* [Overview](#gjdgxs)
* [Preparation Reading](#30j0zll)
* [Extra Help](#1fob9te)

# CSE 450 - Machine Learning & Data Mining

## Module 03 — Decision Trees

 [*Photo by veeterzy on Unsplash*](https://unsplash.com/photos/sMQiL_2v4vs)

## Overview

Estimated Reading Time

Plan on around 90 - 120 minutes for this preparation reading, which consists of textbook reading.

There are also some optional online materials you may want to review.

The objective of this module is to provide a real-world scenario in which you can practice the following data science / machine learning skills:

* Decision Trees and the ID3 Algorithm
* Testing a machine learning model
* Evaluating how well a model makes categorical predictions

## Preparation Reading

Reading the Text

As you read the chapters in the textbook, you may find it useful to start with the chapter summary at the end of the chapter.

This summary provides an overview of the key concepts presented in the chapter, and shows how each is connected together.

Sensitivity and Specificity

In section 8.4.2.1 about confusion matrix-based performance measures, you'll see the terms "True Positive Rate" and "True Negative Rate" (among others).

It is common to refer to "True Positive Rate" as *Sensitivity*.

It is also common to refer to "True Negative Rate" as *Specificity*.

The text mentions this in a footnote on page 414, but they are such common terms, we want to make sure you don't miss this definition.

Complete the following preparation reading:

* Read pages 117 - 144 of your text (Chapter 4 until section 4.4), which describes Decision Trees and the ID3 algorithm.
* Read Pages 397 - 413 of your text which describes ways to verify how well a machine learning model works. (Chapter 8 until section 8.4.2)
* Read the following sections from chapter 8 which each describe a specific model performance measure:
  + 8.4.2.1 - Confusion Matrix-based Performance Measures
  + 8.4.2.2 - Precision, Recall, and $F\_1$ Measure.
  + 8.4.3 - Prediction SCores
  + 8.4.3.1 - Receiver Operating Characteristic Curves (ROC Curves)

## Extra Help

Below you'll find some optional videos and other resources that help supplement the reading.

You should absolutely still do the reading above. One technique would be to read the text, paying particular attention to new concepts (usually written in bold), then research those concepts using videos or other articles until you're confident you understand them. Afterwards, circle back to the text to pick up extra details you might have missed the first time.

Learning Complex Technical Information

Reading technical information can be difficult and is an acquired skill that you absolutely should develop if you're planning to work in data science. New research papers and algorithms are released constantly in this field that require you to parse through information and formulas.

This helps you to not only understand how the algorithm works, but which types of problems the algorithm would and would not not be suited for.

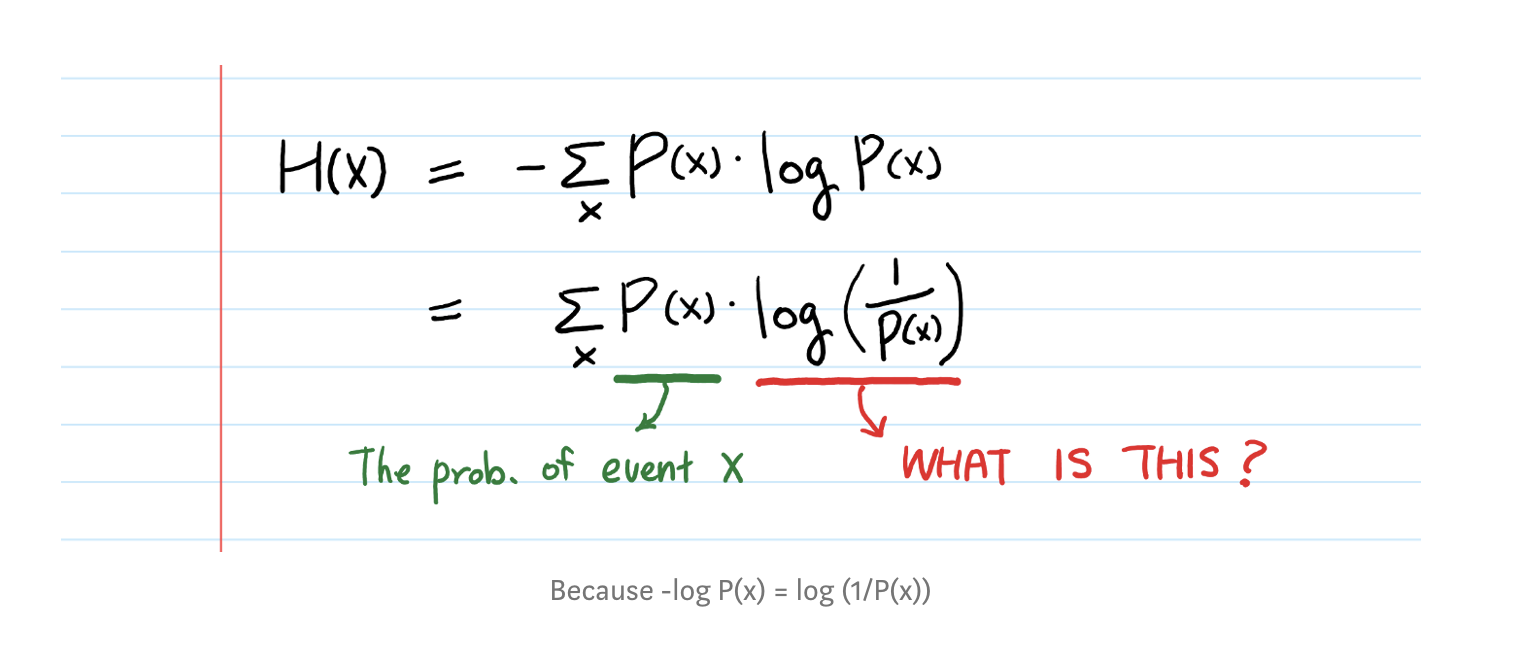
However, sometimes it's nice to have a different perspective. Some people learn better visually, through videos, interactively, or by example.

In some cases, a superficial understanding of an algorithm and its parameters may be good enough for what you need to do. But you'll always benefit from a deeper understanding of how the tools and algorithms you're using actually work, and the reasons they behave better in some situations than others.

#### Decision Trees

* [Here's a cool interactive demo](http://www.r2d3.us/visual-intro-to-machine-learning-part-1/) of decision trees.
* [This follow-up demo](http://www.r2d3.us/visual-intro-to-machine-learning-part-2/) shows the bias-variance trade-off in decision tree models.

#### Shannon's Entropy Model

 [*Image from The Intuition behind Shannon’s Entropy*](https://towardsdatascience.com/the-intuition-behind-shannons-entropy-e74820fe9800)

* [This video](https://www.youtube.com/watch?v=2s3aJfRr9gE) has a nice visual way of explaining the relationship between entropy, questions, and information.
* [The Intuition behind Shannon’s Entropy](https://towardsdatascience.com/the-intuition-behind-shannons-entropy-e74820fe9800) has a nice discussion of the reason behind the log transform part of the entropy formula (equation 4.1 in the text).
* [Why take the log of a continuous target variable](https://towardsdatascience.com/why-take-the-log-of-a-continuous-target-variable-1ca0069ee935) has another nice explanation of why we often care about the log transform of a feature, something that's going to come up quite often.

#### Model Evaluation

* [About Train, Validation and Test Sets in Machine Learning](https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7) is an article that explains why we split the dataset up to evaluate our machine learning models.
* [Accuracy, Precision, Recall or F1?](https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9) explains a bit about each of these performance metrics, and when it is best to use each one.
* [ROC and AUC, Clearly Explained!](https://www.youtube.com/watch?v=4jRBRDbJemM) is a video visually describing how to build and use an ROC curve. (Don't let the corny intro music dissuade you, it's a pretty good video)