**Chapter 2: Background**

1. **Copper Alloys and Some Properties**

Copper alloys are metal [alloys](https://en.wikipedia.org/wiki/Alloys) that have [copper](https://en.wikipedia.org/wiki/Copper) as their principal component(1), and are useful in a wide range of applications in the electrical industry because of their desirable physical properties such as high conductivity, strength, ductility and resistance to corrosion.(5.intro)

Some important properties of materials that are relevant to this study are -

* 1. **Tensile Strength :**

The tensile strength of a material is a measure of the maximum [stress](https://en.wikipedia.org/wiki/Stress_(mechanics)) that it can withstand while being stretched or pulled before breaking. In technical terms, the tensile strength of a material is the force per unit area at which it breaks in two.  
Tensile strength is measured in units of force per unit area. The unit is newton per square meter (N/m^2), kilogram (force) per square centimeter (kg/cm^2) or kilopounds per square inch (ksi).  
  
The tensile strength of an alloy is dependent not only the composition of the alloy but also the processing methods used during manufacturing of the alloy, among other things.

* 1. **Thermal Conductivity :**

The thermal conductivity refers to the intrinsic ability of a material to transfer or conduct heat. In technical terms, it is defined as the amount of heat per unit time per unit area that can be conducted through a plate of unit thickness of a given material, the faces of the plate differing by one unit of temperature. In the International System of Units (SI), thermal conductivity is measured in watts per meter-kelvin (W/(m⋅K)). In imperial units, thermal conductivity is measured in BTU/(h⋅ft⋅°F).

The thermal conductivity of an alloy, at a given temperature, is fixed for a given composition. However, it is independent of processing methods, unlike tensile strength.

* 1. **Electrical Conductivity :**

The electrical conductivity of a material refers to its ability to conduct electricity.

In technical terms, it is the ratio of the current density in the material to the electric field which causes the flow of current.

According to the Wiedemann–Franz law , at a given temperature the electrical conductivity of a material is directly proportional to its thermal conductivity. Since, there is a direct relationship between thermal conductivity and electrical conductivity, we will be focusing on modelling only a) the tensile strength and (b) the thermal conductivity of the copper alloys in this study.

1. **Machine Learning**

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine Learning tasks can be broadly divided into the following few categories :

* 1. Supervised Learning
  2. Unsupervised Learning
  3. Semi-Supervised Learning
  4. Reinforcement Learning
  5. **Supervised Learning**

The objective of supervised learning algorithms is to infer a function from *labeled*[*training data*](https://en.wikipedia.org/wiki/Training_set) consisting of a set of *training examples*. In supervised learning, each example is a *pair* consisting of an input object (typically a vector) and a desired output value (also called the *supervisory signal*). These training data and their labels are used to producs an inferred function, which can be used for mapping new examples. In an optimal scenario, a supervised learning algorithm will then be able to correctly determine the class labels for unseen instances.

Supervised Learning can be further divided into two broad class of problems –

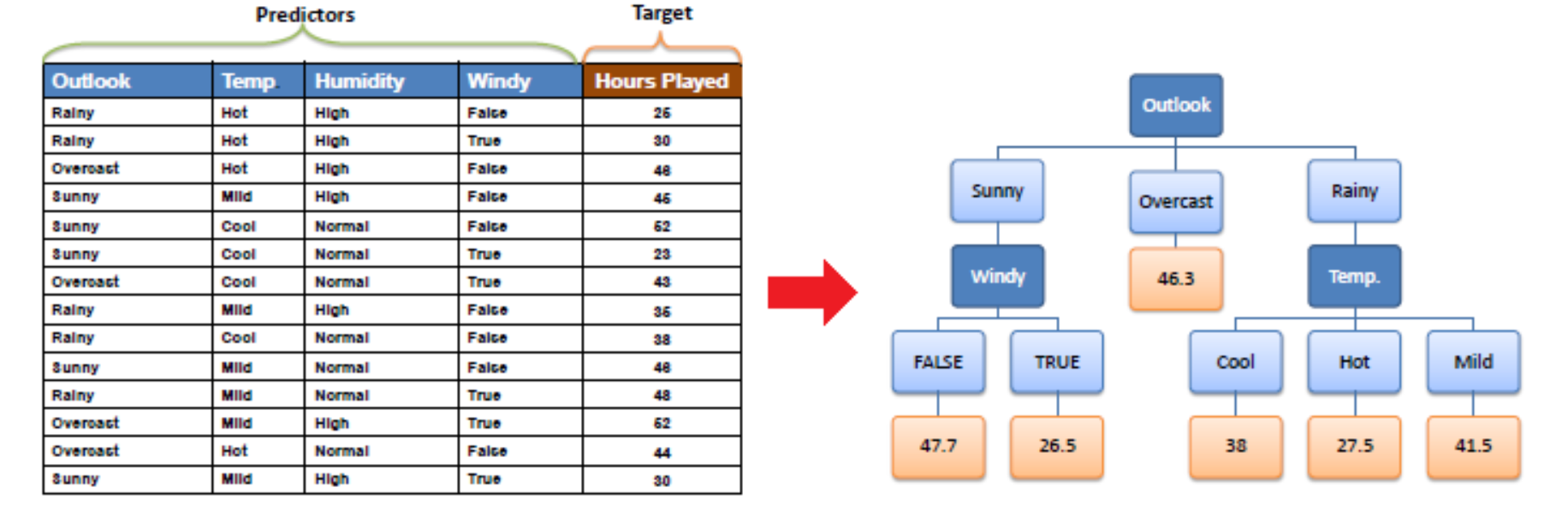
1. Regression
2. Classification
   1. **Regression**

If the labels provided to training examples in a problem are discrete values of a variable, then the problem belongs to category regression or prediction. In other words, regression finds a function that takes an input vector and maps it to a discrete value.

There are several regression models available today, however we will limit our discussion only to those models which are used in this study viz. Random Forest Regressors. But before we discuss Random Forests it is important to understand their building blocks which are Decision Tree Regressors.

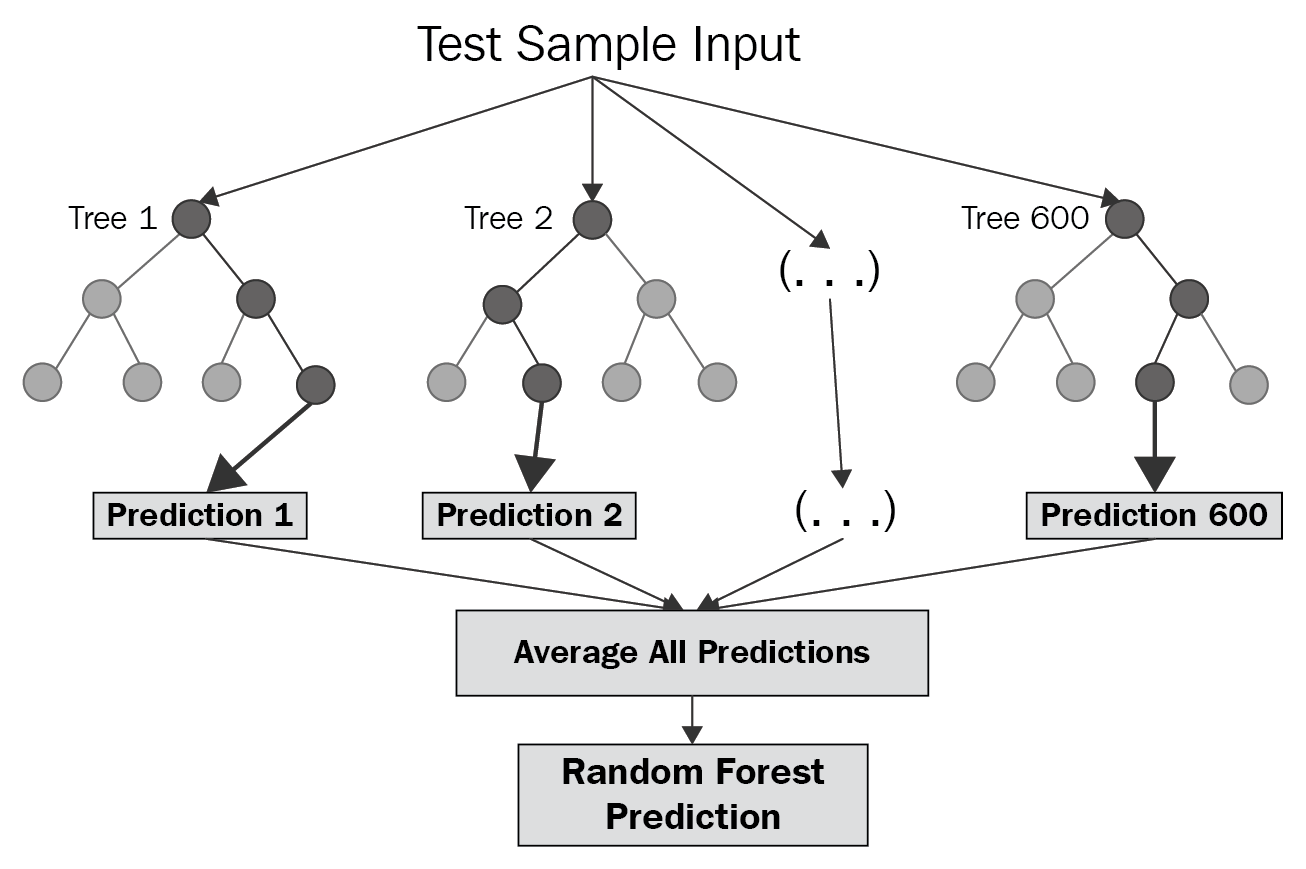
* + 1. **Decision Tree Regression**

Decision Tree Regression is a predictive modelling approach in machine learning which uses decision trees to go from observations about a training item (represented as branches) to model and predict a real-valued target label. They are among the most popular techniques because of their simplicity and interpretability.



* + 1. **Random Forest Regression**

Random Forest Regression is another powerful predictive modelling approach that builds up on the concept of decision trees and tries to address the problem of overfitting associated with them. A Random Forest Regressor is an ensemble bagging approach, which means that it creates several decision tree regressors at training time and its prediction is the mean value of the all the decision tree predictions.



1. **Overview of the Copper Alloy Dataset**

The data used in this study has been collected from multiple sources including the public databases(ref), research papers and educational software.

The dataset contains a total of 34 columns that contain the following information for each alloy instance –

* *2 Identifier Columns viz. Alloy Name and Temper Code*
* 2 columns for Processing Conditions viz. *Form* and *Temper*
* 28 columns for constituents of the alloy (each of these columns contains the contribution of the element to the alloys’ percentage by mass composition)
* Tensile Strength - measured in kilopounds per square inch (ksi) at room temperature (68 °F)
* Thermal Conductivity – measured in BTU/(h⋅ft⋅°F) at room temperature (68 °F)

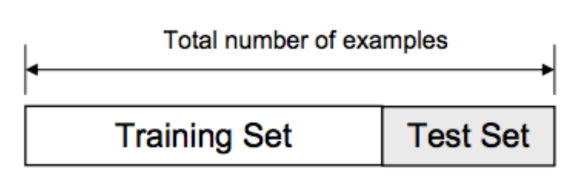
1. **Data Pre-processing**

Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we pre-process data before feeding it into our model. Some common data pre-processing are –

* Handling missing values
* Transformations such as data normalization
* Encoding categorical data
* Handling duplicate data

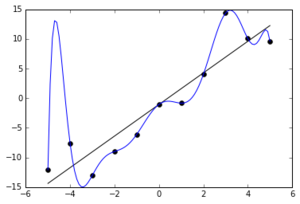
1. **Model Selection/Evaluation**
   1. **Train Test Split**

Train-test split is an essential part of the data science process where we train models on certain part of the data and test the model on unseen data to verify its generalization capabilities. Some models can have amazing scores on training data but poor performance on test data. This is an indication of overfitting by the model on the training data. In other words, it indicates that the model has trained on the noise in the data and memorized it



* 1. **Overfitting Example**

An example of overfitting can be seen in Fig. where we can compare the predictions of two different models represented by the red and blue lines on the data (represented by the dots).



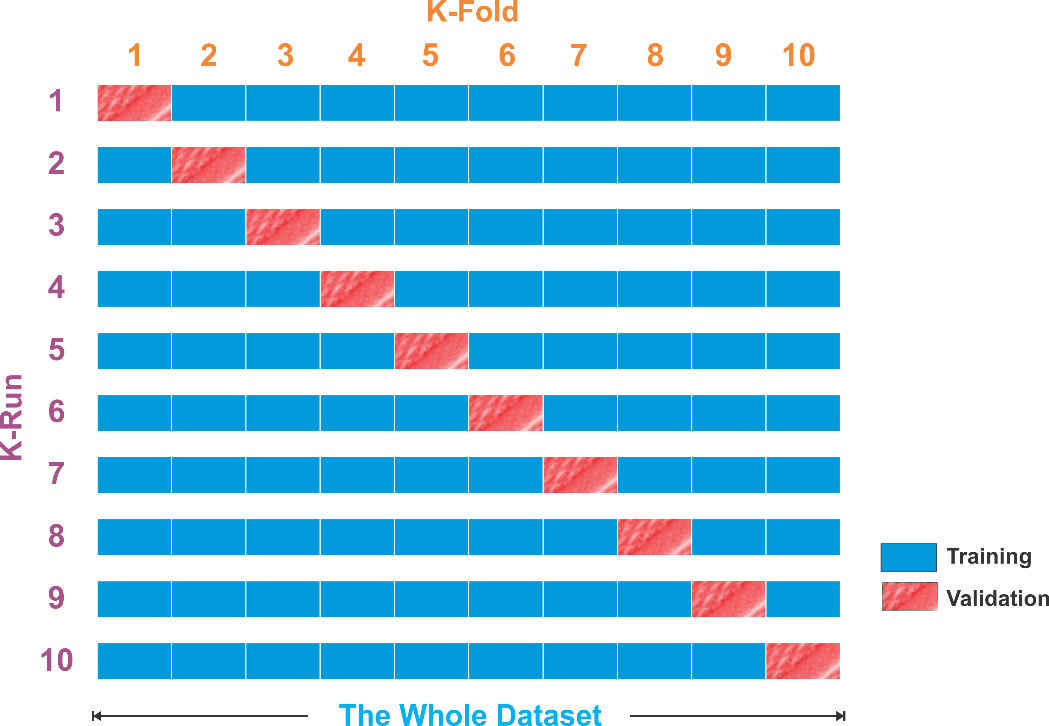
The function identified by model represented by the red line indicates a good fit to the training data and indicates good generalization capabilities. On the other hand, the model represented by the blue line has determined a complex function which, although, predicts all the training data correctly, is likely to fail in predicting unseen data. The blue line model is an overfitted model.

* 1. **10-Fold Cross Validation**

K-Fold Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. It is one of the most reliable methods for evaluating and comparing model performances. In this study we use k = 10 since it has been found through experimentation to generally result in a model evaluation metric with low bias a modest variance.

The general procedure is as follows and can be visualised as in fig:

* Shuffle the dataset randomly.
* Split the dataset into k groups
* For each unique group:
  + Take the group as a hold out or test data set
  + Take the remaining groups as a training data set
  + Fit a model on the training set and evaluate it on the test set
  + Retain the evaluation score and discard the model
* Summarize the skill of the model using the sample of model evaluation scores



1. **Hyperparameter Tuning and Some Approaches**

Hyperparameter tuning is defined as choosing a set of hyperparameters for a given machine learning model to optimise model performance.

* 1. **Grid Search**

Grid search is a traditional way to perform hyperparameter optimization. It works by searching exhaustively through a specified subset of hyperparameters.

Strength - *Guaranteed to find the optimal combination of parameters supplied.*

Weakness - *Can be very time consuming and computationally expensive.*

* 1. **Random Search**

Random search differs from grid search mainly in that it searches the space of hyperparameters randomly instead of exhaustively.

Strength *- Decreased processing time*

Weakness –