Matan Lagnado
Salena Torres Ashton
ISTA 421: Introductory Machine Learning
12 May 2025

Using Decision Trees to Understand Hubble Constant Measurements

The Hubble constant is one of the most important measurements in cosmology today. It relates distances of galaxies to the speeds that they move away from us, or recessional velocity. Developing accurate measurements of this constant will help accurately derive laws in General Relativity which are fundamental to understanding gravity. The laws of gravity are something that are of great importance to understand for present day technology. Things like satellites, GPS, and renewable energy sources are all dependent on gravity. Our society today greatly depends on gravity so measuring the Hubble Constant accurately, gives us the opportunity to create more efficient and precise GPS and satellite systems.

Congress controls the NASA budget. NASA would have the necessary experience and means to construct a telescope with the goal to measure this important constant. However, congress tends to pull funding from NASA projects that do not immediately benefit the U.S. global image. Currently NASA's budget is less than 1% of the total U.S. budget which is following a decreasing trend over the last few calendar years. NASA is vastly important to advances in science and technology because it has the budget to try new things. (Wikipedia 2025) Taking risks on new ground breaking projects is how we developed our space systems and new technology such as GPS, weather satellites, IR thermometers, and MRI imaging. There is no telling what technology might come out of trying to develop new powerful telescopes that help in measuring gravity. Machine learning would be a great tool to connect both experts and those inexperienced in cosmology. Specifically a tool like a decision tree would bridge the gap.

In this project I used the NED-D dataset. A Master List of Redshift-Independent Extragalactic Distances. (Steer 2020) NED is the NASA/IPAC Extragalactic Database. The dataset has 15 columns and 328,318 rows ranging from the year 1963 to 2018. The features in this dataset include: Exclusion Code, D, G, Galaxy ID, m-M, err, D (Mpc), Method, REFCODE, SNID, redshift (z), Hubble const., Adopted LMC modulus, Date (Yr. - 1980), and Notes. However, a lot of these features were omitted such as REFCODE, D, G, and more, this is because they are identifiers describing what is being measured; either the star, or galaxy being observed. The data comes from sky surveys and studies that attempt to measure galaxy distances. As a byproduct of measuring Galaxy distances and also redshift, which can tell us about recessional velocity, we can make predictions on the Hubble Constant. Many of the studies however do not use the data to make a prediction that could be because the data was too error prone, or the method is known to be unreliable for this kind of prediction. Whatever the reason may be it presents the opportunity to differentiate between studies that do and do not make these predictions.

Differentiating between true and false would be a lot simpler to understand than making a numerical Hubble Constant prediction. This project thus would show a congressman how small differences in the construction of a telescope and thus the measurements they can make will help in accurately predicting the Hubble Constant.

To explore this dataset I first confirmed the simplified linear relationship between the distances and velocities (linear regression homework). However velocity is not a given column so we must add a new column that converts the redshift into a velocity. Redshift is related to recessional velocity via the following relationship.

$$z = \frac{v}{c}$$

Where z is redshift, v is recessional velocity, and c is the speed of light. Using a linear regression we can then make a prediction on the value of the Hubble Constant using the Hubble's law which is:

$$v = H_0 D$$

Where v is the recessional velocity of a galaxy, D is the distance to that galaxy, and H_0 is the Hubble Constant. However to benefit the future of instrumentation we continue to make the decision tree.

The dataset contained a few columns that were not usable for this project, specifically the identifier columns that describe what star was being measured, what galaxy the star was in, and various others. Due to the differences between studies many rows also contain missing values. This could be because the goal of the study was to measure a distance and not the Hubble Constant. For missing values in the Hubble Constant column we simply set them to zero which comes in handy later on. However, in other important columns if there is a missing value we remove that row. This is because if we were to start a mission with the goal to measure the Hubble Constant we would want to compare it to other studies that made similar measurements. This left us with around 30,000 rows to work with. Next, we transformed any nonzero value in the Hubble Constant column to 1 giving a true or false column for the Hubble Constant.

One interesting visualization that was made is plotting the distances versus the velocities. This shows the linear nature of Hubble's Law which is still being explored today.

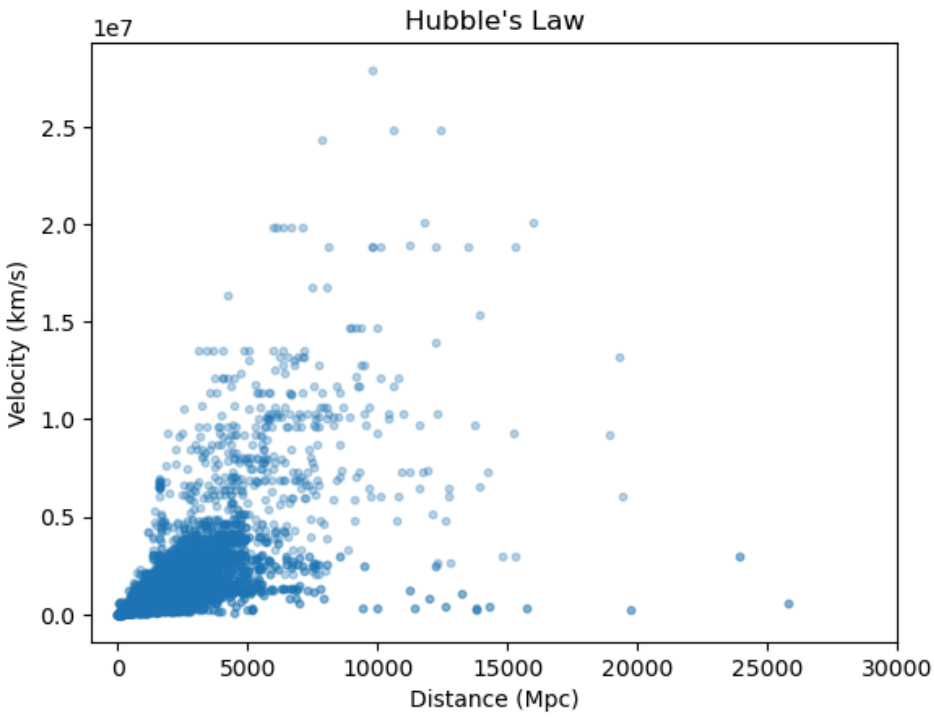

Figure 1 - Plot of all distance (Mpc) and velocity (km/s) measurements in the dataset to show the linear relation of Hubble's Law

Additionally, the plot depicting the linear regression line can be seen which predicts a Hubble Constant of 74.8 km/s/kpc. This is higher than today's estimates but still relatively close to the accepted value of 73.2 km/s/kpc which has decreased over time, first being predicted at 500 km/s/kpc.

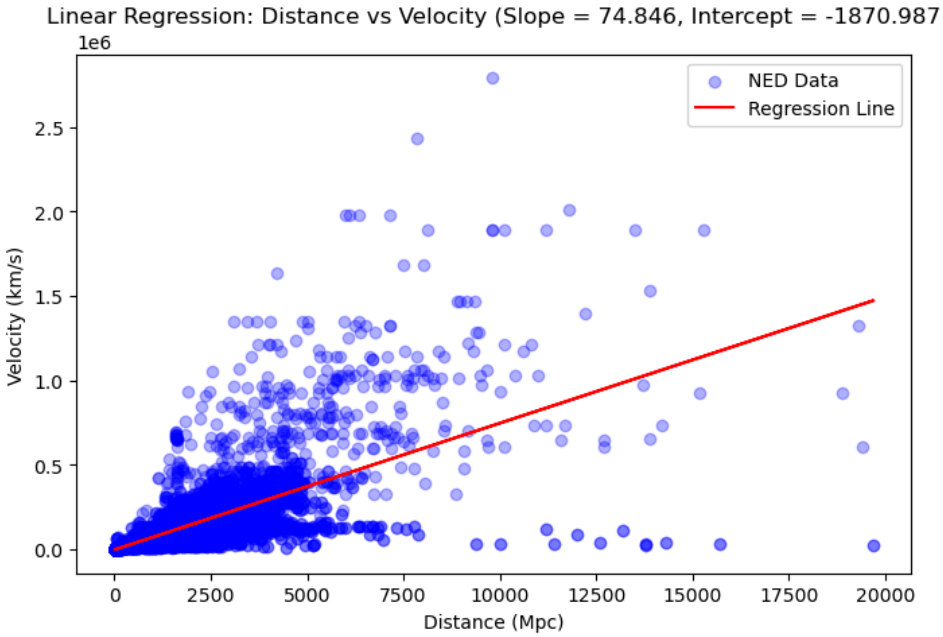

Figure 2 - Plot of all distances (Mpc) and velocities (km/s) in the dataset (blue) including a linear regression line (red) whose slope should be the Hubble Constant.Values beyond 20,000 kpc are not shown as there are few measurements of this nature and they stretch the viewing window of the figure.

To answer this question I utilized a decision tree. A decision tree can be an effective tool in communicating to congress who are not experts in cosmology. This question could have been answered using a logistic regression model as well, however because of the target audience it seemed more valuable to use a decision tree instead of explaining the math behind logistic regression. The model can still utilize the numerical columns to create splits in the data. The decision tree model effectively translates to beginners, splits the data through numerical columns, and is easy to display thousands of data points.

To create the decision tree model we must understand the Gini index quantity. This quantity will be used to justify any splits made on the decision tree. The gini index varies between 0 and 1. A 0 tells us that the set of data is 'pure' meaning that either there is only 1 value in the set or all the data in the set is the same. However, as the gini index approaches 1 it tells us that there are more and more unique values in the dataset. The gini index is calculated with the equation:

$$G = 1 - \sum_{i=1}^{N} P_i^2$$

Where i is one class of data so in this project zeros or ones, and P is the probability of a datapoint being in that class. The probability is calculated by the total count of a class over the total number of data points.

Next we have to understand how a decision tree decides to make a split or node. It will be a computationally expensive process however it will be precise and greedy. A greedy decision tree will look for the current best split and move in that direction. Although this may not be the best possible tree it will save on computation power by not trimming the tree later on. To create a decision tree node first we look at all the features and for each feature we look at every unique value. We will then split the data on each of these unique values where any value less than or equal to the value will be placed in one bin and the rest of the values go in another bin. Now for each of these bins we calculate the gini index. We then weight these two gini indexes by their size and add them together. Now we have one metric that describes the entire node. This lengthy iterative process then takes the smallest gini index that was calculated and takes that as the best possible split.

To actually build the tree we create a function that also uses an iterative process. We have to define a maximum depth that we will allow our tree to be. If it is too large it will become increasingly computationally expensive. Additionally the tree should also not split data if all the data on a node or leaf is identical. The tree first takes all the data and makes the greediest split, then stores the information from the split. Next, the subsets created from the split data calculate the next greediest split and store that data.

This process continues until either the maximum depth is reached or we have a well defined tree.

To evaluate the model we want to take a look at the accuracy on a random set of data. To do this we will train the tree using 80% of the data and test it on the remaining 20%. This is simple to do but we also want to understand how adding extra depth to the trees impacts our results. On top of calculating accuracy we can just as easily calculate the precision, recall, and f1 score.

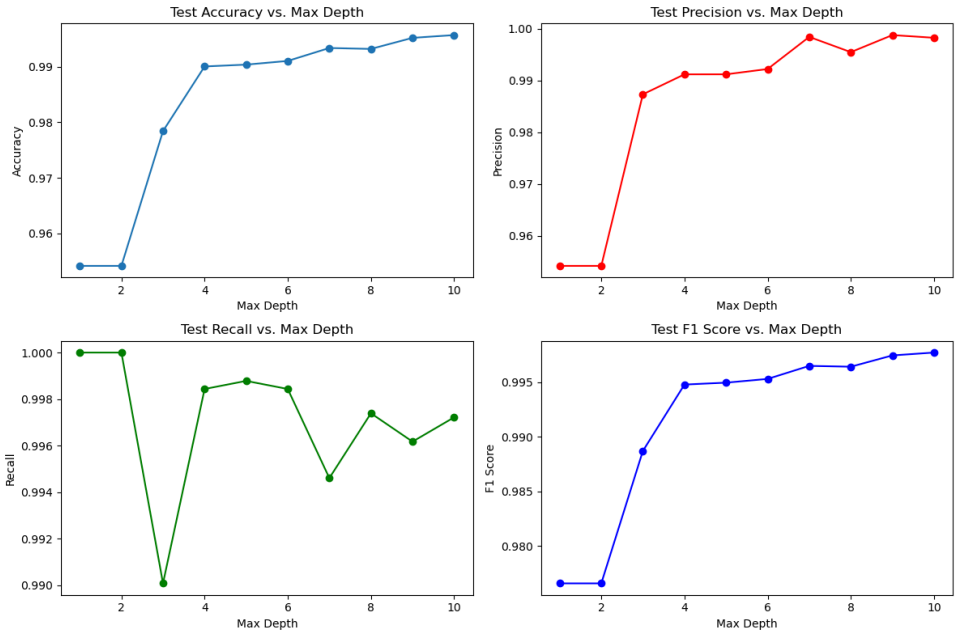

Figure 3 - Four plots each depicting how a different performance metric varies as the decision tree increases in depth. The decision tree in these plots are based on data that includes the year the datapoint was published.

There are a few interesting things to consider with these plots. First, all of the metrics are very high even with only a depth of 1. Next, we may notice that there is a large jump in the metrics at a depth of 3-4. Specifically, looking at the accuracy tells us that the decision tree has done an excellent job at making predictions on data even when we don't spend a lot of computation time on the project.

The results shown above are promising but also concerning. It is suspicious that after only one split the decision tree can reach in the high 90's of accuracy. To try and understand why this occurred I looked into what the first split was. The first feature that the data split on was the year. Perhaps this was the result of technological advancements or a new standard set by cosmology committees. To deal with this possibility I removed the date column from my dataset and reconstructed the decision tree.

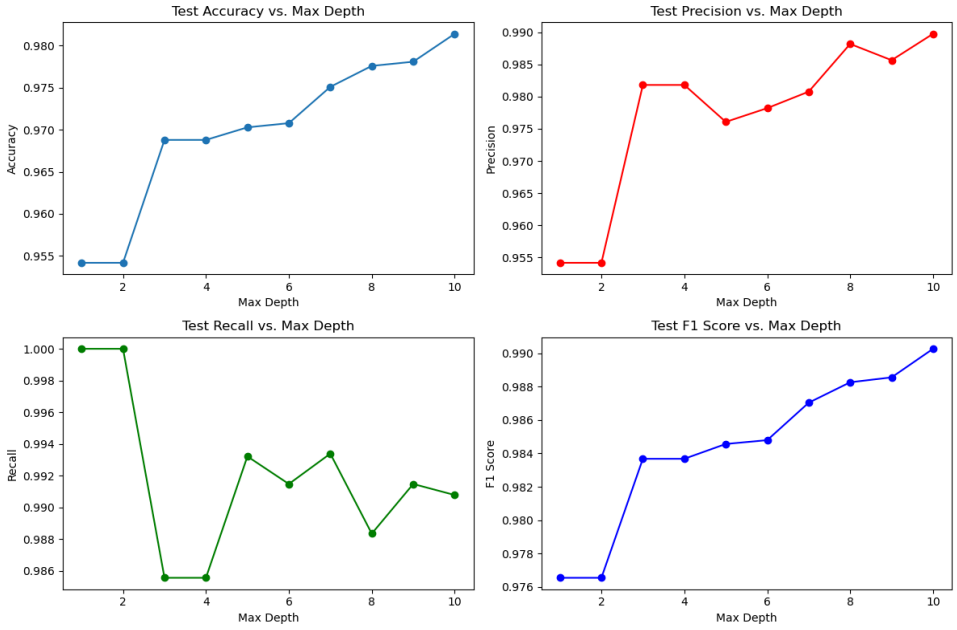

Figure 4 - Four plots each depicting how a different performance metric varies as the decision tree increases in depth. The decision tree in these plots are based on data that does not include the year the datapoint was published.

Once again the decision tree performs exceptionally well on a test data set. In this case the first feature that was split on was redshift which leads me to believe that the decision tree is working as intended.

The performance of the decision tree on each dataset can be evaluated with a confusion matrix. The first dataset which includes the date, by year, yielded 5,998 data points on the diagonal and only 26 data points that are off the diagonal. Of the correctly predicted points 266 are true and 5,732 are false. Where true means the point predicts a value for the Hubble Constant.

It can definitely be seen in the previous plots that the date does have an impact on the accuracy of the decision tree. After removing the date column the number of values on the diagonal drop to 5,912. This accuracy is still very strong; it shows that the information can still be communicated without the date. Another interesting change in the data is that neither the true nor the false predictions were significantly impacted. In the second dataset there are 49 fewer correct true predictions and 37 fewer correct false predictions.

This decision tree is a unique approach to communicating cosmology measurements and methods to those unfamiliar with the topic. With a high accuracy of predicting whether or not a measurement will be able to produce a value for the Hubble Constant it is relatively simple to show someone in congress how certain improvements to telescopes can relate to measuring the natural constant. One example of this might be that there is a distant galaxy whose redshift is only measurable in the infrared

spectrum of light thus a larger telescope would be needed to make this measurement which would result in a refined value for the Hubble Constant.

One way this project was limited was by the dataset. There is definitely more that goes into making these measurements that can't be broken down into one column as this dataset does. We could have a more well-rounded understanding if the dataset also included columns like telescope diameter, wavelength observed, and whether the telescope was ground or space based.

One of the most important skills that I was able to practice on this project was cleaning data. There was a lot that went into understanding what each feature was and why it would or would not be relevant. Additionally, I had practice modifying the data to go from numerical to binary. I feel that this aspect of the project although not difficult is something I will continue to utilize, not this exact process but looking at data as a whole as true or false instead of continuous.

My background is not strictly programming so this course, not only this project, reinforced my ability to take mathematical concepts and apply them to code. This is a skill I have used on multiple personal projects so being even more confident in my abilities only opens new opportunities for me to put my skills into practice.

Citations

ChatGPT (2025) Decision Tree for Hubble. Available at: https://chatgpt.com/share/6816d11c-ce24-8013-b57d-b6a4eec22488 (Accessed: 12 May 2025).

Özeren, E. (2024) Building a decision tree from scratch with python, Medium. Available at: https://medium.com/@enozeren/building-a-decision-tree-from-scratch-324b9a5ed836 (Accessed: 12 May 2025).

Steer, I. and Madore, B.F. (2020) NED-D: A Master List of Redshift-Independent Extragalactic Distances, Ned Redshift-independent distances (ned-D). Available at: https://ned.ipac.caltech.edu/Library/Distances/ (Accessed: 12 May 2025).

Wikipedia (2025) Budget of NASA, Wikipedia. Available at: https://en.wikipedia.org/wiki/Budget_of_NASA (Accessed: 12 May 2025).