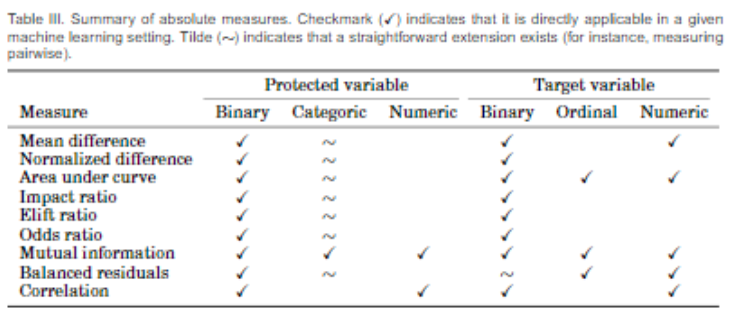
Methodology:

1. Statistical method [(refere nce)](https://clarkuedu-my.sharepoint.com/:b:/g/personal/jazhang_clarku_edu/EWZ_MhuwdvFBqI5ZO6F8DX8Bzaz3W2ZbR-J8Vr7Cvo5QdQ?e=agKJ5E) ([reference](https://clarkuedu-my.sharepoint.com/:b:/g/personal/jazhang_clarku_edu/EX942CtXSrBFjJbWApXYC0EBeNx1xLyT9mkNI61P3EEUjg?e=Ln12FI))
   1. Applying the machine learning algorithms to the original data set and obtaining the result.
   2. Selecting the features/attributes/parameters we would like to focus on. Separating the groups in these features.
   3. Run statistical test for each feature that we would like to analyze.
      1. Regression slop test – test if the regression coefficient of protected variable is significantly different from zero (what if the relationship is non-linear?).
      2. Difference of means test – test if the means of the two groups are equal (assumes independent samples, normality and equal variances, test required)
      3. Difference in proportions for two groups or many groups - test if the rates of positive outcomes within the two/more groups are equal.
   4. If the statistical test is applicable, we calculate the absolute measures and the conditional measures.
      1. Absolute measure – capture the magnitude of the differences between (typically two) groups of people.



* + 1. Conditional measures – capture how much of the difference between the groups is explainable of other characteristics of individuals

**2.Evaluate the TPR disparity** – the difference in true positive rates (TPR)，the false positive rate of a non-diagnosis – among different protected attributes such as patient sex, age, race, and insurance type. Reference:(<https://arxiv.org/pdf/2003.00827.pdf>)

First, we examine the differences in true positive rate (TPR) across different subgroups per attributes: A high TPR disparity indicates that sick members of a protected subgroup would not be given correct diagnoses—e.g., true positives—at the same rate as the general population, even in an algorithm with high overall accuracy.

TPR disparities for binary attributes. For binary attributes, we quantify the TPR disparity as the difference between TPR of sexд and∼д,per label yi,i ∈ {1,...}.Then, with random variablesYˆ and Y denoting the predicted and ground ii truth labels for yi , the TPR of sex д per disease yi , is TPR д, yi = P[Yˆ = y |G = д,Y = y ], and the associated TPR sex disparity is, Gapд,yi = TPRд,yi − TPR∼д,yi

TPR disparities for non-binary attributes. For non-binary attributes, we use the difference between a subgroup’s TPR and the median (as measure of central tendency) of all TPRs to de ne TPR disparity, GapSj,yi = TPRSj,yi −Median(TPRS1,..,TPR SNk )i.

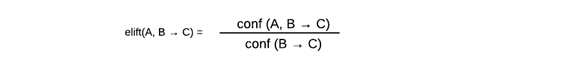
Second, we measure the FPR of various subgroups, where they are predicted not to have any diagnoses despite showing signs and symptoms of conditions

Association rule:

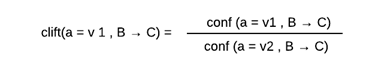
[*Reference: Measuring Discrimination in Socially-Sensitive Decision Records*](https://clarkuedu-my.sharepoint.com/personal/jazhang_clarku_edu/_layouts/15/onedrive.aspx?ct=1587756053530&or=OWA%2DNT&cid=51b7a8e4%2D60a0%2D3fbf%2D0dd6%2Daa8417de29a7&originalPath=aHR0cHM6Ly9jbGFya3VlZHUtbXkuc2hhcmVwb2ludC5jb20vOmY6L2cvcGVyc29uYWwvamF6aGFuZ19jbGFya3VfZWR1L0VqbFA3ODBadTBCSnFaRUUydGlwTWw4Qk1CS21fNDUyOVZnRnQ4YXNLNlRmR3c%5FcnRpbWU9c2ZsWm1ZVG8xMGc&id=%2Fpersonal%2Fjazhang%5Fclarku%5Fedu%2FDocuments%2FMachine%20Learning%20Discrimination%20Analysis%2FLiterature%2FMeasuring%20Discrimination%20in%20Socially%2DSensitive%20Decision%20Records%2Epdf&parent=%2Fpersonal%2Fjazhang%5Fclarku%5Fedu%2FDocuments%2FMachine%20Learning%20Discrimination%20Analysis%2FLiterature)

Extend lift: Let A, B → C be an association rule such that conf (B → C) > 0. We define the extended lift of the rule with respect to B as: conf (A, B → C) / conf (B → C). We call B the context, and B → C the base-rule.

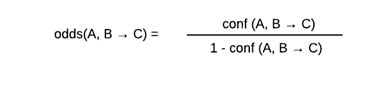
Extend lift express the relative variation of confidence due to the extra item in the premise of the base rule B → C:

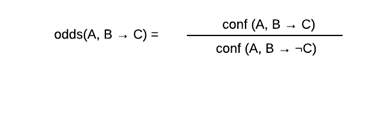


Let a = v1, B → C be a classification rule, and v2 ∈ dom(a) with conf (a = v2, B → C) minimal and non-zero. Classified lift:



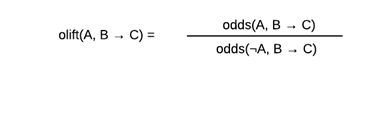
Odds ratio:





Let A, B → C be a classification rule with conf (¬A, B → C) > 0 and conf (A, B → C) < 1.

The odds lift of the rule is:



Difference measures: in the U.K., a difference of 5% in confidence between female (A is sex=female) and male (¬A is sex=female) treatment is assumed by courts as significant of discrimination against women.

eliftd(A, B → C) = conf (A, B → C) − conf (B → C)

sliftd(A, B → C) = conf (A, B → C) − conf (¬A, B → C)