CS584 DATA MINING HW4: Fairness and Classification

\* Name: Shiva Shukla \* Miner Username: Vaathi \* GMU ID: G01321322

\* Rank: 23 \* Public Score: 0.86

# **Steps to run the code**

1. Download HW4Fairness\_Classification.ipynb file
2. Open the notebook using Jupyter Notebook or Google Colab
3. Upload the train.csv and test.csv file to Google Colab or Jupyter Notebook
4. Run the headers: “Import Required Libraries” to “Test Data Prediction”

# **Introduction**

The assignment handles imbalanced census dataset to classify if the person has income greater than 50000 or less than 50000. The data set has 32561 rows, 13 features, and label(income). There are protected features that may provide a bias in prediction: gender, race, and features correlated to them. In this assignment, I have predicted the accuracy using the features(part-1), analysed three fairness metrics: demographic parity, equal opportunity, and equality of odds, to identify the impact of protected features(part-2), and tried to mitigate the bias(part-3).

# **Part-1**

1. Import required packages and libraries and read the data.
2. **Pre-processing**:
   1. The data set contains a lot of missing values marked with ‘?’. First, I replaced these values with NaN and dropped them. But due to a decrease in the training dataset, there was a reduction in accuracy. Then, I replaced the ‘?’ values with 0. This will be handled in the categorical encoding of the features.
   2. The education feature is a categorical feature that is the same as education-num. So, I dropped it.
   3. Dataset has features with string values. Since the values are categorical, I used an Ordinal Encoder to encode all the features except race, gender, and income. In race and gender, I manually mapped majority class(based on count: white, male) to 1 and minority(based on count: others, female) to 0. In income, I mapped ≤50000 to 0 and >50000 to 1.
   4. The features have data with varying ranges. I used a standard scaler to get a similar data distribution across all features. I also tried the min-max scaler, but the accuracy didn’t improve(85%). With standard scaler, I got 86%.
3. **Observation**: I used 6 classifiers: Decision Tree, AdaBoost, Random Forest, Logistic Regression, K-Nearest Neighbour, and Linear SVM. I got the best accuracy using Random Forest Classifier(86%). Along with retesting with different parameters, I used Grid CV to find the best hyper parameters: {'criterion': 'entropy', 'max\_depth': 15, 'n\_estimators': 150}. To handle class imbalance, I used class\_weight = {0:1, 1:1.5}.
   1. **Metrics**: I calculated the accuracy, f1-score, and recall for all the classifiers. Using the best classifier random forest, I also calculated the metrics for two cases: dropping protected attributes: race and gender(RFC\_1) and dropping protected and correlated features(RFC\_2)(Fig:1, Fig:2, Fig:3).

**Chart, bar chart

Description automatically generated**

**Chart

Description automatically generated Chart

Description automatically generated**

1. **Results**: I got the best results for Random Forest Classifier. Accuracy and F1-Score were highest but recall for 0 was slightly low for Random Forest. Decision tree gave an accuracy of 86%, but F1-Score and recall were low.

1. **Part-2**
2. **Pre-processing**: The same dataset is used to analyse the fairness of all the six classifiers. To analyse the bias on majority and minority, I classified the race as the majority(1, white) and minority(0, others) and gender as the majority(1, male) and minority(0, female). The pre-processed data from part-1 is used to analyse the fairness metrics: demographic parity, equal opportunity, and equality of odds.
3. **Observation**: These are the results for attributes gender(Fig:4) and race(Fig:5). The opportunity is the equality of the true positive rate, and odds is the equality of the true positive rate, and false-positive rate. The results showed the difference between the true positive rate as the opportunity and the false positive rate as the odds. Random forest predicts with maximum accuracy but has some disparity. The metrics on the prediction show that there is some gender bias. The male population has a higher chance of getting an income of >50000 than the female population. The opportunity results show that the model correctly predicts the male population having income greater than 50000, a bias. Logistic regression has the worst prediction. From Fig:4, logistic regression has the maximum difference in opportunity showing heavy bias for the majority gender(male). Decision tree and Adaboost have low bias than random forest but poor accuracy and f1-score.

From Fig:5, we can see that race also introduce some bias as white people have a higher chance of having an income > 50000. But it has less impact on fairness as compared to gender.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

1. **Results**: There is gender bias in the data due to which males have a higher chance of having an income > 50000. Race also has some bias in favour of the majority class, but it is less comparing to gender.
2. **Part-3**
3. **Pre-processing**: To find out the features that correlated the most with the protected attributes gender and race, I used mutual\_info\_classif from Sklearn feature selection. It shows that marital-status, occupation, and relationship, correlate the most with the protected features gender and race.

Chart, bar chart

Description automatically generated

1. **Observations**: I calculated the results using two strategies:
   1. **Removed Race and Gender:** In Fig:4 and Fig:5, RFC\_1 represents fairness metrics when I removed the race and gender on the best classifier from part-1, Random Forest Classifier. There is a minor reduction in gender bias. The impact of race also reduced, but it was less comparing to gender. Accuracy, F1-score, and recall obtained here are almost comparable to part-1 since, except for race and gender, I used all the remaining features.
   2. **Removed Race, Gender, Marital-Status, Occupation, Relationship**: From Fig:6, the features: marital-status, occupation, and relationship has the maximum correlation with the protected attributes. After removing the mentioned features, fairness improved, as shown in Fig:4 and Fig:5 under RFC\_2. There was a significant decrease in the demographic parity and opportunity. The odds(false positive rate) approximately decreased to 0.
2. **Results**: After removing the protected attributes and correlated features, bias reduced significantly for gender and slightly for the race. Accuracy was reduced to 83% when I reduced the features for training the model showing a fairness-accuracy trade-off.

|  |  |
| --- | --- |
| **Models** | **Accuracy** |
| Random Forest Classifier( max\_depth = 15, n\_estimators=150,random\_state=42, class\_weight={0:1, 1:1.5}, criterion ='entropy'), removed education feature, Standard Scaler, Ordinal Encoder | 0.86 |
| Random Forest Classifier( max\_depth = 15, n\_estimators=150,random\_state=42, class\_weight={0:1, 1:1.5}, criterion ='entropy'), removed education feature, Min-Max Scaler, Ordinal Encoder | 0.85 |
| Random Forest Classifier( max\_depth = 15, n\_estimators=150,random\_state=42, class\_weight={0:1, 1:1.5}, criterion ='entropy'), removed sensitive features, Min-Max Scaler, Ordinal Encoder | 0.83 |
| Random Forest Classifier( max\_depth = 15, n\_estimators=150,random\_state=42, class\_weight={0:1, 1:1.5}, criterion ='entropy'), removed education feature, Standard Scaler, SMOTE Oversampling Ordinal Encoder | 0.77 |
| Decision Tree(max\_depth = 7), removed education feature, Standard Scaler, Ordinal Encoder | 0.86 |

# **Results**

1. Random Forest and Decision Tree gave an accuracy of 86%. But the decision tree had a low F1-Score and Recall.
2. The decision tree had a better result in few fairness metrics in the validation set, but the random forest gave the best results in the test data. Also, after removing the protected and correlated attributes, random forest reduced the bias most effectively in both validation and test dataset.
3. After removing protected attributes, accuracy was reduced to 83%, showing a fairness-accuracy trade-off.

# **Conclusion**

Successfully developed and analysed a model that reduced the bias in the data.