Team Members: Chuan Chen (cc6580) and Yihang Zhang (yz2865)

Date: 3/24/2021

Project Title: Investigating Bias in Insurance Premium Prediction

Data:

The Data we will be looking at is found here https://www.kaggle.com/noordeen/insurance-premium-prediction. This dataset contains 1338 samples and 7 features. Four of the features are numerical including age, BMI, children, and expenses. Three of the features are nominal including sex, smoker, and region. The goal of this system is to analyze the relationship between different features and how they impact medical expenses, in order to help insurance companies compute predicted medical expenses in the future and price their premiums accordingly.

ADS:

The code we will be analyzing for this project is found here https://www.kaggle.com/klmsathishkumar/predicting-insurance-premium. The code implemented several different machine learning algorithms but ultimately suggested the Random Forest model as it gave the best accuracy. Thus, we will be focusing our analysis on the potential bias given by the chosen Random Forest model.

To fairly access the bias of the ADS, we may require literature references from outside sources on relationships between protected attributes and relative medical conditions and other relevant information.

Motivation:

We chose the insurance premium data because its features include several sensitive attributes, including age, sex, and region, of which there exhibit known minority groups. We have seen several examples of bias in ADS in class, and It will be interesting to investigate if there is known bias in charging different groups different premiums and analyze what factors might have caused such bias. We chose our ADS because, despite its thorough analysis of performance for different machine learning models, it used accuracy as the performance metric. Yet, from the examples we have seen in class, accuracy can sometimes mask the presence of bias in the ADS. For example, two groups can exhibit similar accuracies in their predicted medical expenses, yet one group is always priced lower but the other group is always priced higher. All of the reasons above motivated our group to select this ADS for our bias analysis.