Reichman University

**Location:** Address or Room Number
**Date:** Date
**Time:** Time

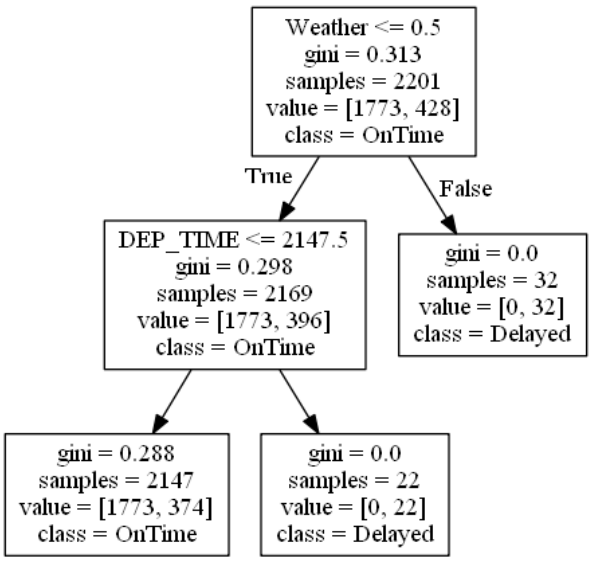
# CLASSIFICATION with DECISION TREE

## Dr. Gail Gilboa Freedman

In the forthcoming sessions, you will discover how to **solve the problem of CLASSIFICATION for data on FLIGHTS' DELAY with DECISION TREE algorithm**

**Python material:** 'DS Decision Tree.pynb'; 'DS Decision Tree flights data prepared.csv'

**Key terms:** supervised learning; the classification problem; regression; spam detection; spam detection; handwriting recognition; training testing data; accuracy; confusion matrix; sensitivity; specificity; ROC curve; The flight delayed problem; Excel Advanced.Unique; Excel match-function; DecisionTreeClassifier; numpy.ndarray; Graphviz; sklearn.tree.export_graphviz




In the upcoming sessions, we focus on **Supervised learning models** (tuned against labeled data).

Supervised problems include **classification**, **regression**

Specifically, we will focus on the problem of classification.

Business applications of classification problems: spam detection, handwriting recognition

Reichman University

**Location:** Address or Room Number
**Date:** Date
**Time:** Time

## I. Problem: CLASSIFICATION

a) **Definition**

**The classification problem** is to produce a grouping of data points with the purpose of predicting the label.

b) Evaluation measures:

Why is it important to evaluate the classification model? To compare different models; to select the right features; to decide whether to use a model or not.

Desired properties (of the classification model):

1. High True Positive
2. High True Negative

### Evaluation Measures

**The overall classification Accuracy (Error) rate** is the proportion of samples that have been correctly (incorrectly) classified.

$$Accuracy = \frac{actual == predicted}{total}$$

**Confusion matrix** is a table that demonstrates the number of correct and incorrect predictions categorized by type of outcome.

| | **Actual = Negative** | **Actual = Positive** |
|---|---|---|
| **prediction=NO** | True Negative:<br>Labeling Negative as No | False Negative<br>Labeling Positive as No |
| **prediction=YES** | False Positive<br>Labeling Negative as YES | True Positive<br>Labeling Positive as YES |

Reichman University

**Location:** Address or Room Number
**Date:** Date
**Time:** Time

**Sensitivity (Se) or True-Positive-Rate** is the proportion of identified positives among all the positives.

$$Sensitivity = \frac{true\ positive}{true\ positive + false\ negative}$$

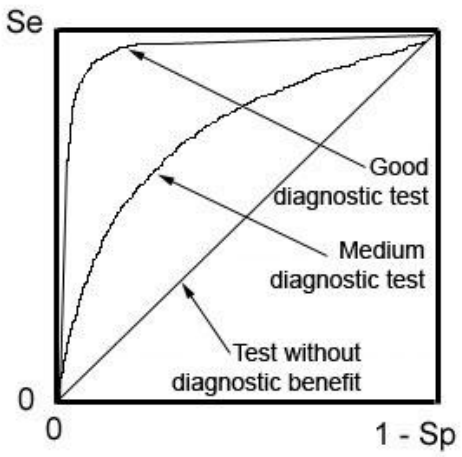
**Specificity (Sp) True-Negative-Rate** is the proportion of identified negatives among all the negatives.

**Receiver Operating Characteristic (ROC) curve**

If you have a parameter that influences the performance of the model, you aim at:

- low/high true-positive-rate
- low/high false-positive-rate

For a variety of parameter values, compute TPR & FPR to generate the curve.




c) Algorithms (that solve the classification problem)

decision tree, naive Bayes, support vector machine (SVM), least-squares

d) Algorithm of the hour: **Decision Tree**

In the forthcoming lessons, we will get familiar with the Decision Tree. It is an algorithm that is an algorithm that solves the classification problem, predicting or classifying future observations based on a set of decision rules.

Reichman University

**Location:** Address or Room Number
**Date:** Date
**Time:** Time

## XV. Data Y

The data in 1 file:
***'DS Decision Tree flights data.csv'***

Each row represents a flightand characterizes by 8 features (columns):

**'DAY_OF_WEEK', 'DEP_TIME', 'Weather', 'CARRIER_i', 'DEST_i', 'ORIGIN_i', 'Delayed', 'Train'**

**The flight delayed problem** is to produce a grouping of data points with the purpose of predicting whether a flight will be delayed

## XVI. Tools:

**Excel: for data preparation**

We have column of values, that we need to transform into categorical columns of indices.

E.g. for the strings column 'blue' 'blue' 'yellow' 'red', we will have the integers column '1' '1' '2' '3'

**How?**

**Step by Step:**

(1) Select column

**(2)** Filter for unique values, by Data > Sort & Filter > Advanced.**Unique**

(3) Insert a new column

**(4)** Use **match-function**

Reichman University

**Location:** Address or Room Number
**Date:** Date
**Time:** Time

**Practice:** Solving X for Y with Z

*'DS Decision Tree.ipynb'*

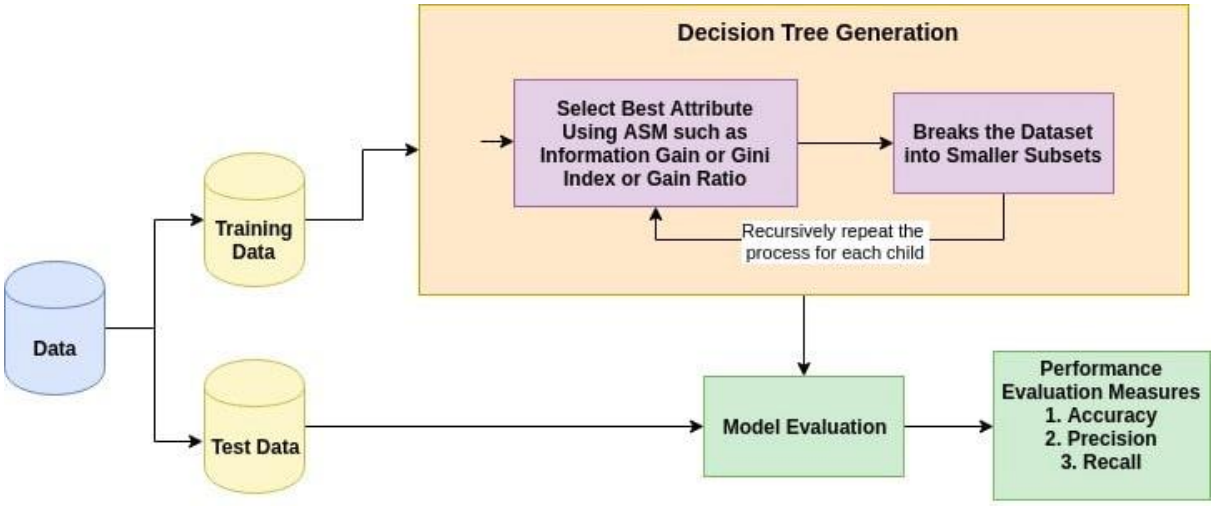
b) Install and Import

c) EDA

d) Building and Evaluation of the model

**Python: for running the Decision Tree algorithm**

**How does decision-tree-classification work in Python?**




**Step by Step:**

(1) Saving data in numpy.ndarray to represent a multidimensional array of fixed-size items

(2) Split array to into train and test

(3) Predictors and label separations

(4) Build model with training data

(5) Make predictions for testing data

(6) Evaluate model)e.g. by accuracy)

**Python: for running a visualization of the Decision Tree model**

with Graphviz, visualization software, and sklearn.tree.export_graphviz

## XVII. Summary and Discussion

We learned theory and practice of how to build a classification model that classify objects into the same class when they are similar in terms of their tendency to have the same label. Specifically, we solved the delayed flight problem of classifying flights by similarity in terms of their characteristics. We utilized Python libraries to implement Decision Tree and Accuracy for modeling and evaluating our model, respectively. The data process described is an example of common supervised data analysis.

## XVIII. I AM VERY CURIOUS!

Machine learning-based survival rate prediction of Korean hepatocellular carcinoma patients using multi-center data