# Project Proposal

ML 1 - Yilmaz
2024 - 2025
Period 4
Members: Abhinav Palikala, Justin Lee

**Motivation: What problem are you tackling?**

GainRatio is not the best heuristic for selecting what attribute to split a decision tree on. Using this heuristic can result in inefficient trees that are less accurate as well as more computationally expensive compared to  better trees. We should define a heuristic that factors in what attributes are left to split on to make more informed decisions about what attribute a tree should split on. This would result in faster, simpler trees, which would help accuracy, time complexity, and interpretability.

**Method: What machine learning techniques are you planning to improve upon?**

We are planning to improve upon the decision tree method. In particular, we are planning to improve upon the method in which it determines the optimal attribute to split upon. In particular, the current method, Info Gain, attempts to quantify the amount of information gained from analyzing a particular attribute. Instead, we may improve our choice by weighting future attributes in our initial consideration. In a sense, this will also value the ‘distinctness’ of attribute knowledge.

**Intended experiments: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?**

We want to test decision trees on a variety of different datasets, some with a lot of attributes and some with a few. After constructing decision trees using several different heuristics, like our heuristic, GainRatio, and InfoGain, we want to see which one produces more accurate trees as well as more efficient trees. We can check the efficiency of trees by seeing the depth of each branch and prioritizing trees with lower average depths. We will also construct random forests with our decision trees to see if a more efficient decision tree also results in more accurate

random forests, or if decreasing randomness through a better heuristic negatively affects random forests.

**Presenting pointers to one relevant dataset and one example of prior research on the topic are a valuable addition.**

**Dataset:**

**Any dataset will be valid, as long as we remain consistent**

**Relevant Research:**

[1206.4620] Improved Information Gain Estimates for Decision Tree Induction

(PDF) The Use of Heuristics in Decision Tree Learning Optimization

More effective way to improve the heuristics of an AI... evolution or testing between thousands of pre-determined sets of heuristics? - Artificial Intelligence Stack Exchange

Improved Information Gain Estimates for Decision Tree Induction