







Bookmarks



Bookmark

- ▶ Start Here
- ▶ 1. The Big Picture
- ▶ 2. Data And Features
- ▶ 3. Exploring Data
- ▶ 4. Transforming Data
- ▶ 5. Data Modeling
- ▼ **6. Data Modeling II**
 - Lecture: SVC**
Quiz 
 - Lab: SVC**
Lab 
 - Lecture: Decision Trees**
Quiz 
 - Lab: Decision Trees**
Lab 



6. Data Modeling II > Lecture: Decision Trees > Video

How Do Decision Trees Work?

MSXPPDSX2016-V004100

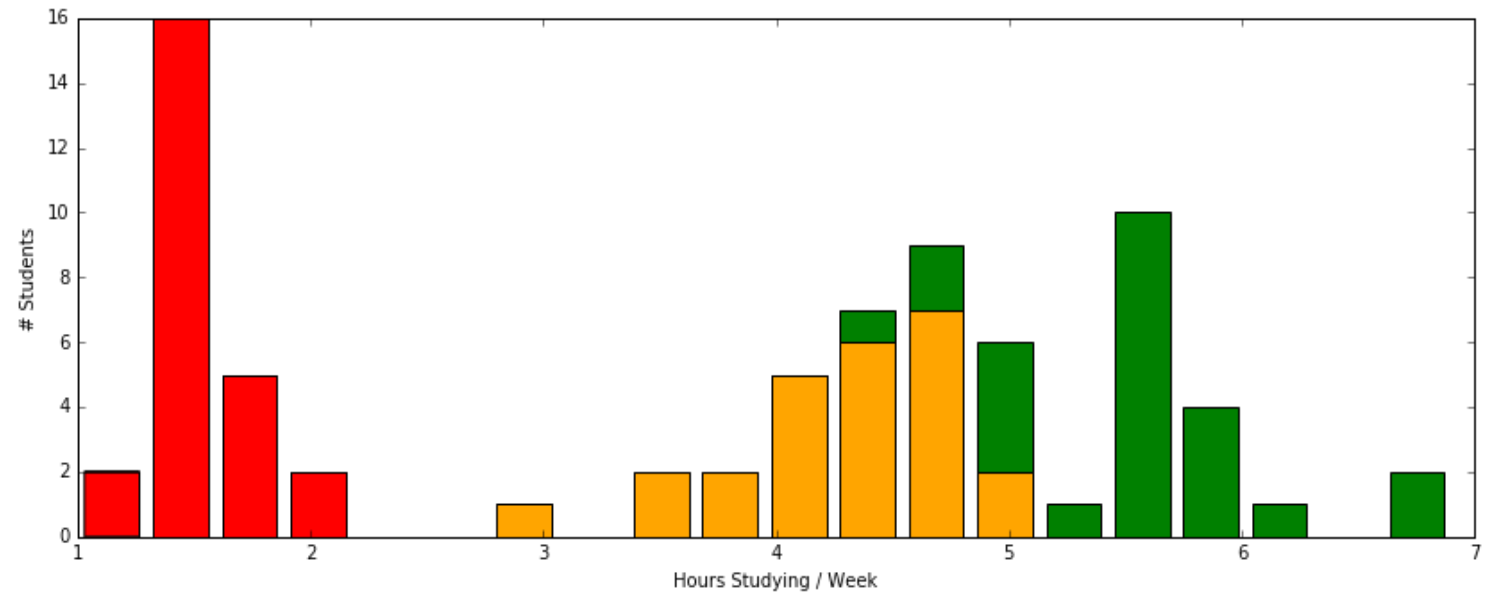


Dive Deeper

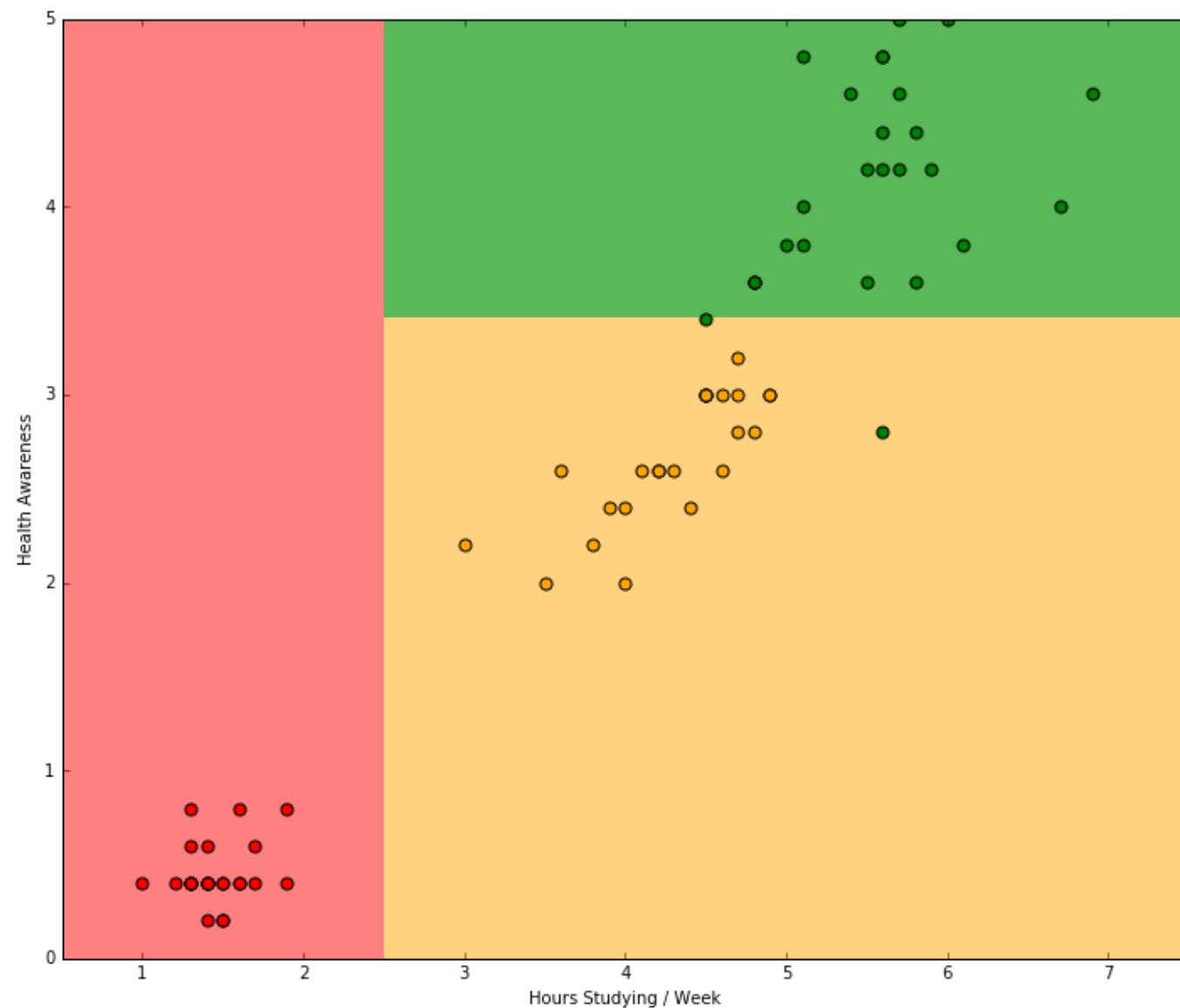
 0:00 / 3:29 1.0x

Let's start with a quick review of all the machine learning algorithms you know. PCA and Isomap work by intelligently simplify your data by either variability or intrinsic geometry. K-Means clusters your data based on feature similarity. K-Neighbors classifies your data by feature similarity. Linear regression models a continuous, linear correlation in your data. And now decision trees classify your data probabilistically based off of entropy, which can be thought of as 'purity' or information gain.

Each decision tree models sequential consequences based off of their chance of occurrence, their resource cost, and their information gain, with the goal of maximizing the overall purity of class the samples stored in each leaf node. To illustrate this, consider the following example. A teacher feels that there must be some relationship between how healthy his students eat, how many hours they spend studying, and their final grades. To discern the correlation, he asks 75 of his students how many hours a week they spend studying. It seems reasonable to start with this question, as it should have the most impact on their grades.

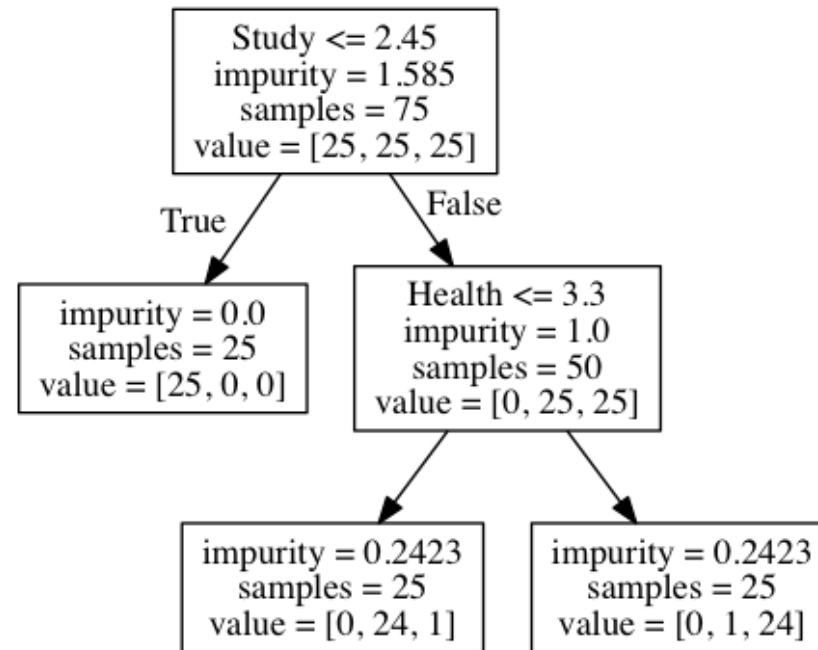


Fascinating. The teacher notices that 100% of the failing students spend less than 2.5 hours per week studying. Perhaps equally interesting is that of the students passing the course, some of them study a *lot*, yet others study just at the average level. To properly classify these students, more data is certainly needed. Being curious, the teacher asks the students another question: how health conscious are they with regards to their food, on a scale from 1-5. Would they prefer pepperoni pizza (1) or reach an apple (5)?



Various answers come back, and something catches the teacher's eye. With only a *single* exception, every single students who doesn't take their health and eating habits seriously, by responding with less than 3.3 on the 5-scale survey, has lower ranking grades. Just as the teacher keeps asking questions to

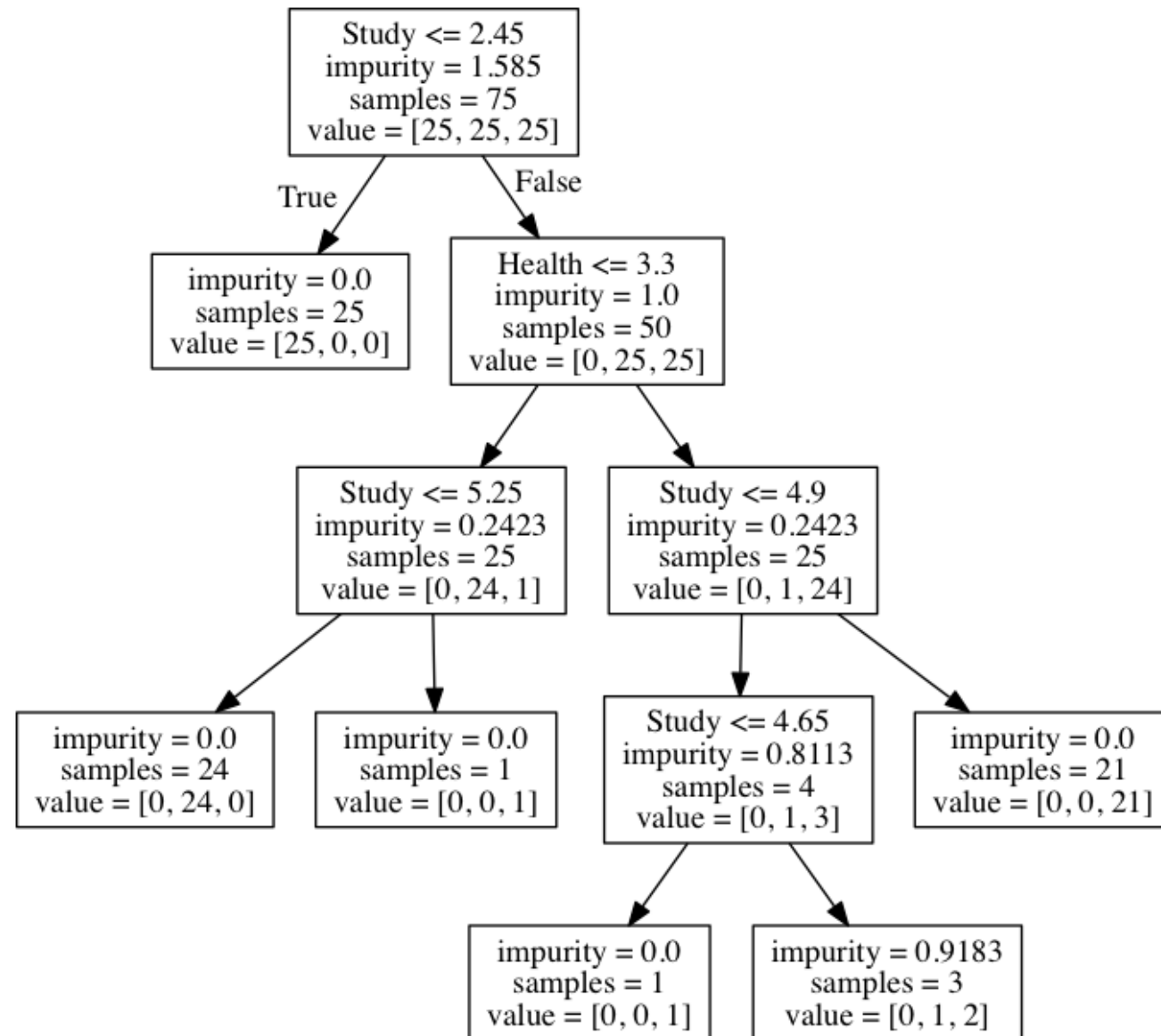
find out more details about his students so he can properly classify them, so to do decision trees like cleanly chopping away at your feature space to **purify the class** of your samples, based off of linear or straight decision boundaries. This is what a decision tree would look like for the above example:



The root node starts with all samples, intermezzo. Value = [25, 25, 25] represents all the teacher's students who belong to three classes: slackers, status-quo, and over-achievers, in that order. At the first node, there is a high level of impurity. Actual calculations for impurity and entropy have been included in the Dive Deeper section. The feature being considered by the root node is the number of hours spent studying per week, and decision being made is ≤ 2.45 or greater. In this tree, each left branch is a true response to the decision question, and each right branch is a false response.

Those students that spend ≤ 2.45 hours per week all belong to the first class, value = [25, 0, 0] and so that branch has an impurity of 0.0. Students that spend more than 2.45 hours per week studying consists of the rest of the student body. Exactly half of the students belong to one class, and exactly half belong to another class. In other words, this branch is as thoroughly mixed up as a binary classification can possibly be. This is why impurity is set at 100% here. But by testing the health awareness feature, the teacher is able to considerably *purify* the resulting selections, such that only a single student sample is incorrectly classified on either side.

If the teacher knows what's good for him, he'll probably stop at this point. But if he let's curiosity get the better of him, he might continue asking the students more questions, or alternatively, dig deeper into the data he already has in order to further derive means of classifying his students based on his existing features. This would be a good example of overfitting data. Doing so leads to very intricate and complicated trees:





© edX Inc. All rights reserved except where noted. EdX, Open edX and the edX and Open EdX logos are registered trademarks or trademarks of edX Inc.

