**Classification & clustering methods**

**Decision Trees & Random Forests**

**Jonathan Brown**

# 

horizontal line

# Introduction

Clustering is an unsupervised machine learning approach which involves attempting to cluster similar data together without knowing its labels. Conversely, classification is a supervised machine learning approach that attempts to identify what class or category new data should fit in by learning from training data (which contains the class/category labels for all the data points).

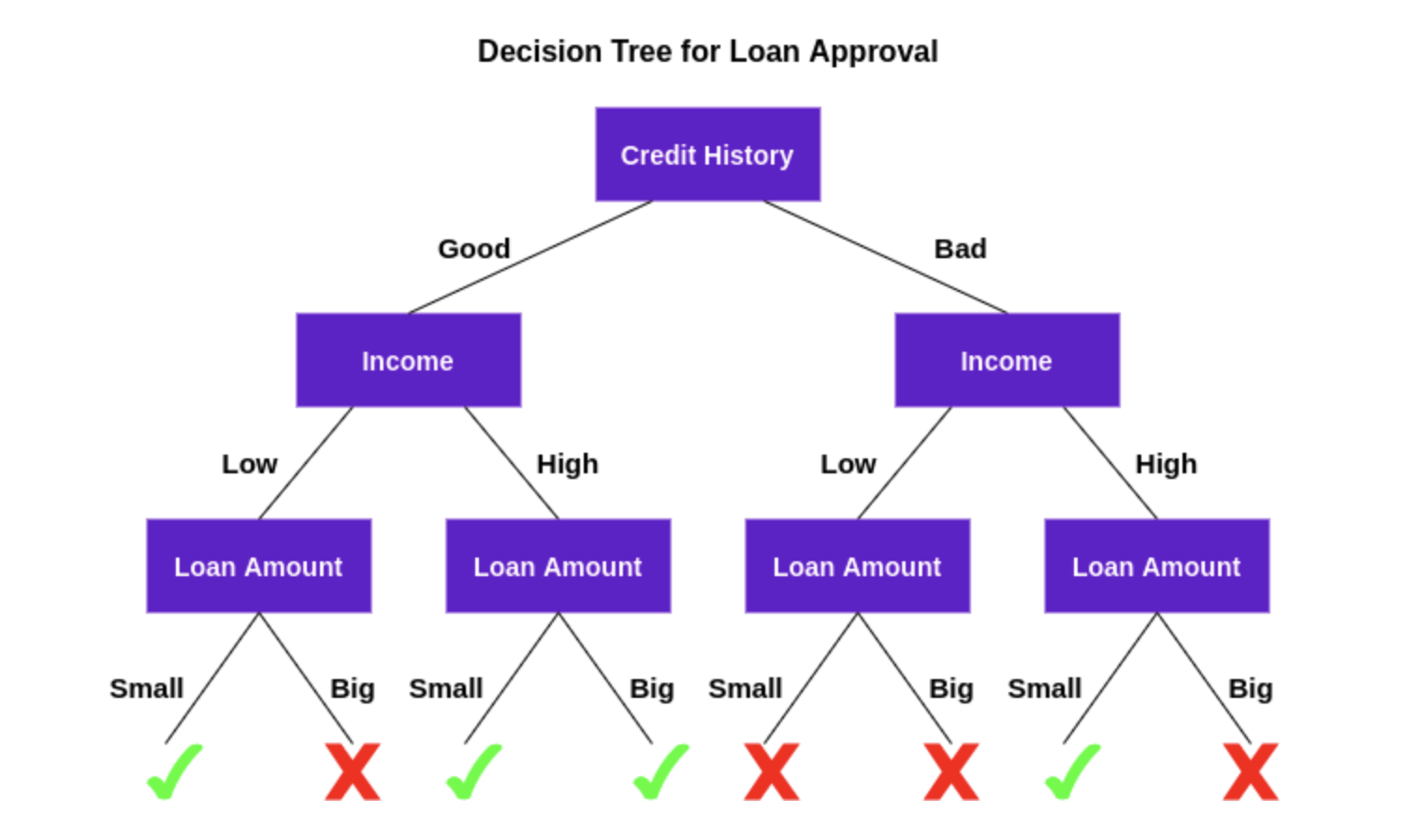
In this brief report I will look into the machine learning model of decision trees and Random Forest. I will provide a definition for these models and a list of strength and weakness, along with an example of where these models might be used.

## 

## Decision trees

A decision tree is a supervised machine learning algorithm that can be used for both classification and regression problems. This type of model asks a series of questions and these sequential decisions are then used to reach a specific result. An example of this type of model could be a bank's decision to provide a customer with a loan.

The model will ask a start question ie. Does the customer in question have a good or bad credit history. This decision will then provide the bank with another decision to be made and so on until the final decision is provided. An example of this is illustrated in the below diagram.



**Pros & cons of Decision Tree**

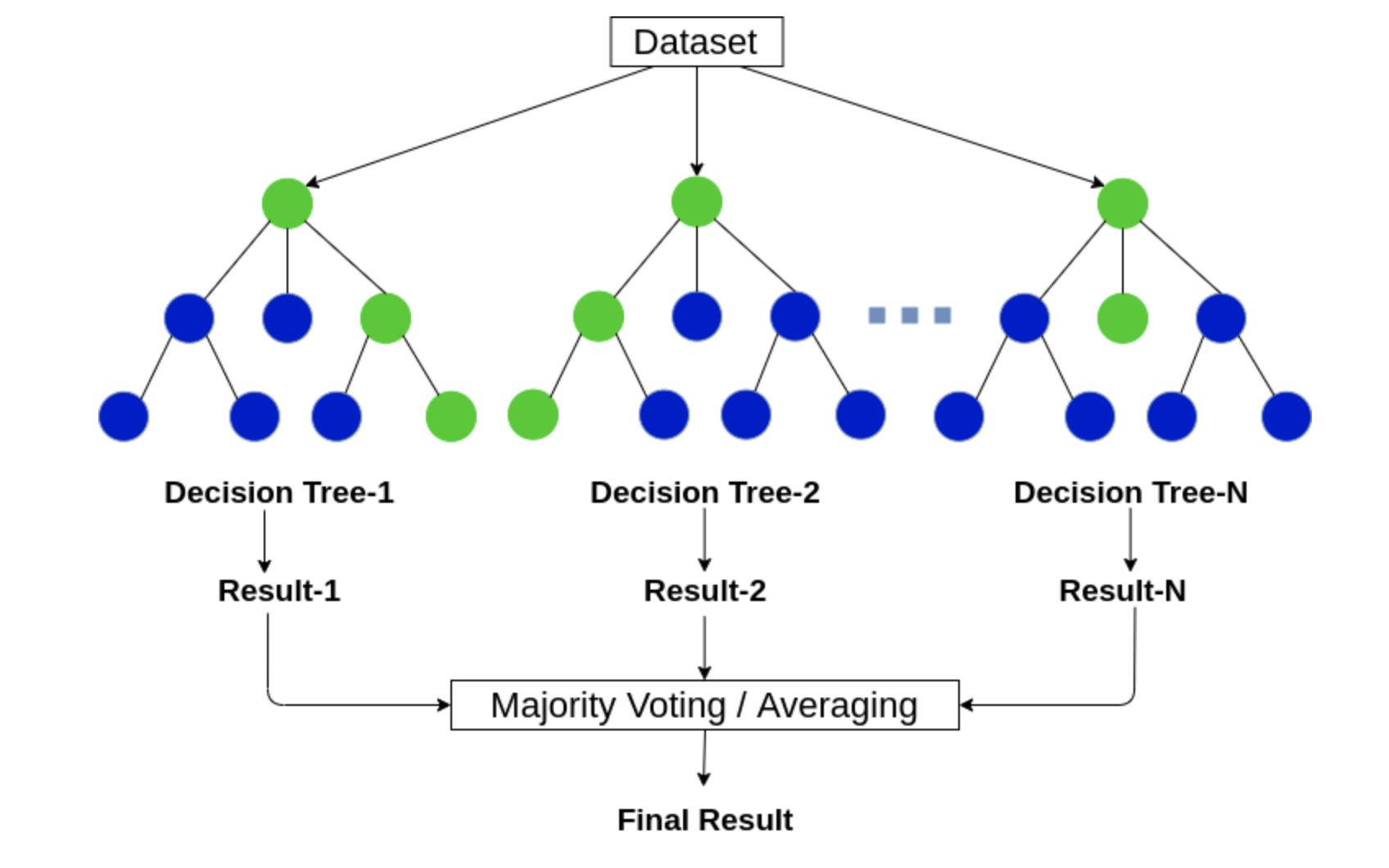
One strength in using a decision tree model is its overall simplicity. This makes it easy for users to follow these models.

On the flip side this simplicity can lead to very simple decisions which can be hard to explain in a working environment. Ie its a yes or no with no explanation.

**Random Forest**

Random Forest is a machine learning algorithm that comes to a final decision by using multiple decision trees in order to come up with the best possible decision.

Hence this is where the name “random” forest comes from. The forest is made up of random decision trees which all work with a random subset of different features to produce and output. This output is then combined to generate a final output



**Pros & cons of Random Forest**

One advantage is that this model can be used to fit large amounts of data.

A disadvantage of this model is that it does not give an accurate prediction when used for regression. Furthermore it can be prone to over-fitting and a too busy model to deal with.