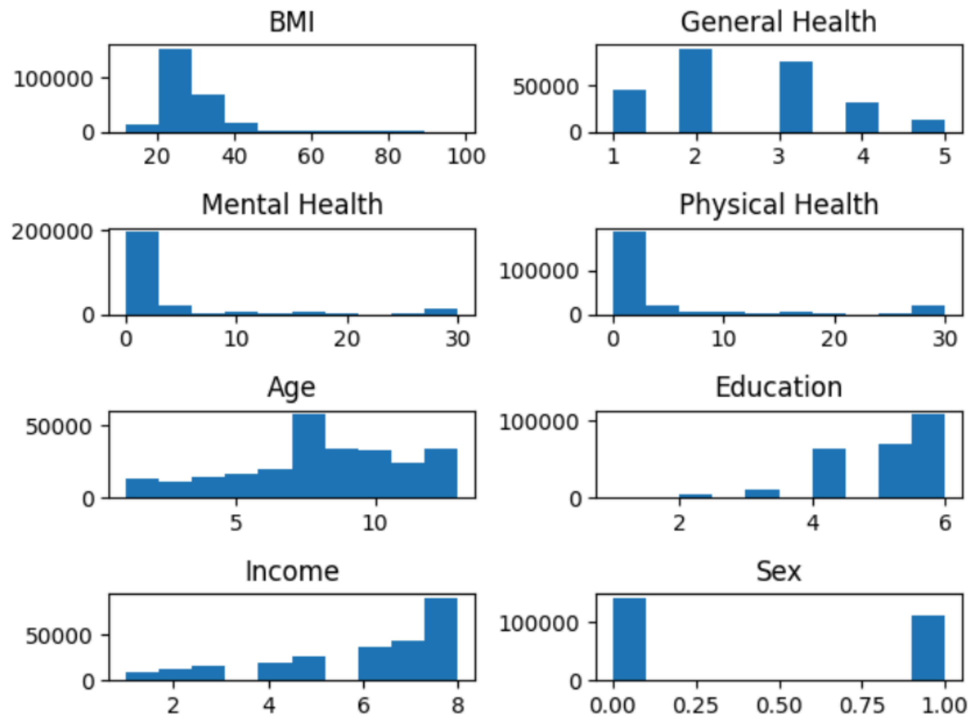
Categorical variables:

BMI - Body Mass Index

GenHlth - Would you say that in general your health is: scale 1-5, 1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor
MentHlth - Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days
PhysHlth - Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
Age - scale 1-13, 1 = 18-24, 9 = 60-64, 13 = 80 or older
Education - scale 1-6, 1 = Never attended school or only kindergarten, 2 = Grades 1 through 8 (Elementary), 3 = Grades 9 through 11 (Some high school), 4 = Grade 12 or GED (High school graduate), 5 = College 1 year to 3 years (Some college or technical school), 6 = College 4 years or more (College graduate)
Income - scale 1-8, 1 = less than 10,000, 5 = less than 35,000, 8 = 75,000 or more

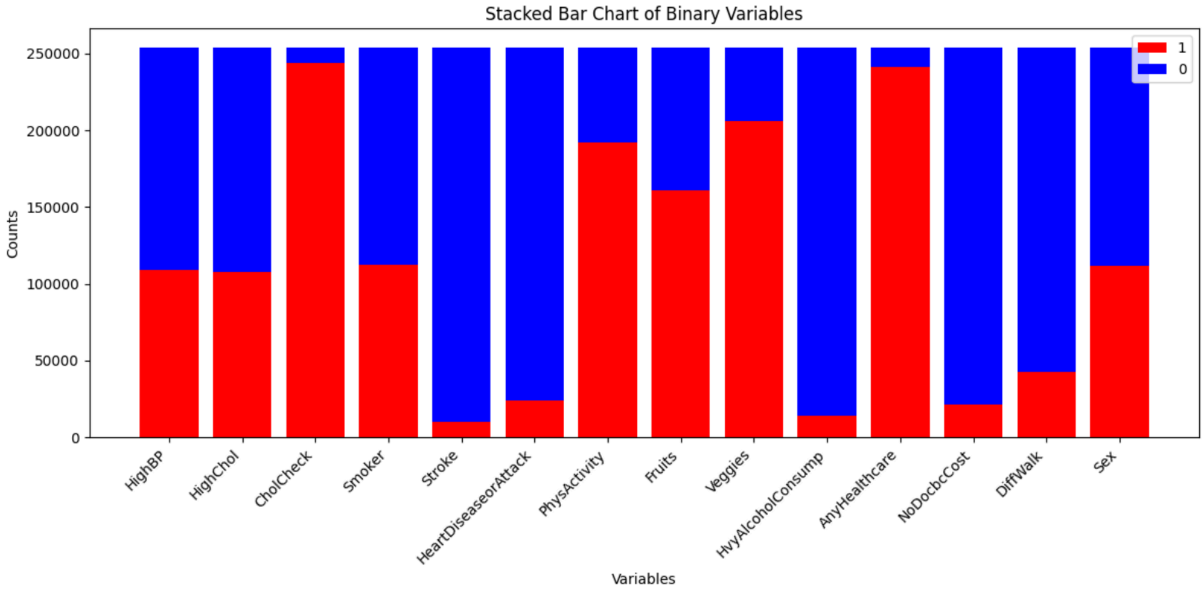
Response Variable:
Diabetes_binary - 0 = no diabetes, 1 = prediabetes or diabetes

EDA: Categorical Variables



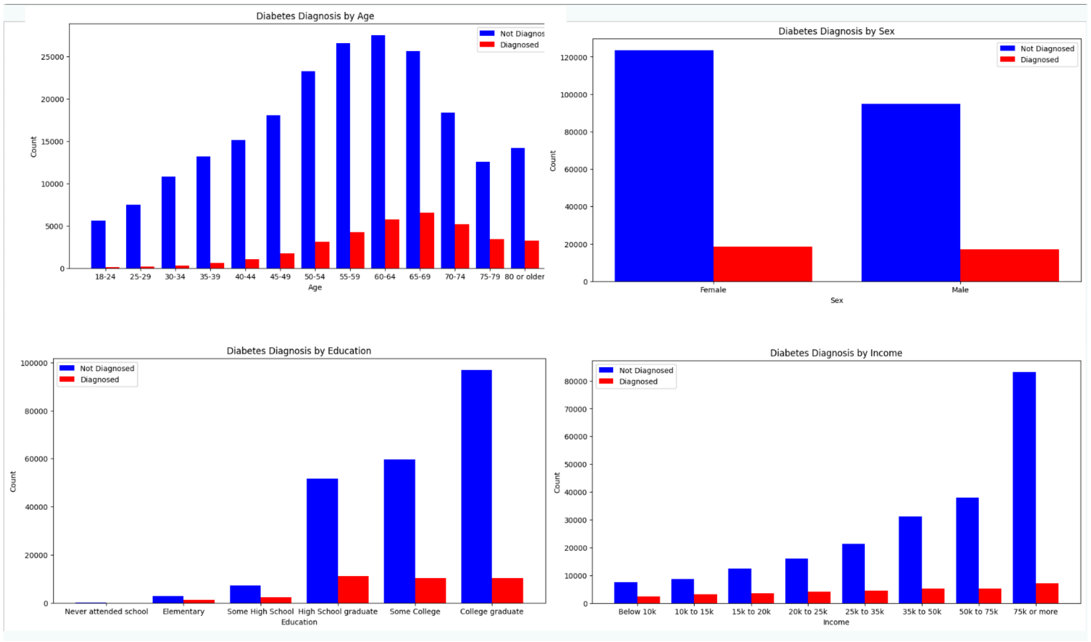
We see BMI, Mental Health, and Physical Health are all skewed to the right, while Education and Income are skewed to the left, and Age and General Health are roughly normally distributed. There are slightly more females than males.

EDA: Binary Variables



There seems to be a roughly even amount of those with high blood pressure, high cholesterol and have smoked at least 100 cigarettes in their life, and males. Slightly more people who have done physical activity outside of work in the past 30 days, eat fruit 1 or more times a day, and eat vegetables 1 or more times a day. A large majority have checked their cholesterol in the past 5 years and have some kind of health care coverage. A small minority have had a stroke, have coronary heart disease (CHD) or myocardial infarction (MI), are heavy drinkers, weren't able to afford to go to the doctor in the past year, or have serious difficulty walking or climbing stairs.

EDA: Response of Categories Intended to Protect



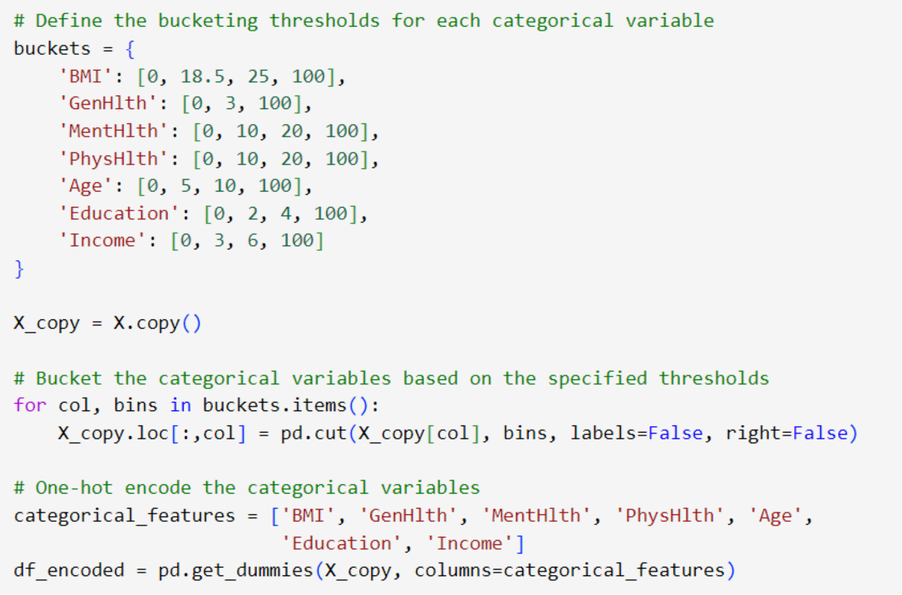
Ethical Considerations:

Seeing as our dataset does include sensitive information that may contain bias (Age, Sex, Education, and Income), we will keep in mind that after building the model, we will have to test it for fairness in these categories we would like to protect.

Approach:

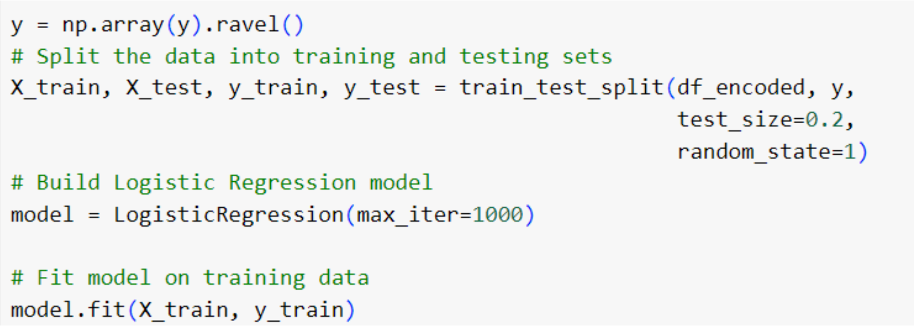
Since our data is mostly categorical, we will build a logistic regression model, as this type of model is best suited for categorical data.

We first one-hot encoded our categorical data to binary, with the idea that doing so would increase the model performance.



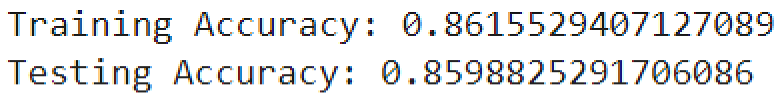
```
# Define the bucketing thresholds for each categorical variable
buckets = {
    'BMI': [0, 18.5, 25, 100],
    'GenHlth': [0, 3, 100],
    'MentHlth': [0, 10, 20, 100],
    'PhysHlth': [0, 10, 20, 100],
    'Age': [0, 5, 10, 100],
    'Education': [0, 2, 4, 100],
    'Income': [0, 3, 6, 100]
}

X_copy = X.copy()

# Bucket the categorical variables based on the specified thresholds
for col, bins in buckets.items():
    X_copy.loc[:,col] = pd.cut(X_copy[col], bins, labels=False, right=False)

# One-hot encode the categorical variables
categorical_features = ['BMI', 'GenHlth', 'MentHlth', 'PhysHlth', 'Age',
                        'Education', 'Income']
df_encoded = pd.get_dummies(X_copy, columns=categorical_features)
```

Then, we built our model and ran it with a 80:20 train test split.



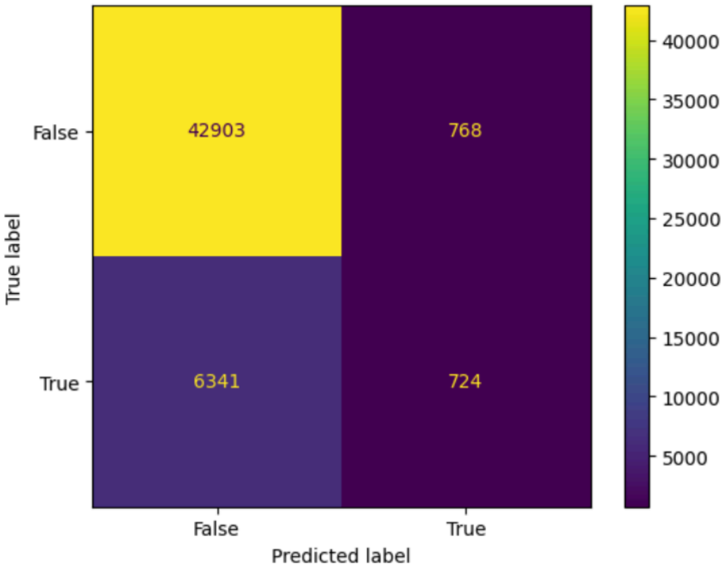
```
y = np.array(y).ravel()
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(df_encoded, y,
                                                    test_size=0.2,
                                                    random_state=1)
# Build Logistic Regression model
model = LogisticRegression(max_iter=1000)

# Fit model on training data
model.fit(X_train, y_train)
```

Our training accuracy was 86.1%, and our testing accuracy was 85.9%.



```
Training Accuracy: 0.8615529407127089
Testing Accuracy: 0.8598825291706086
```

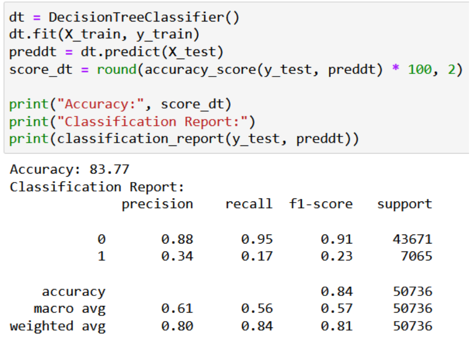
Which looked pretty good, however, after looking at our confusion matrix:



We found we had a true positive rate (TPR) of 0.1025, which means that only 10.25% of those with diabetes are identified as having diabetes in our model, which is not good.

Challenges:

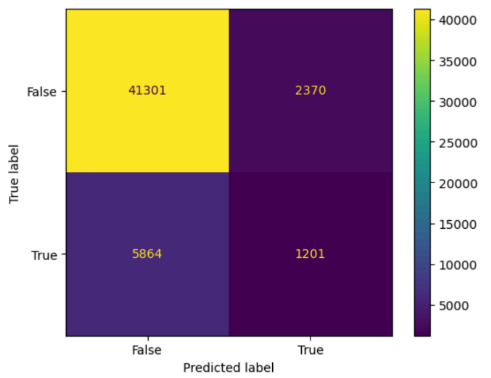
In order to tackle the issue of such a low TPR, we tried different models: a decision tree and a Multinomial Naive Bayes model:

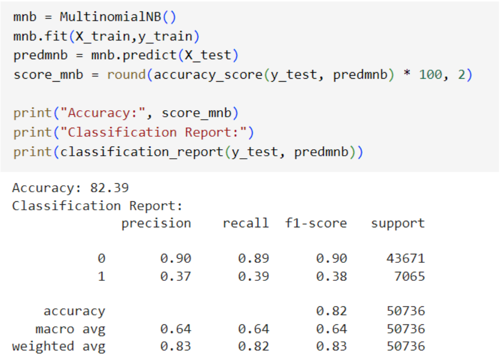


```
dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)
preddt = dt.predict(X_test)
score_dt = round(accuracy_score(y_test, preddt) * 100, 2)

print("Accuracy:", score_dt)
print("Classification Report:")
print(classification_report(y_test, preddt))
```

```
Accuracy: 83.77
Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.95      0.91     43671
           1       0.34      0.17      0.23      7065

    accuracy                           0.84     50736
   macro avg       0.61      0.56      0.57     50736
weighted avg       0.80      0.84      0.81     50736
```

Confusion Matrix for Decision Tree:

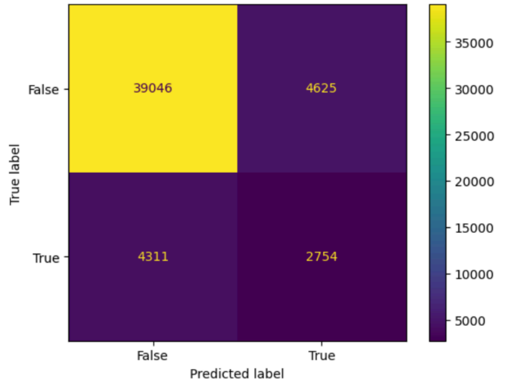




```
mnb = MultinomialNB()
mnb.fit(X_train,y_train)
predmnb = mnb.predict(X_test)
score_mnb = round(accuracy_score(y_test, predmnb) * 100, 2)

print("Accuracy:", score_mnb)
print("Classification Report:")
print(classification_report(y_test, predmnb))
```

```
Accuracy: 82.39
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.89      0.90     43671
           1       0.37      0.39      0.38      7065

    accuracy                           0.82     50736
   macro avg       0.64      0.64      0.64     50736
weighted avg       0.83      0.82      0.83     50736
```

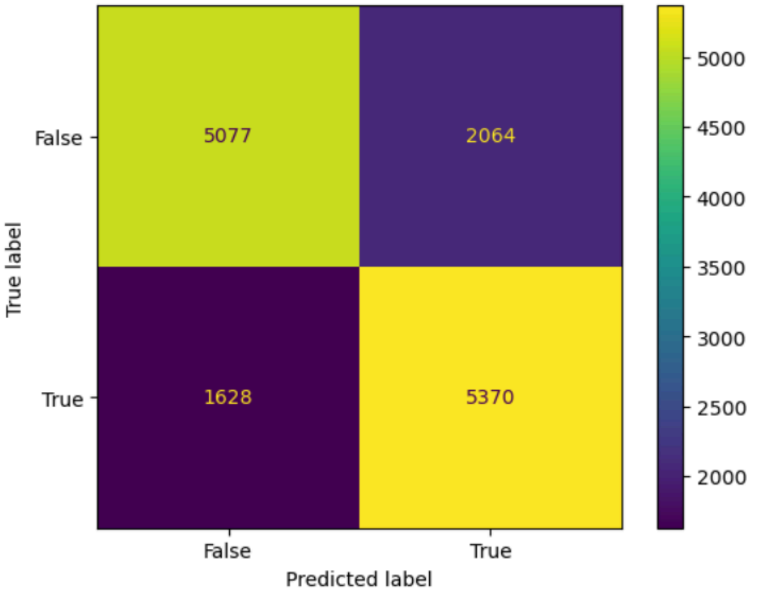
Confusion Matrix for Multinomial Naive Bayes:



The accuracies were fairly close to the logistic model, but the true positive rates were not to our liking. Only 0.1700 for the decision tree and 0.3898 for the Multinomial Naive Bayes model.

Seeing as different models were ineffective we opted to perform resampling. We decided to resample with respect to the diabetes_binary feature, and undersample the rows that were not diagnosed with diabetes to make things more even. Essentially, we would take all positive diabetes diagnoses and take a sample of those diagnosed to not have diabetes that is equal to those with diabetes and combine the 2 datasets. Then, we re-ran logistic regression with our new dataset.

Confusion Matrix for Logistic Regression Model w/ Undersampled Data:



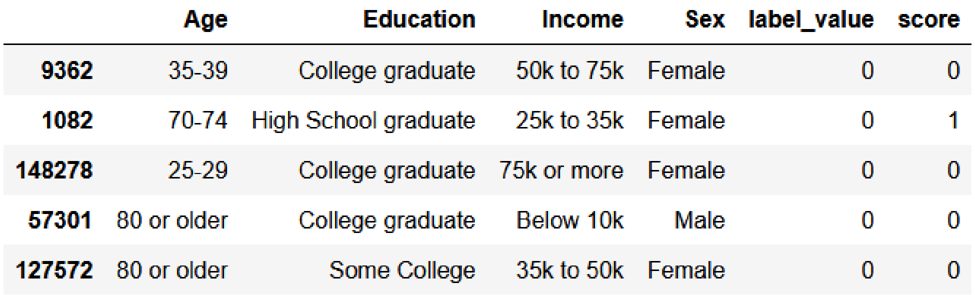
The TPR was successfully increased to 0.7674, which we were very happy with.

After undersampling we now have a new dataset that has 70,692 rows, and after retraining and rerunning the logistic regression model, we got a training accuracy of 73.9% and a testing accuracy of 73.8%. Although our model accuracy is lower, we were able to successfully increase our true positive rate.

Approach Continued:

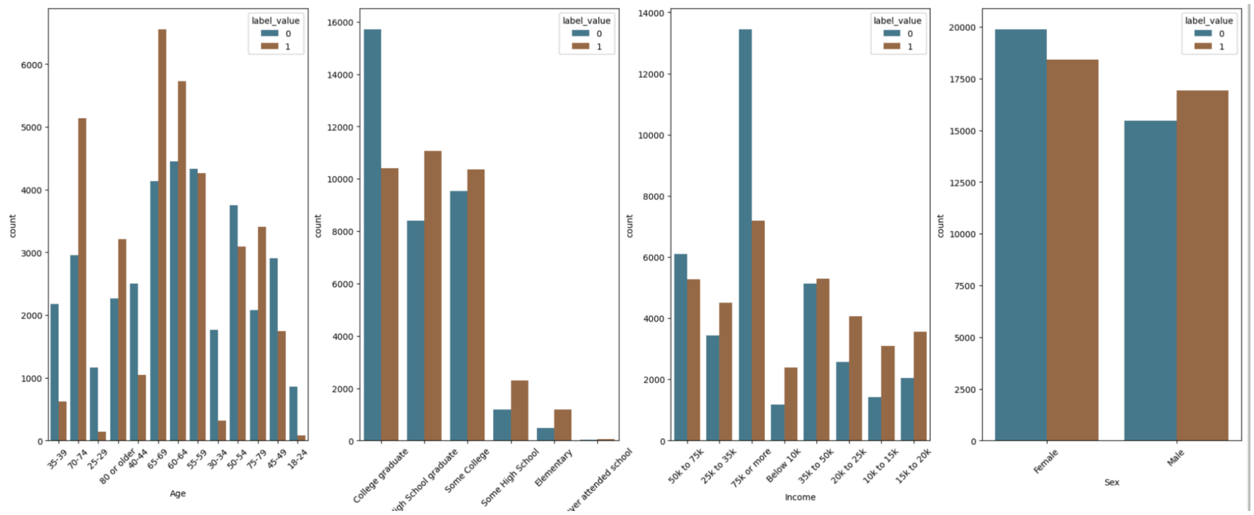
Now that we have our model, we will move on to analyzing fairness and bias in our model. We used the Aequitas python library to perform the analysis, and decided that the age, sex, education, and income features would be our features of interest, as age and sex are protected classes, and income and education may be sensitive classes, so we will look for bias in them as well.

First, we needed to create a dataset that Aequitas could use for the analysis. We recategorized our features and added label_value and score columns, which contained the actual diabetes_binary results and the predicted results, respectively

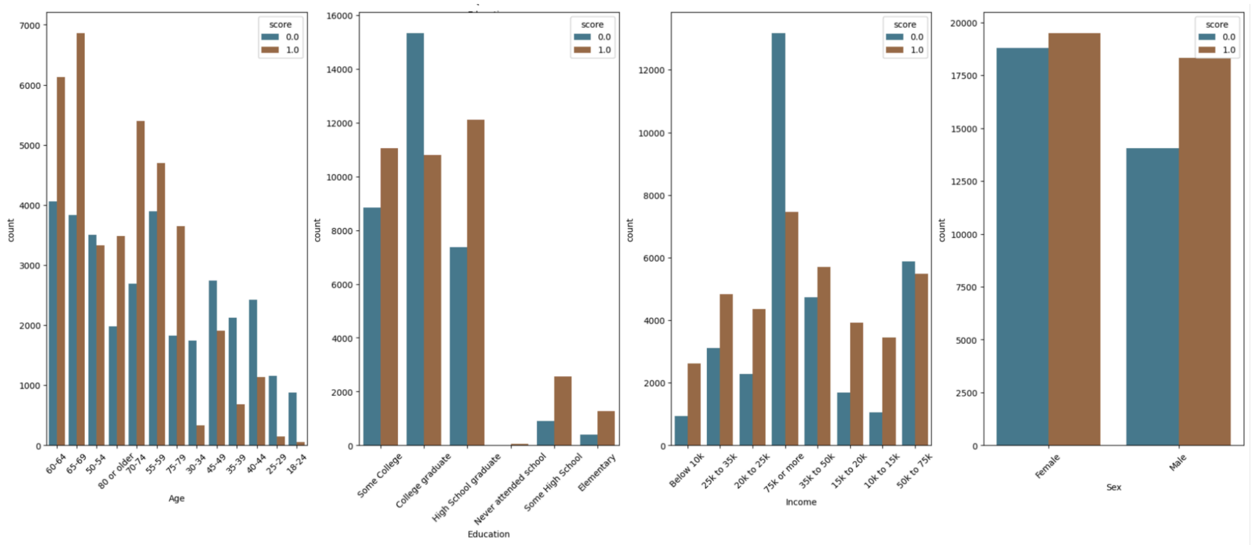


	Age	Education	Income	Sex	label_value	score
9362	35-39	College graduate	50k to 75k	Female	0	0
1082	70-74	High School graduate	25k to 35k	Female	0	1
148278	25-29	College graduate	75k or more	Female	0	0
57301	80 or older	College graduate	Below 10k	Male	0	0
127572	80 or older	Some College	35k to 50k	Female	0	0

Distributions of Label (True) Values:



Distributions of Score (Predicted) Values:

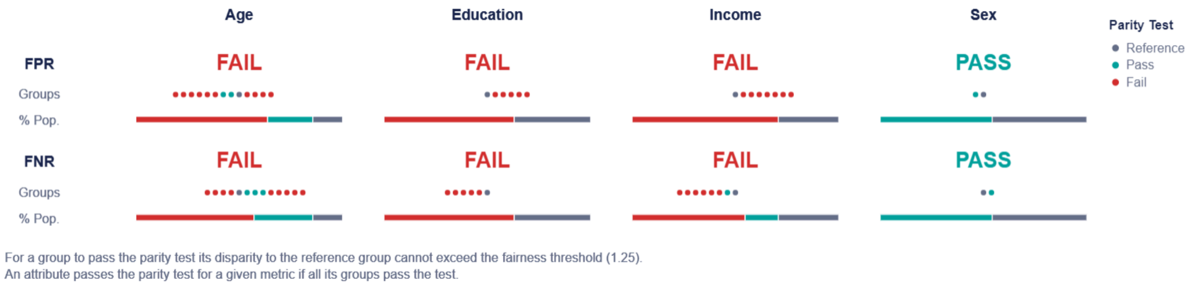


We can see that distributions of the label and score values are quite similar for our categories of interest. Age seems skewed to the lower ages with most of our sample being in their 60s. Education also seems skewed towards the lower side, with most of our sample having gone to college. Income seems roughly normally distributed, and sex is fairly even with slightly more females than males.

We choose to focus on the metric of misclassification rate in order to determine fairness. This means we will look at the false positive rate (FPR) and the false negative rate (FNR) to see if they are equal between different groups in a particular variable. This is to determine if there are certain groups that tend to be labeled inaccurately which would indicate bias against that group.

We used the summary plot feature in Aequitas to get a summary of these metrics and used a threshold of 1.25, meaning that if the FPR or FNR differs by more than a factor of 1.25 for different categories of a variable then that variable will be deemed unfair and fail the parity test.

Fairness Metrics of Unchanged Model:



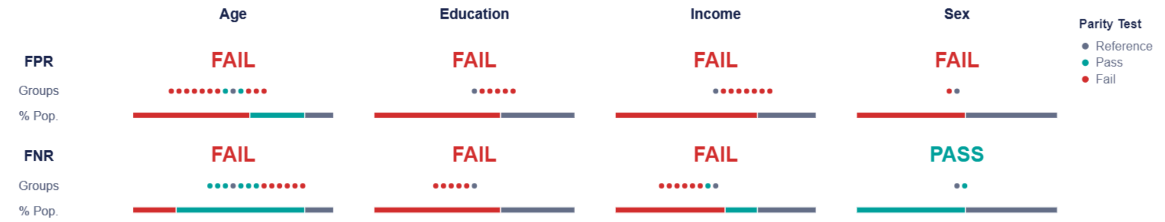
From our summary output we can see that Age, Education, and Income all fail both parity tests for FNR and are therefore unfairly treated in our model, whereas Sex is fairly treated and passes both parity tests.

We will attempt to remove the bias in Age, Education, and Income by dropping these columns from our model and reevaluating the model using the score values from a model built without these columns. We will try dropping the columns individually and all together as well.

Results:

Unfortunately, we were not able to remove bias from our model.

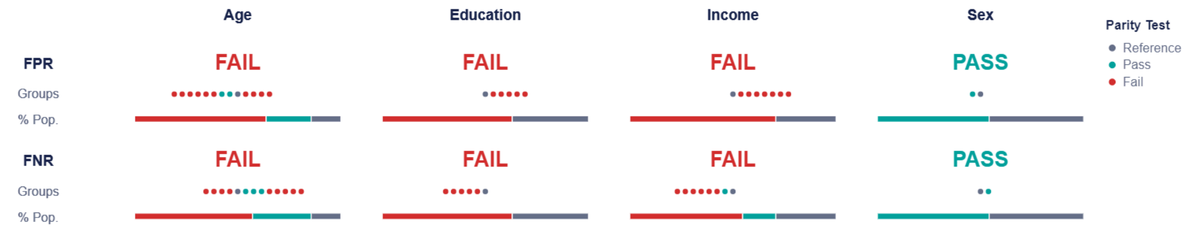
Fairness Metrics after dropping Age column:



For a group to pass the parity test its disparity to the reference group cannot exceed the fairness threshold (1.25).
An attribute passes the parity test for a given metric if all its groups pass the test.

After dropping the Age column, we see that the bias in the model still exists as Age, Education and Income fail our parity tests for FPR and FNR, and even Sex which used to be fair for both now becomes unfair in the FNR. Therefore dropping the Age column does not remove the bias from our data.

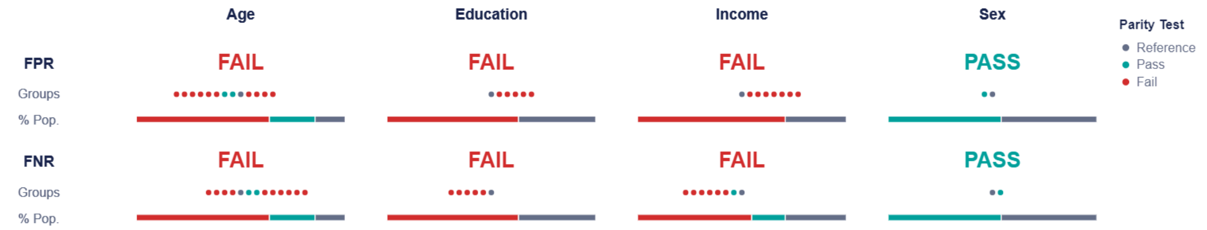
Fairness Metrics after dropping the Education column:



For a group to pass the parity test its disparity to the reference group cannot exceed the fairness threshold (1.25).
An attribute passes the parity test for a given metric if all its groups pass the test.

After dropping the Education column, we see that bias still remains in our model. Dropping the Education column does not help mitigate bias either.

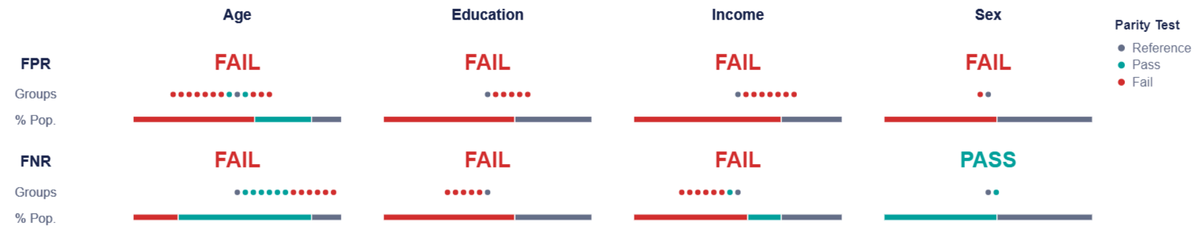
Fairness Metrics after dropping the Income column:



For a group to pass the parity test its disparity to the reference group cannot exceed the fairness threshold (1.25).
An attribute passes the parity test for a given metric if all its groups pass the test.

After dropping the Income column, bias still remains in our model. Dropping Income also does not remove bias from our model.

Fairness Metrics after dropping Age, Education, and Income columns:



For a group to pass the parity test its disparity to the reference group cannot exceed the fairness threshold (1.25).
An attribute passes the parity test for a given metric if all its groups pass the test.

We see that bias still remains in our model after dropping all 3: Age, Education, and Income. Therefore we can conclude that bias in our model cannot be removed by merely dropping columns and were thus unable to remove the bias.

Closing Thoughts:

While we were not able to remove the bias from our model we believe that it could be possible that the model is not what was being biased, but it may in fact be the data itself that was biased. Age, Education, and Income are all important factors that may be involved in an individual's likelihood of getting diabetes. As diabetes is a progressive disease, it would make sense that as one gets older, the probability of having diabetes would increase. Additionally, those with less education and income may be in a position where they either cannot afford to take care of their health as closely or are unaware of how to, which could also contribute to an increase in the likelihood of having diabetes. If this is the case, then our model would not be able to remove this bias as this is more of a societal issue which would need to involve changes in society to fix.

However, it is also important to note that merely dropping variables is not a rigorous enough approach to guarantee fairness of our model. Therefore, it could very likely be our model, which was not tuned properly to avoid bias, that caused the bias we detected in our analysis.