Dawei Huang
05/05/2020
Ethics Project Proposal

Project Proposal: Interpretability of Machine Learning

For this project, I hope to consider a new definition for the interpretability of machine learning where interpretability is the accuracy to which the machine learning models may predict results when compared to real results. Therefore, the interpretability of our machine learning models are high when it can accurately predict results that align closely with real world phenomena. The metric that I will propose to compare the interpretability of each machine learning model will simply be the accuracy of the model when compared to real world results. I hope to compare this definition to the most common definition where the integrability of machine learning is determined by how closely a human will predict the same outcomes.

For these possible data sets that I may use could be college admissions data where machine learning models may be used to predict whether or not certain applicants were accepted or not over others. I plan to use the most common types of machine learning models such as decision trees, logistic regression, and Random Forest and I will be testing their interpretability. I will be using python to code these different types of machine learning models and for analysis of their results.