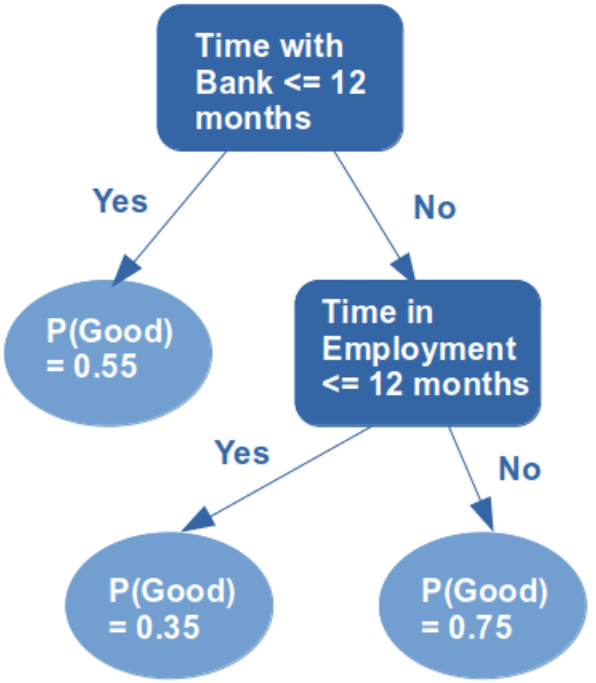
# Machine Learning:

# Decision Tree Classifier

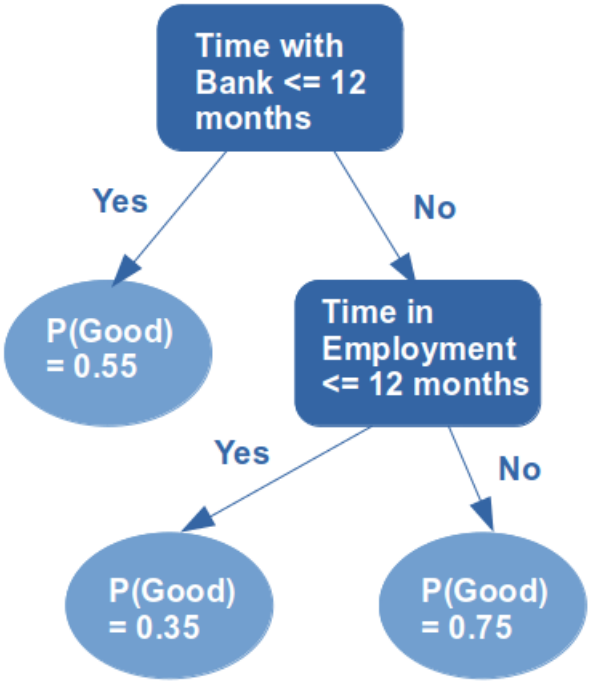
# 1. What is a Decision Tree Classifier?

- Decision Tree Classifier is a Machine Learning algorithm suitable for classification problems.
- A Classification problem is a problem where the aim is to predict a category.
- The output of a Decision Tree classifier returns the probability that an element belongs to a target category
- In Example: Predicitng if an applicant is Good or Bad in Credit Risk Modelling



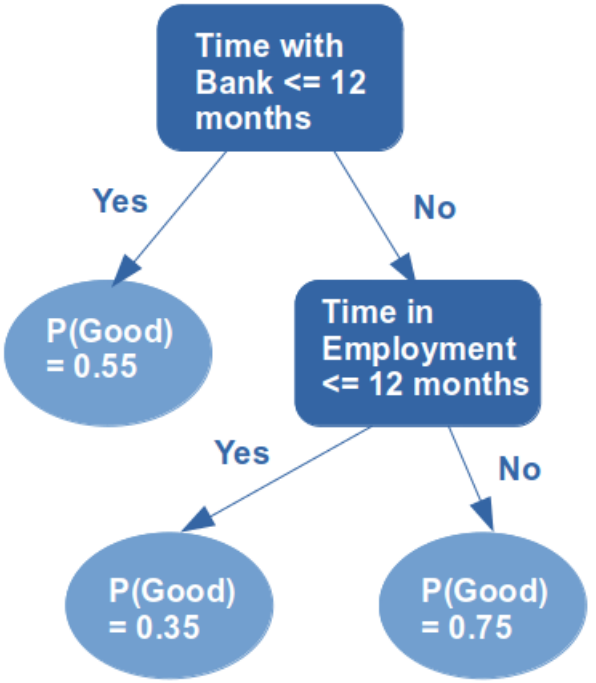
# 1. What is a Decision Tree Classifier?

- A Decision Tree Classifier consists of 3 components:
- Root Node
- Internal Nodes (Branches)
- Leaf Nodes



# 2. How is a Decision Tree Classifier created?

- Decision Trees segment the data by using predictive variables.
- Predictive variables with high predictive power are used early on.
- Predictive power can be measured by **Gini Impurity** or by **Entropy**
- Decision Trees always use the most informative variable splits.



# 2. How is a Decision Tree Classifier created?

- **Gini Impurity – Formula:**

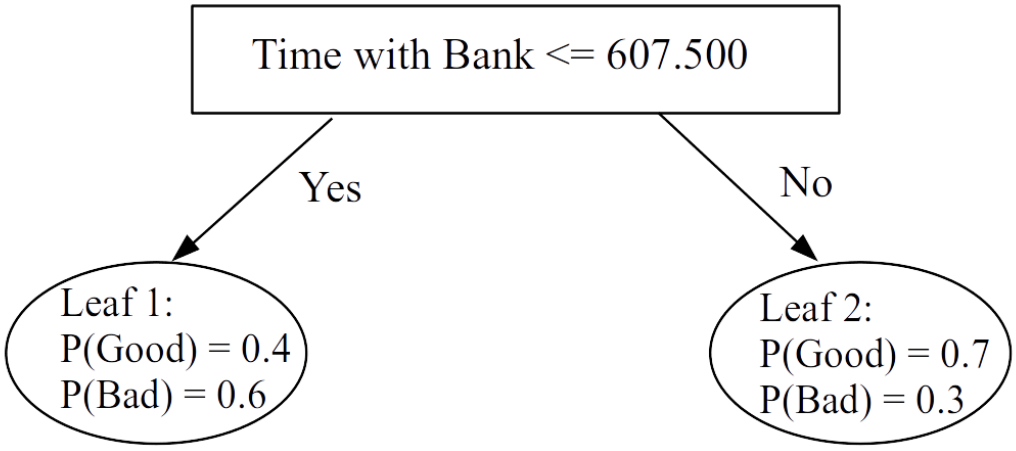
Gini Impurity (Leaf) = $1 - P^2(\text{Good}) – P^2(\text{Bad})$

Gini Impurity (Total) = $\Sigma$ (Weight * Gini Impurity (Leaf)) / Number of Leaves

Weight = Number of elements in Leaf / Total number of elements

# 2. How is a Decision Tree Classifier created?

- **Gini Impurity – Example:**
- Gini Impurity (Leaf 1) = 0.48
- Gini Impurity (Leaf 2) = 0.42




- Assuming we have 100 elements in
- each Leaf:

- Total Gini Impurity = 0.5(0.5*0.48 + 0.5*0.42) = 0,225

- Splits with **lower** values of **Gini Impurity** have **higher predictive power**.

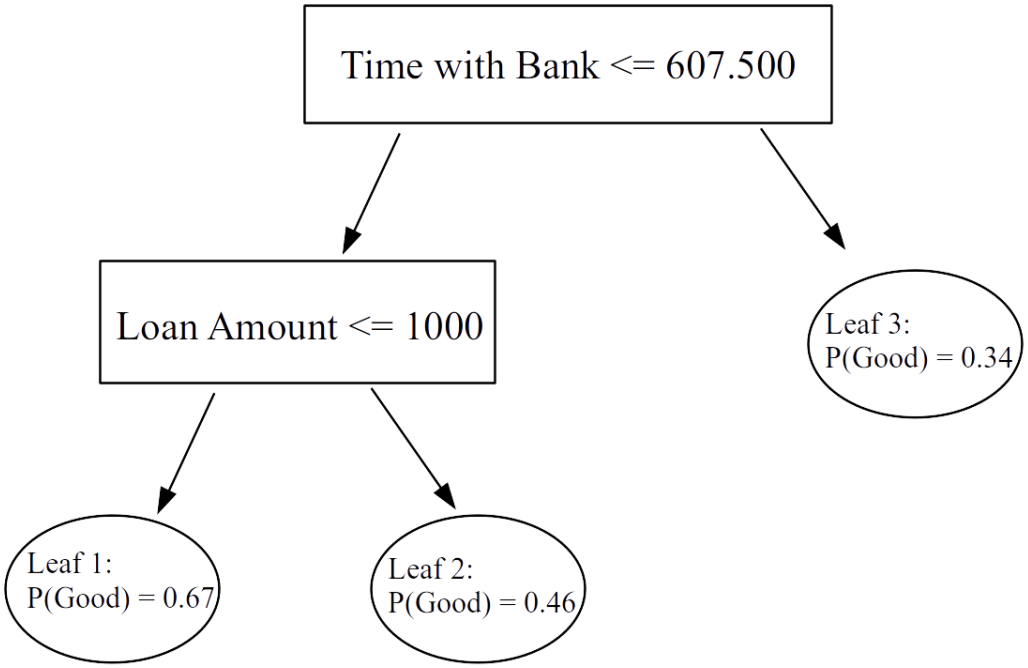
# 2. How is a Decision Tree Classifier created?

- Numerical variables are sorted in ascending order. The average of each pair is calculated and used as a test split. The Gini Impurity is calculated for each possible test split and the split with lowest Gini Impurity is used in the Decision Tree.

| Gross Annual Income | Average Value | Gini Impurity |
|---|---|---|
| 20.000 | | |
| | 22.500 | 0.273 |
| 25.000 | | |
| | 27.500 | 0.302 |
| 30.000 | | |

# 2. How is a Decision Tree Classifier created?

- When calculating further splits, the existing splits are considered as well.
- The Gini Impurity of the Loan Amount split is calculated only on the population that has Time with Bank <= 607.500



# 2. How is a Decision Tree Classifier created?

**How is missing data handled?**

- Missing values in a column could be replaced with the mode value for categorical variables and with the mean / median value for numerical variables

- Alternatively, a highly correlated variable could be used to build a linear regression fit and to predict the missing value

| Unique ID | Gross Annual Income | Time with Bank |
|---|---|---|
| 1 | 25,000 | 24 |
| 2 | 20,000 | 12 |
| 3 | 50,000 | 72 |
| 4 | ? | 48 |

# 3. Hyperparameters

- Hyperparameters are like settings of the model. They are used to define model behavior:
- **Maximum Depth** – how many splits can a Decision Tree make before comming to a prediction
- **Minimum Child Sample** – how many elements at least are required in each leaf
- **Number of Leaves** – how many segments can the Decision Tree split the data into
- **Criterion** – Statistical Metric used to determine the best splits (Gini Impurity or Entropy)

# 4. How is data segmented in a Decision Tree Classifier