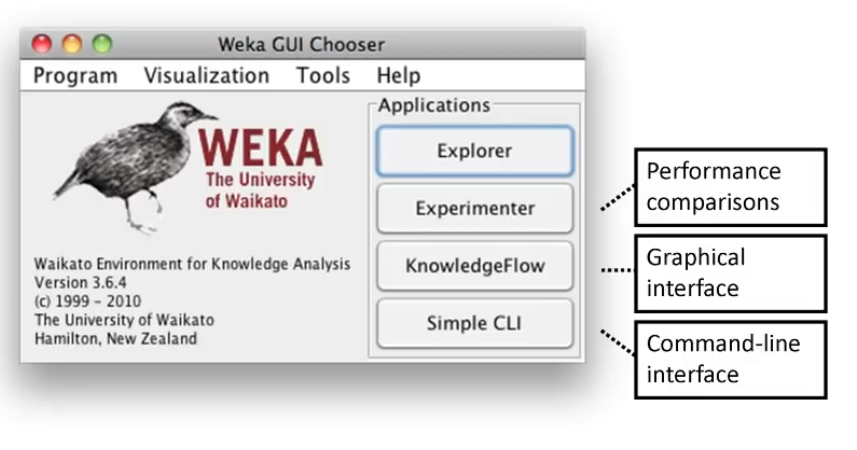
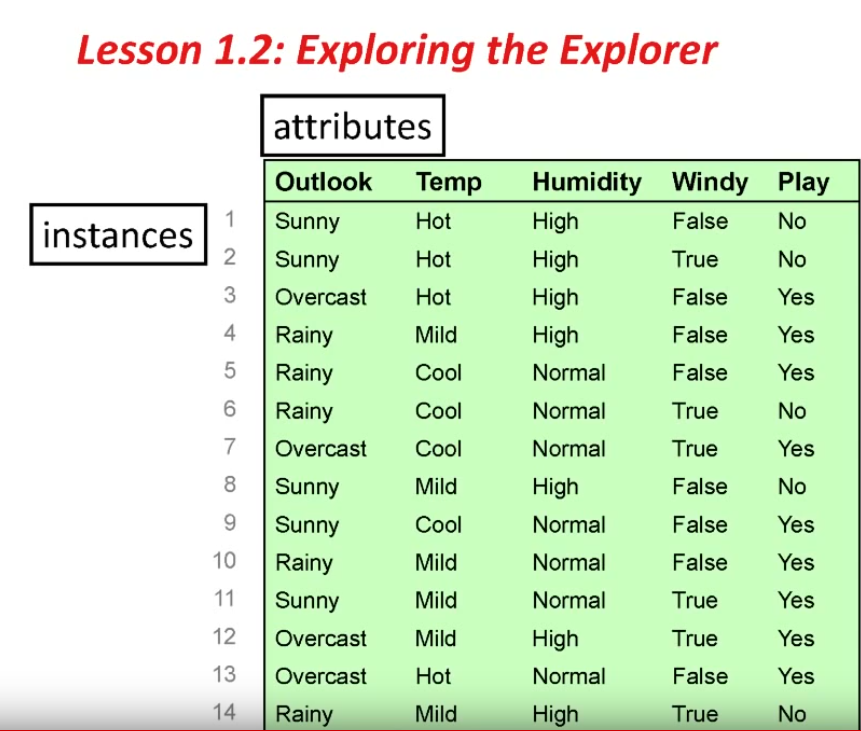
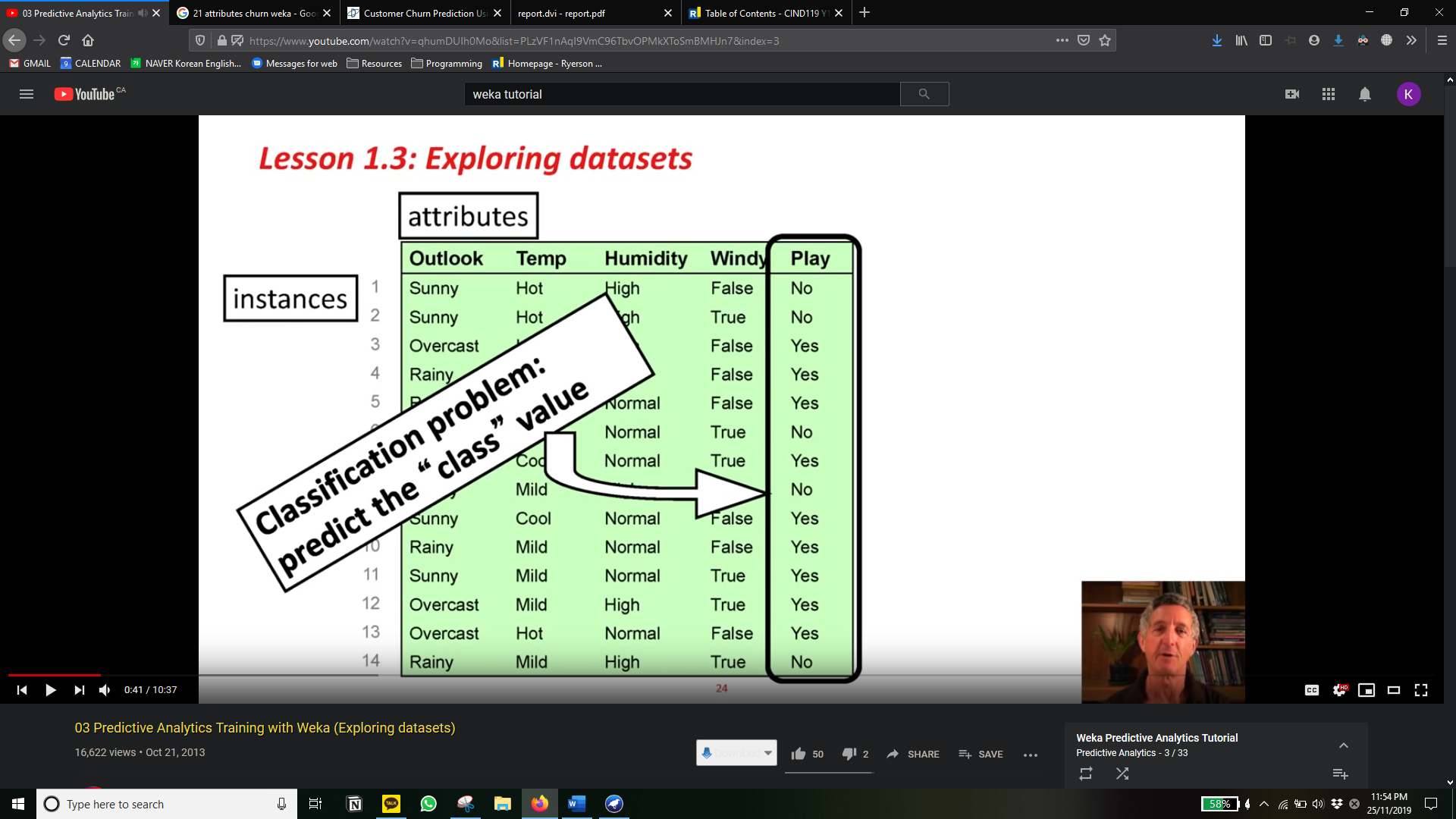
WEKA: Waikato Environment for Knowledge Analysis

developed at the University of Waikato, New Zealand

Tutorial created by Kyuhwan Kim







**J48 is a tree classifier.**

**Classify -> trees -> J48**

**pruned**

**Try: pruned vs unpruned (91.95%)**

**What’s the difference between pruned and unpruned?**

**minNumObj (larger # to have larger leaves)**

**(5.0/1.0) 5 correct and 1 incorrect instances got to the leaf**

**<To visualize tree>: right click -> visualize**

**To fit, enlarge, right click -> fit to screen**

**<Filters>**

**To remove an attribute**

**Preprocess -> Unsupervised (supervised is for class: rare) ->**

**attribute->Remove**

**Click remove. -> attribute indices -> type the attributes to remove -> OK -> apply (right hand side)**

**EASY WAY: select an attribute and click remove button right below**

**To remove an instance**

\*this will create a reduced dataset

Preprocess -> unsupervised -> instance-> removewithvalues (click it)

Click more for more details

Attribute index: which attribute

Nominal index: which instance (put #)

<Visualization>

Classify-> visualize classifier errors

Result: class vs predicted class

\*square boxes represent errors

**<add classification>**

**Preprocess->Choose->classifier to J48 -> outputclassification(TRUE)**

**Naïve Bayes**

[**https://www.youtube.com/watch?v=PRatbc8lOU8&list=PLzVF1nAqI9VmC96TbvOPMkXToSmBMHJn7&index=7**](https://www.youtube.com/watch?v=PRatbc8lOU8&list=PLzVF1nAqI9VmC96TbvOPMkXToSmBMHJn7&index=7)

**7:46**

**Classifier-> bayes-> NaiveBayes-> Start**

**Other: supplied test set -> set**

**++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

**Class 2:**

**Classify -> Choose…-> Tree-> User Classifier**

**(we will not likely use user (self) classifier)**

**\*really important that training and test data are different**

**\*for our project, use % split (make a random split on the dataset)**

**\*always guess negative: ZeroR classifier (baseline accuracy)**

**To change random seed:**

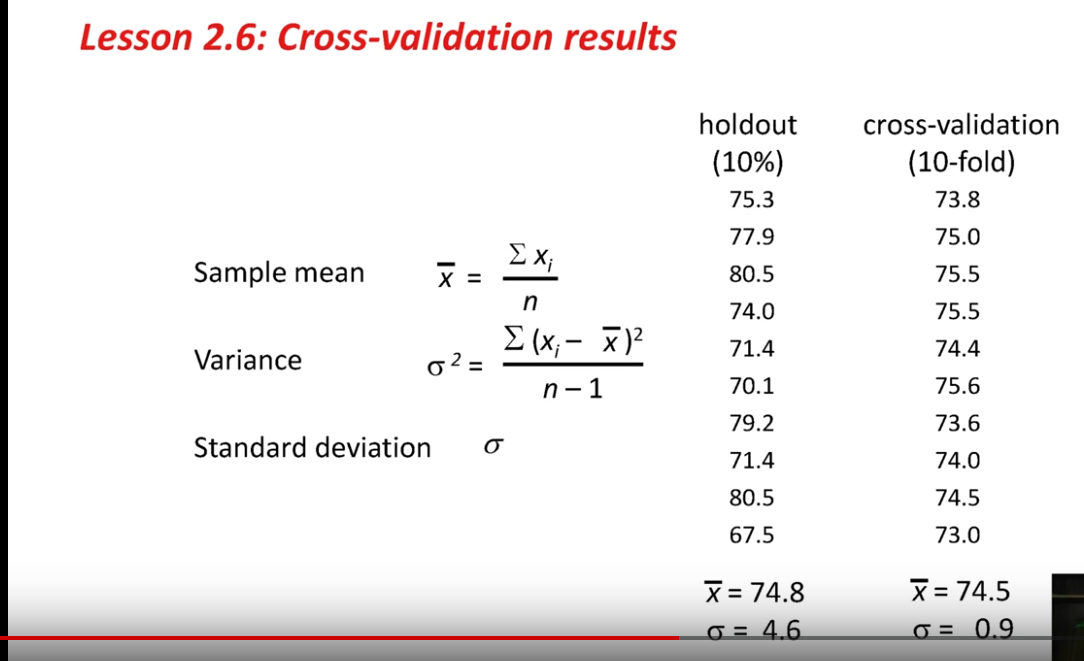
**Classify -> More Options -> change # for random seed**

**Cross Validation**

**Repeated holdout, holdout 10% repeat 10 times**

**Cross validate, 10 different pieces, use 9 for training, repeat until all done. Then average the 10 results**

**Stratified cross validation: same except that each fold has right proportion of each class value (helps reduce variance of the estimate)**



**After cross validation, 11소 time use 100% of datasets**

**Practical rule of thumb. If you have lots of data, you can use percentage split.**

**Class 3:**

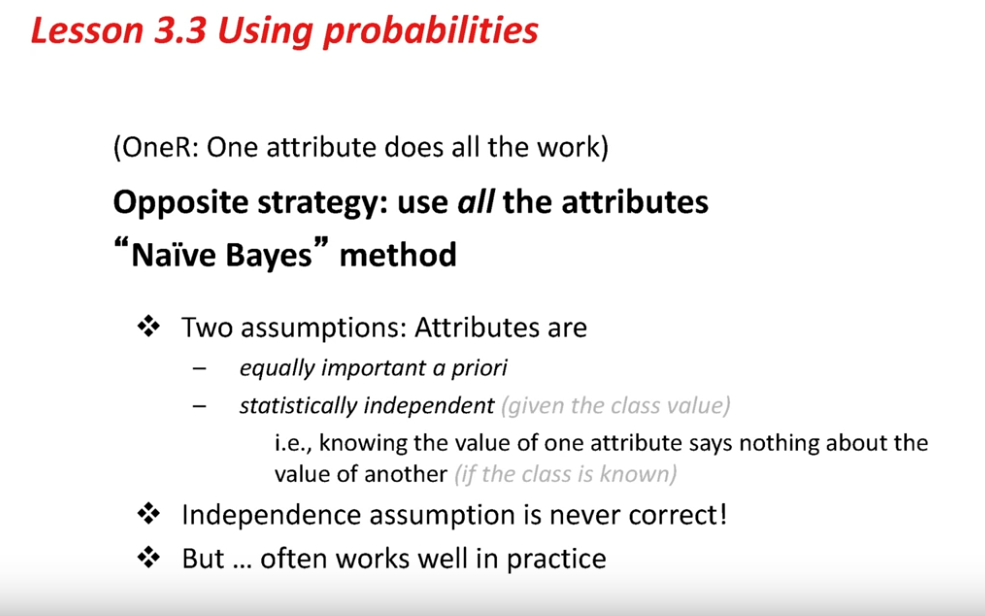
OneR: one attribute does all the work, chooses attribute with the smallest error rate

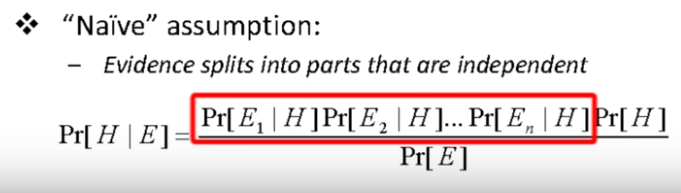
**Model Overfitting**

Any machine learning method may “overfit” the training data. By producing a classifier that fits the training data too tightly. Therefore, works well on training data but not on independent TEST data.

**Naïve Bayes:**

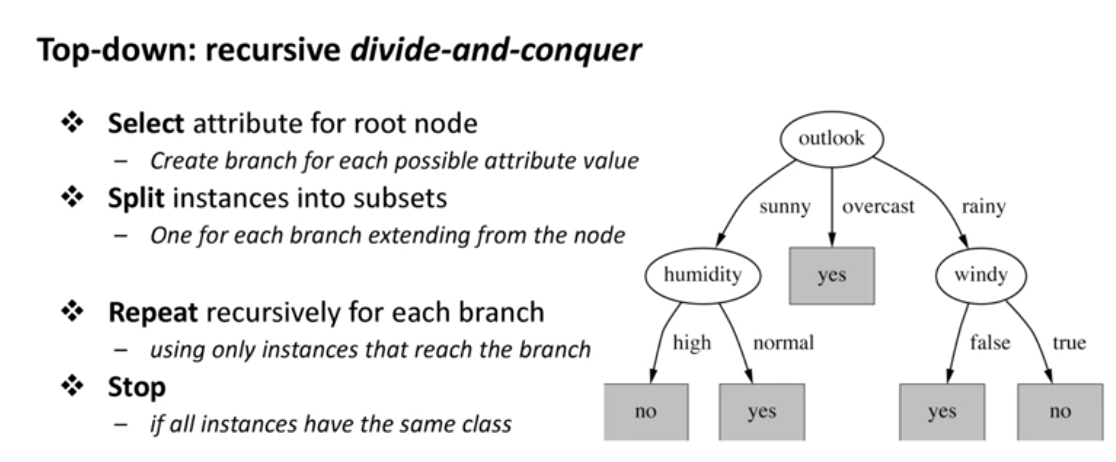
Opposite strategy to OneR. Use all the attributes equally and independently to the decision.





\*WEKA adds 1 to all the counts. Reason is to get rid of the zeros (sometimes the count can have zero counts; zero multiplications ruin the calculations)

**Decision Trees: J48**



**Root node is the attribute which splits yes/no the best (info. Gain)**

Aim is to get the smallest tree using information tree

Weather (3,0) -> 3 correctly and 0 incorrect/other

**Pruning Decision Trees**

J48 prunes by default

Unpruned results in bigger tree and worse prediction result