<Computational alloy design >

A thesis submitted in partial fulfilment of the degree of

Bachelor of Engineering (Honours)

by

<Venkata prakash>

<U6420727>

Supervisor: <Nick Birbilis>

Examiner: <Matthew doolan>

Australian National University

College of Engineering and Computer Science

The Australian National University

<june 2020 >

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university. To the best of the author's knowledge, it contains no material previously published or written by another person, except where due reference is made in the text.

<Venkata prakash>

9 June 2020

Acknowledgements

Firstly, I would like to express my honest appreciation to my supervisor Prof. Nick Birbilis for providing this engineering student a fantastic opportunity. His guidance is the reason for me to keep moving the initial research stage to the final stage of the project facing all the odds. Without this project, I wouldn't have even known anything about machine learning, which is the most booming subject around the world. I thank Nick Birbilis again for providing me continuous support and providing a path for me to learn about machine learning.

Secondly, I need to thank my computer science team member of the project, Abhinav Pandey, for constantly helping me in machine learning modeling.

Finally, I want to thank ANU for providing me the opportunity to explore the project that is required to interact with a completely different field of engineering

Abstract

Uncountable alloy combinations are possible with the available elements, and creating new ones require experimentation that costs money and time. Most of the manufacturing industries use this trial and error experimentation method to develop new alloys. This money and time problem is solved by using the machine learning approach to design a new alloy. This project developed two machine learning models using the random forest regression algorithm, one to predict thermal conductivity and to predict the tensile strength of using the chemical composition as input. The accuracy of the thermal conductivity model is 76.9 percent, and thermal conductivity is 89 percent. As the accuracies are good enough, these models are used to develop a reverse model that takes in tensile strength/thermal conductivity as input to produce chemical composition as output. The reverse model, which is the final model produced, can be used by manufacturers/researchers to skip most time spend during the experimentation stage in developing new alloys, as our model can directly provide the alloying elements to be used along with the chemical composition required to design alloy of desired thermal conductivity or tensile strength.

09 -TO DO

Contents

Chapter 1 Introduction

1.1 Motivation

An alloy is basically a blend of two or more elements out if which one is a metal. Alloying enhances the properties of metals often by making them more durable, more corrosion resistant, stronger e.t.c..,. There are various combinations of alloys possible, which can be achieved by combining various metals or elements to have different properties. This is not just limited to a combination of metals or elements but also extended to the composition of elements and metal that needs to be combined to make an alloy. Uncountable alloy combinations are possible by combining the existing metals with different elements and with different compositions. Moreover, all the different combinations will have different properties associated with it, either enhanced or reduced.

The creation of an alloy uses the conventional method used to design a new alloy, which involves a lot of testing, failure, re-testing in the process designing. The creation and manufacturing of new alloys consume a lot of time and money in the experimentation stage, which is a trial and error method. This project helps the life of the researchers easy and help manufacturing industries to manufacture alloys with enhanced properties in a short time. This project mainly focuses on copper alloys and desired properties themal conductivity and tensile strength which will be discussed further in scope of the project section.

This project aims to develop a machine learning model that takes in the desired properties of the alloy as input and provides the chemical composition required to make an alloy as output.

1.2 Research question :

How does designing the cu alloys with high thermal conductivity and strength with the use of machine learning can impact the manufacturing industries in future?

1.2.1 Connection with Real-world problems?

The final output for this project is the creation of a program with the aid of machine learning modeling, that takes in the desired properties of a copper alloy as input which is either thermal conductivity or tensile strength and produce the output as alloy combinations that are possible with the given input properties. This project helps to create new copper alloys with desired properties, which could help in advancement in many fields of industries such as advancements

in integrated circuits and electronic products and also in electrical wiring. This project helps the Research and manufacturing department of the manufacturing industry to reduce their cost, time they spend in coming up with combinations that are possible that match their desired property. This project helps the Research and manufacturing department by providing them possible combinations and the compositions range, which will help them to reduce time in the testing(trial and error) stage. This will save them a lot of time and money for laboratory testing and also helps them manufacture alloys at a cheaper rate. The impact of the project will be medium-high on the manufacturing industries as this leads faster production rate at less cost, and time will be a huge benefit to copper alloy manufacturing industries. This might lead to manufacturing industries to the use the approach of machine learning method for creating new alloys instead of conventional methods.

1.3 Objective:

- Creation a user/interface that takes the desired copper alloy property value as input and generate the chemical composition needs to make that alloy.

1.4 Broad Project context and Scope of the project:

1.4.1 Scope

The main focus of the project is Copper alloys, which are most commonly used in electrical industries as everyone knows that copper is good conductor of electricity. The desired properties that needs to be enhanced are tensile strength and thermal conductivity. Aaccording to Wiedemann-Franz law(Doitpoms.ac.uk, 2019) metals thermal conductivity is proportional to electrical conductivity in metal alloys at constant temperature. As all copper alloy data collected has properties at 273k, electrical conductivity is ignored, as the electrical conductivity of copper alloys increases with thermal conductivity as stated in the Wiedemann-Franz law above.

Exclusions:

Final output limited to Digital model : At the intital stage of the project, the final output is designing of alloys by manufacturing inorder to produce a physical mode, due to the covid19 circumstances, as there is least possibility for mobility(movement), scope is reduced to create a digital model that gives chemical composition as output with inputs being one of the desired properties.Hence there is no experimentation stage for this project, it is limited to digital model.

1.4.2 Broad project context:

Please note that detailed methodology is discussed in the methodology section, this is just a broad context of the project.

Data collection and data preprocessing: The projects starts with collection of all the copper alloy data avaliable from the reliable sources, which are evaluated. The data is collected from various resources like books, websites e.t.c. The data preprocessing steps helps to clear any errors in the dataset and helps to create a data set that is suitable for machine learning modelling.

Selection of model and training data set: These step involves in selecting a model that is suitable for our data set(which is regression in our case), therefore using the selected model to train the data set.

Validation and freezing the model: These step is to evaluate the model that is created and freezes the model to proceed to next phase. This model takes chemical composition of alloy as an input and outputs the tensile strength and thermal conductivity.

Reverse modelling and user interface package: These steps helps in building the final model, the name reverse modelling suggests that it's reverse of the previous model created. So this take the thermal strength or tensile strength as input and generates chemical composition as output. A final step is creation of user-interface script that takes input of any of the desired property and return chemical composition as output.

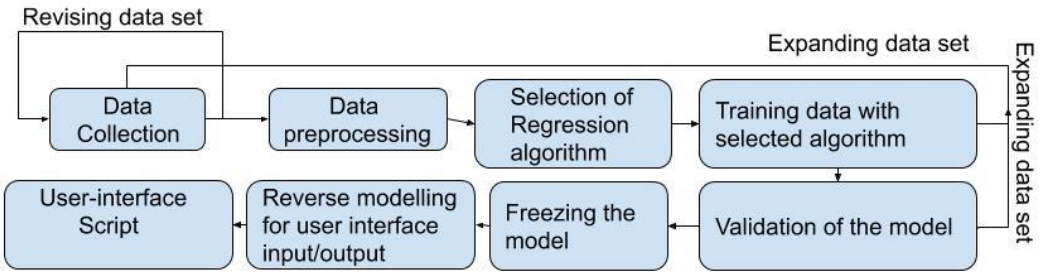

FIG: 1 Methodology of the project

Report outline: Chapter 1 introduces reader to the project, Chapter 2 provides some background knowledge related to this project. Chapter 3 provides you with the methodology followed inorder to achieve the results. Chapter 4, we discuss and analyse the results along with the limitations, future work of the research. Chapter 5 gives the conclusion to the project.

Chapter 2 Background

2.1)Copper alloys :

Copper alloy is a metal alloy of which main metal is copper and rest all are alloying elements. The properties of these alloys changes with the composition of the alloy and alloying elements(which are basically impurities). Copper alloys are wekk know for their high elelctrival conductivity, strength, corrosion resistant therefore used in most of the industrial applications . Unified numbeing system shortly(UNS) is used to notate the copper alloys. There are varies number of alloys available right now, from C100XX-C9XXXX with different properties and compositions. This combines both active and inactive alloys, only active alloys are considered for this project.

2.2)Desired properties:

The desired properties selected for this project are thermal conductivity and tensile strength. These two properties are used as input in our final user-interface model to generate the top 10 alloy combinations possible for that property.

a)Thermal conductivity:

Thermal conductivity is a material property, and it does not depend on the dimensions but depends on the composition and structure of the material. Thermal conductivity is the measure of the ability to allow heat that passes through it by conduction. (Thermal Conductivity of Metals and Alloys – Neutrium, 2020)

Units and ranges used for this project: : 226 - 13.3 **Btu/ sq ft/ ft hr/ °F at 68°F**

b)Tensile strength:

Tensile strength is nothing but ultimate tensile strength, which is maximum strength that the material can bear without breaking. Generally, better-performing copper alloys need to have high tensile strength. Tensile strength not only depends on the composition of the material but also the process it follows to manufacture it.

Units and ranges used for the project: 212 - 16.5 **ksi**

c) **Electrical and thermal conductivity relation**:
According to Wiedemann-Franz law, thermal conductivity is directly proportional to the electrical conductivity at a constant temperature. If the thermal conductivity of an alloy increases, then the electrical conductivity of the alloy increases. (Doitpoms.ac.uk, 2019)

2.3) Machinelearning:

Machine learning is a subset of Artificial intelligence that facilitates the system with the ability to learn automatically and better from experience without explicitly being programmed.
Machine learning algorithms used to train the input data in order to predict the outcomes. The more the data is fed to the computer, the better the mathematical relation between the data is analyzed using a machine-learning algorithm to predict accurately.

In simple words," **Machine learning** is an application of artificial intelligence that uses an algorithm to study the mathematical relationships between the data and help you make better predictions and decisions."
There are three types of machine learning

- Supervised learning
- Non-supervised learning
- Reinforcement learning

Only supervised learning was used for this project; the rest two are out of the scope of this project. The decision was made to use supervised learning, which will help us to use the regression algorithm, which satisfies the requirements of this project.

Supervised learning: In this algorithm, the training data will have the inputs correctly paired with the outputs. In the training stage, the algorithm used will search for mathematical relations and patterns in the data to the outputs in order to use the relation for predicting the new data. The main goal of this is to predict the new output based on the input given.

$$Y = f(x)$$

is the continuous output that needs to be predicted based on the input x, F(x) is the function created by the machine learning algorithm during its training stage. (A Brief Introduction to Supervised Learning, 2020)

Types of supervised machine learning:

- Classification algorithms
- Regression algorithms:

For classification algorithms, The output y is not continuous for this algorithm; it's discrete, which is categorical. For example, the output for a classification algorithm can be yes/No based on the inputs. It is not numerical/continuous.

Classification cannot be used for our project, as the output of our data set is a numerical value that can be satisfied by the regression algorithm. (A Brief Introduction to Supervised Learning, 2020)

NOTE: All the sample graphs that will be used in this section 2.4 are referred from various source, to explain the regression models. All these are sample graphs doesnot contain the actual/real data, but helps in learning about the alogorithm.

2.4) Regression algorithms:

All the regressions models are the part of supervised learning, where the input can be categorical/ numnerical columns but the output is continuous or numerical value.

2.4.1)Simple Linear regression:(SLR)

SLR is a machine learning algorithm that performs a basic regression with single feature input and output. The regression line is just like y=ax+b, a linear equation with only one independent variable X and one dependent variable y.

The independent variable is either continuous/discrete, but the dependent function is continuous.

For example consider a data set with input experience and output as salary which is described in the image below, The red points in the image shows the experience,salary pairs on the graph. The black linear line shown in the image is the best fitting linear line for the input and output relation.

Here **For the graph below the equation is → salary = experience*a + b**

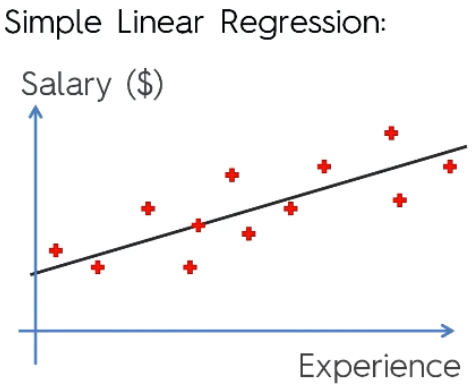

Fig:2 Simple linear example (**Kirill Eremenko**.2020)

2.4.2) MULTIPLE LINEAR REGRESSION:

Multiple linear regression belongs to the family of linear regression, but this has more benefits over the Simple linear regression when there are more independent features in the dataset. MLR can be used to model the data with multiple input features that unlike linear regression, which only maps a single input to a single output. MLR models the mathematical relation between the multiple input features and the dependent variable and tries to fit the data to a linear equation.

General equation: $Y = Ax_1 + Bx_2 + Cx_3 + \ldots\ldots\ldots + Nx_n$

Here y is the dependent variable, X1, X2, X3 ...Xn are independent variables.

.4.3) Polynomial Linear regression :

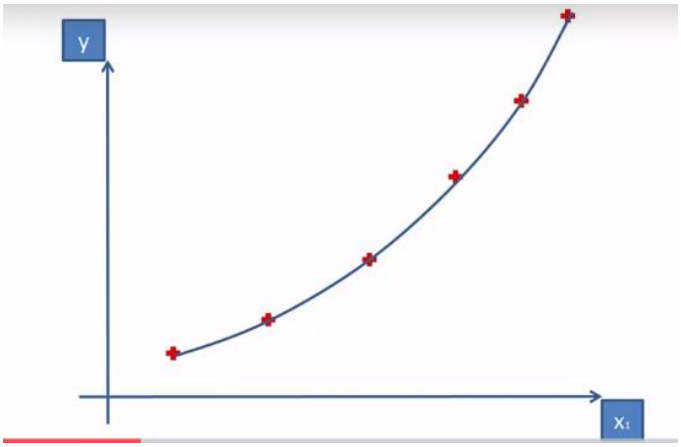

Fig: 3 Polynomical linear regression example (**Kirill Eremenko**,2020)

This regression is used mostly in cases where multiple linear equations don't fit the data. Polynomical regression finds the the mathematical relation ship between the independent and dependet variable and fits the data to a polynomial curve with a degree n.

Consider an example dataset with x1,x2…xi as inputs, and y as output if you refer to the image. The red point points are the input and output coordinate pairs if you refer to the image below. The blue curve joins the points is the best-fitted equation modeled by the regression algorithm.

General equation = $y = Ax^1{}_1 + Bx^2{}_2 + Cx^3{}_3 + \ldots\ldots\ldots + Nx^n{}_n$

Here y represents the dependent variable on the multiple features and x1, x2, x3….Xn represents the independent variables. The equation is the curve that passes through the data, which will created by finding on the mathematical relationship between the data.

2.4.4)Decision trees :

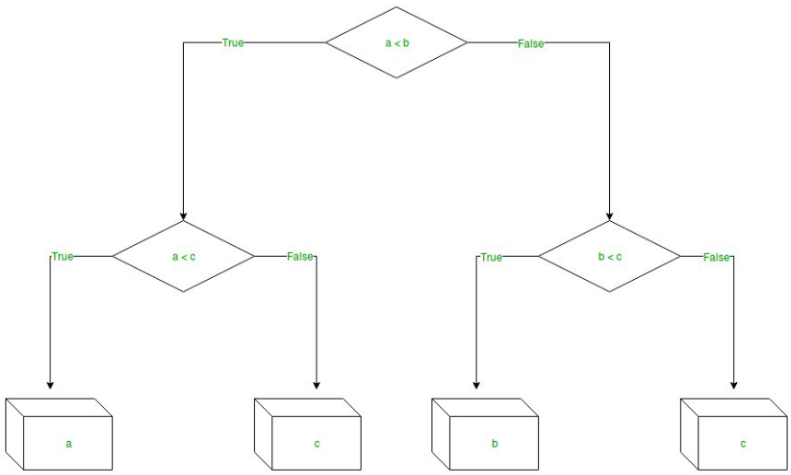

Fig: 4 Decision tree example (Decision Tree Regression in Machine learning, 2020)

The decision tree regression is the easiest, yet the most dominant machine learning algorithm.

Decision tree regression uses tree structure which is similar to a flow chart to predict the output based on input features. as said earlier, even this falls under supervised learning

Decision tree regression is used for the numerical output problem. Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.

For example consider the sample image above, which is the representations of a decision tree. The entire tree is bases on the decisions made by the algorithm. Nodes present in the tree represents the condition, based on the condition it goes either right or left.
The end nodes which are called as leaves represents the ouput of the alogorithm based on the decisions made.
The decision tree used here looks simple whis is used to explain how the algorithm works, but for actual dataset created in this project there will huge number of branches and decisions. (Decision Tree Regression in Machine learning, 2020)

2.4.5) Random forests :

Random forests is based on the decision tree which acts as a building block for this regression.

Random forests is fast, powerful and simple supervised M L algorithm. It dominates decision trees as this depends on the many decision tress rather than depending on a single decision tree. This algorithm works for large range of datasets.

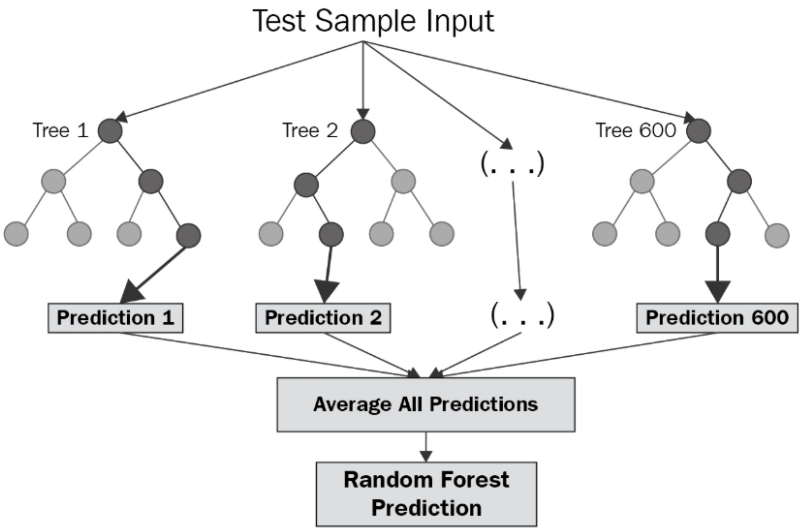

Fig 5 : Random forest structure. (Random Forest and its Implementation, 2020)

The idea of random forests to combine as many random trees as possible in a single model to yield the output. If you look at the example decision tree above, it shows how the random forests works. As we have seen in decision trees section the final output of the decision tree is the final node/leaf of a tree. For random forests, all the final outputs from huge number of decision trees are averaged in order to find to final output. This is more reliable than the decision trees as multiple decision trees are used. (Random Forest and its Implementation, 2020)

2.5)Hyper parameter tuning:

This is more practical than conceptual unlike the regression models explained above. For regression algorithm in python(scikit-learn), each algorithm comes up certain parameters. In order to have best functioning model we need to have the best parameters inorder to produce better output and enhance the performance of the model. By hyper-parameter tuning method, we will be able to find the best parameters for the algorithm.

For example random forests, these are the hyper parameters for random forests algorithm in from scikit-learn.

```
'n_estimators': n_estimators,
'max_features': max_features,
'max_depth': max_depth,
'min_samples_split': min_samples_split,
'min_samples_leaf': min_samples_leaf,
'bootstrap': bootstrap
```

All these parameters will have some range of values and different values make algorithm perform differently, the hyper parameter tuning is performed by running the algorithm through all the selected range of possible values for the parameters and gives the parameters that perform best with the model.

Please refer to the appendices C, to know how hyperparameter tuning is implemented using randomized search

2.6) Validation:

Generally, we don't know how a model perfrm withoit testing it with a new data. Overfitting is the result of the model fitting data at it's best but performing worst with the new data set.

This may be caused due to the noise in the data or lower data size. In order to overcome the problem of overfitting there are two methods used in this project:

2.6.1) K-Fold Cross Validation:

In this validation technique, the dataset is divided inko some k-subsets(k= number of subsets). This method performs the validation k-times, each time one of the k subsets is used as a test set and rest of the subsets together put as train set. Each run gives a validation score maximum is one ,maximum score means least overfitting and best performing model. This k-trails produces K- cross validation scores and the average value of all the scores is K-fold cross validation score.

The benefit of this k- trails, at every trail, there will be different combinations of train and test sets, which helps the model to test on the new dataset eachtime. The image below shows us how the data is divided into training and test sets base on the K-value. K-value considered for this example is 10, so there are 10 sub datasets and 10 trails, each trails give a cross validation score and all the scores for 10 trails are averaged to give the final value. If the K-fold value is near to one, then it performs better, if it's very low, the model doesnot perfoem well on new dataset which is ultimately overfitting. (Cross-Validation in Machine Learning, 2020)

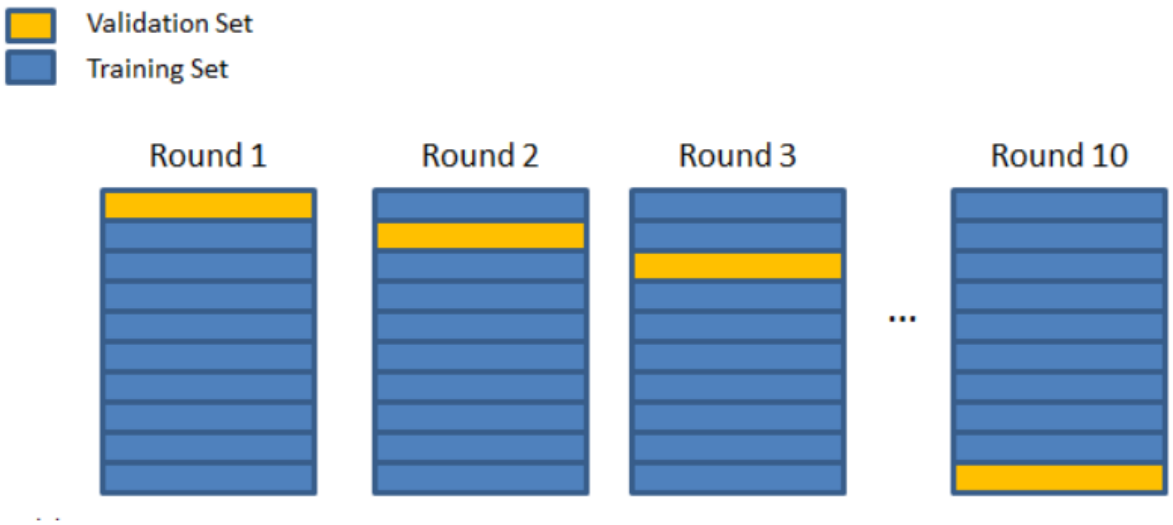

FIG 6 : Cross validation process(Cross-Validation in Machine Learning, 2020)

2.6.2) Addition of more data points:

One method to reduce the overfitting is by adding more data, over fitting can also be caused due to the insuffucient data size. By adding more data to the data set, the performace of the model can be imporved by reducing the overfitting. (Cross-Validation in Machine Learning, 2020)

Chapter 3 Methodology

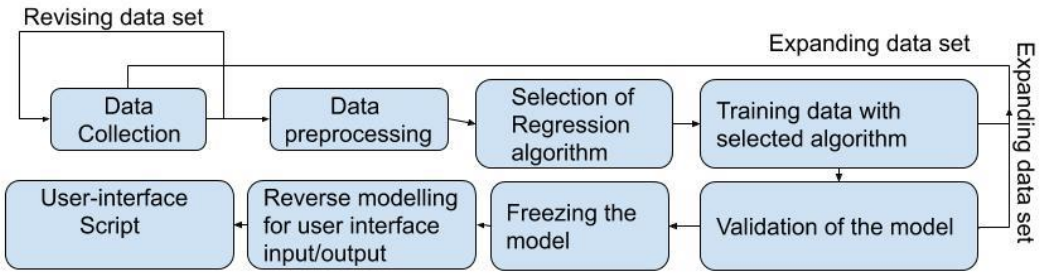

Fig7: The flow chart represents the methodology followed for finishing this project.

3.1) Data collection method

The most important stage of this project machine learning methodology is the data collection stage. The data collected in this stage, will help the supervised machine leaerning model to predict the outputs. This stage needs to be completed first inorder to move to the next stage. This stage involves collecting the copper alloy data from variable reliable resources. The main features of the data are chemical composition, thermal conductivity, tensile strength and temper method of an alloy.

For the final objective of this project, it requires creating an interface that takes in thermal conductivity or tensile strength and output possible compositions of an alloy. Hence we are using the machine learning algorithms to find the mathematical relationships/ patterms between independent variables chemical composition,temper code and dependent variables tensile strength and thermal conductivity.

Data set format:

The image below shows the format of the data set that is used for the project. Out of all these coulumns Name and temper will be left unused for the modelling. As name column will not have any relation to data and hardly used to determine anything , this is ignored during

modelling. Whereas, temper code is used instead of the temper which tells us the temper method it follows to manufacture.

Name	Form	Temper	Temper Code	Tensile Strength	Thermal Conductivity
C10100 Oxygen-Free-Electronic	Flat Products	1/2 Hard	H02	42	226
C10100 Oxygen-Free-Electronic	Flat Products	1/4 Hard	H01	38	226
C10100 Oxygen-Free-Electronic	Flat Products	1/4 Hard	H01	38	226
C10100 Oxygen-Free-Electronic	Flat Products	1/8 Hard	H00	36	226
C10100 Oxygen-Free-Electronic	Flat Products	1/8 Hard	H00	36	226
C10100 Oxygen-Free-Electronic	Flat Products	As Hot Rolled	M20	34	226
C10100 Oxygen-Free-Electronic	Flat Products	As Hot Rolled	M20	32	226
C10100 Oxygen-Free-Electronic	Flat Products	Extra Spring	H10	57	226
C10100 Oxygen-Free-Electronic	Flat Products	Hard	H04	50	226
C10100 Oxygen-Free-Electronic	Flat Products	Noal Grain Size 0	OS025	34	226
C10100 Oxygen-Free-Electronic	Flat Products	Noal Grain Size 0	OS050	32	226
C10100 Oxygen-Free-Electronic	Flat Products	Spring	H08	55	226
C10100 Oxygen-Free-Electronic	Rod	As Hot Rolled	M20	32	226
C10100 Oxygen-Free-Electronic	Rod	Hard	H04	55	226
C10100 Oxygen-Free-Electronic	Rod	Hard	H04	45	226
C10100 Oxygen-Free-Electronic	Rod	Hard	H04	48	226
C10100 Oxygen-Free-Electronic	Rod	Noal Grain Size 0	OS050	32	226
C10100 Oxygen-Free-Electronic	Shapes	As Hot Extruded	M30	32	226
C10100 Oxygen-Free-Electronic	Shapes	As Hot Rolled	M20	32	226
C10100 Oxygen-Free-Electronic	Shapes	Hard	H04	40	226
C10100 Oxygen-Free-Electronic	Shapes	Noal Grain Size 0	OS050	32	226
C10100 Oxygen-Free-Electronic	Tube	Hard Drawn	H80	55	226
C10100 Oxygen-Free-Electronic	Tube	Light Drawn, Ligh	H55	40	226
C10100 Oxygen-Free-Electronic	Tube	Noal Grain Size 0	OS025	34	226

Thermal Conductivity	Cu(1.2.3	Pb	Zn	Fe	P	Ag	As	O	Sb	Te	Ni	Bi	Cd	sn	S	Mn	Zr
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0
226	99.99	0.0005	0.0001	0.001	0.0003	0.0025	0.0005	0.0005	0.0004	0.0002	0	0	0	0	0	0	0

Fig: 8 Data set (Copper.org, 2019)

Why do we have the temper code and form columns in the data set?

As the thermal conducctivity is directly dependent on the composition and structure, it only depends on the composition of the alloy. Whereas the tensilse strength depends on the form and the temper process it follows, so the form and temper columns are necessary in order to predict the tensil strength property.

Early trends from data set:

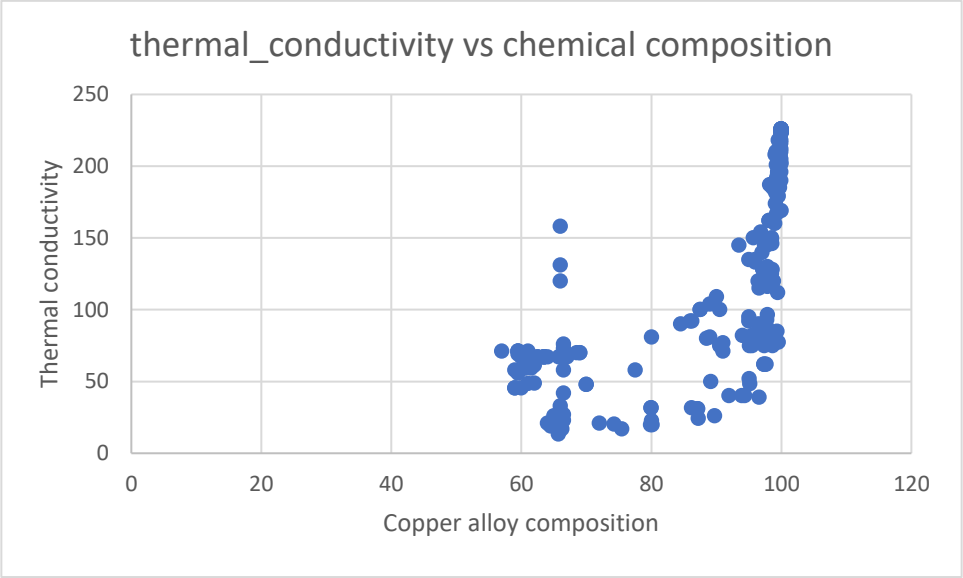

Fig: 9 Thermal conductivity vs copper composition

From the data set available, from the image it can be seen that the thermal conductivity of the alloy increases with the chemical composition of the copper in the copper alloy. Which also states that electricial conductivity increases with the copper composition, as the thermal conductivity is directly proportional to electrical conductivity at constant temperature.

Reliability of the dataset:

Most of the copper alloy data collected is from the copper alloy organisation which is a member of interanational copper alliance. This tells you the reliability of the data we have collected.

For tensile strength 1900 data points are collected from the reliable resources, for the thermal conductivity, 210 distinct data points are collected removing the similar values.

For a single copper alloy assume C10100 in the data set shown in the fig:8, there are distinct values for tensile strength as it varies with temper code, but only one value for thermal conductivity”226” as it is dependent only on the chemical composition. So there will be only one thermal conductivity value for an alloy but multiple tensile strength values for an alloy. This is reason for the huge difference in the collection of data points of thermal conducitivity (210 data points) and tensile strength(1900 data points).

3.2) Data pre-processing stage:

The first stage of data pre processing is filling the missing values in the dataset, at the initial stage there were some missing numerical values in the data set, all these values were filled by taking mean of the columun such that it will not cause any deviation in the dataset. If there are any missing values, then it is not a suitable dataset for machine learning modelling. So, this step can also be considered as cleaning of data set in order to convert it into the useful and efficient format required for machine learning.

3.2.1)Dummy variable for categorical columns:

There are some categorical columns in the data set referred in Fig 8

For example, conside the form column in the origibal data set added in data-collection phase. It has different types of forms and all these are categorical which could not be recognised by the machine learning algorithm. In order to use this data in the modelling phase, the categorical column in converted into dummy variables.

as centrifugal cast	as sand cast	bar	die forging	flat products & rod	flat products & wire	flat products, rod & bar	forgings	pipe	plate	plumbing fittings	products	rod	rod & bar	rod & wire	rod, bar &
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig:10 Dummy variables in data set

If you look at the figure 10 it shows all the different types of forms like bar, plate, forgings e.t.c.., as a separate column. For the first row, it shows "1" for flat-products columns(products in image), which says the form is flat product for the first row which is true according to the original data set. This is how dummy variables are used to replace the categorical data by converting it into a numerical value.

This dummy variable creation step is very important for the data inf there is a categorical data present in the dataset, as the regression analysis requires all the independent to be numerical.

Once these two processes are completed, then the the data set is ready for the modelling phase.

3.3) Selection , Training and validation of the algorithm:

Note:All the modelling processess have been performed using the SCIKIT-LEARN which is a machine learning library available for python.

The regression model is used when the output is continuous/ numerical. Based on the data set format, it can be clearly seen that the final output needs to be a continuous/numerical value I,e, thermal conductivity or tensile strength. So, regression model is the only possible algorithm used to model this data.

3.3.1) Selection of regression algorithm for tensile strength:

For tensile strength , the input parameters (independent) variable are the composition, temper code and the form of the alloy and the output(dependent) is the tensile strength of the alloy.

Note: Validation of the results: K-fold cross validation which is described in the background section is used to test the accuracy of the model. This model will also test the model in terms of overfitting/ underfitting. The selection of the model is soley based on the cross-validation score as better Cross- validation score will make sure that the model performs better with new data.

From the table: 1, C.V mean score represent the Cross validation score that will be used in to select the final algorithm.

The C.V mean scores from the above table is used for selecting the model.

Tensile Strength

	R-squared on Test	CV mean	CV median	CV std
Random Forests	0.918125	0.896573	0.892660	0.034438
Decision Trees	0.905646	0.851480	0.839740	0.054939
Multiple Linear Regression	0.769285	0.826907	0.813903	0.066037
SVR	-55109.750385	-0.066808	-0.071163	0.034479

Table 1: Cross validation score for tensile strength regression models(pandey,2020)

3.3.1a)Simple Liner regression:

This linear regression cannot be used for the data set as this is only for single independent variable and sigle dependent variable. Whereas out data has multiple input features which is impossible using this Linear regression method.

Result : Simple Linear Regression method is not eligible for the modelling, so not used for the modelling.

3.3.1.b)Multiple Linear regression: (MLR)

MLR is the third best performing model for predicting the tensile strength with CV mean score of 82.6 percent. The score is above 70 percent, so this model is eligible for proceeding to the next stage which is reverse modelling.

The graph shows the predicted and the actual values of the tensile strength, the blue point represents the predictions, it can be seen that this MLR performed well in this stage. Both the C.V mean score and the the graph shows positive results.

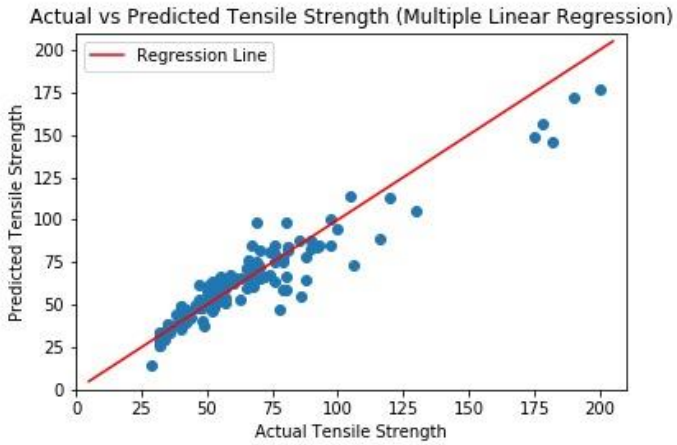

Fig 11: Predicted vs actual values graph for MLR :

3.3.1.c) SVR(support vector regressions):

This regression is not discussed in the background section, as this is not commonly used regression algorithm, this is mostly commonly used for classifications. This model peforms very bad in both the cases, so restricted the use of this. Presenting the graphs for this just to use this a reference to compare the other algorithms.

If you look at the table:1 the CV average score for this model is very low -0.006 out of all the tested algorithms.

The graph shoes the unrelaibale results for the predicted values when compared to the actual values, as the C.V score is very low for this, the model is eliminated from the selection.

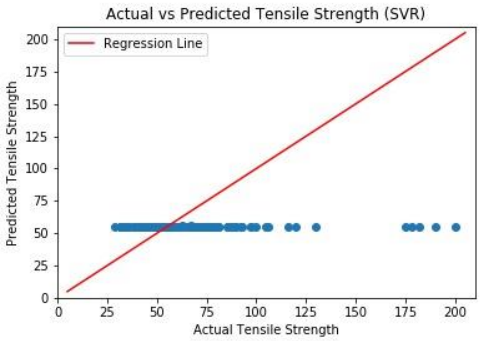

Fig12 : Predicted vs actual values graph for MLR

3.3.1.d) Decision trees:

Decision trees is the 2nd best performing model out of all the models that were used to train the data set. The C.V mean score is 85 percent, which is good enough to proceed to the next step. The graphical representation also shows us that the most of the predicted values of nearby the acutal data line.

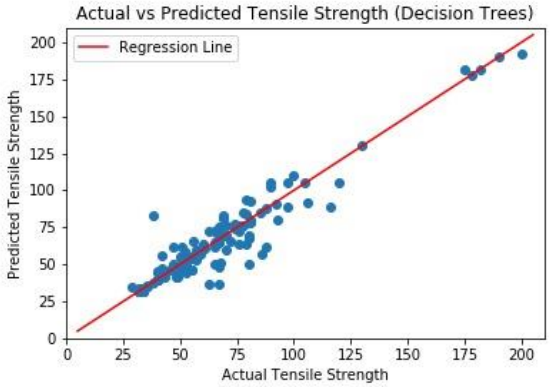

Fig13 : Predicted vs actual values graph for MLR

3.3.1.e) Random forests:

Random forests are the best performing model out all models that were used to train the data. The C.V mean score for this is 89 percent, which is the highest of all the models. The graph shows pretty good results for the predicted values. Out of all the graphs, this shows the best output for the predicted values.

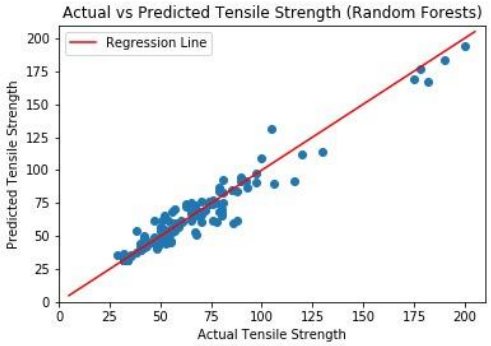

FIG:14 Predicted vs actual values graph for Random forests

Final selection: RANDOM FORESTS REGRESSION MODEL

The graphs can't judge the model's selection, but the selection of the model is judged by the cross-validation mean score used to measure the accuracy. As the best model from the C.V score is **Random forest** with 89 percent accuracy, this tensile strength model is frozen and proceded to the Reverse modelling step.

3.3.2) Selection of regression algorithm for thermal conductivity:

For tensile strength , the input parameters (independent) variable are the composition, temper code and the form of the alloy and the output(dependent) is the thermal conductivity of the alloy. Below graphs and score are generated from the python, using scikit learn.

Thermal Conductivity

	R-squared on Test	CV mean	CV median	CV std
Random Forests	0.808500	0.764179	0.802906	0.176773
Decision Trees	0.840471	0.538287	0.664599	0.377613
SVR	-9.370615	0.235385	0.226133	0.112811
Multiple Linear Regression	0.494418	-194.629872	0.519930	388.637058

Table: 2 **Cross validation score for thermal conductivity regression models(pandey,2020)**

For thermal conductivity, there is the only model that scores above 70 percent of the Cross-validation mean score. As this is the only value that is reliable for proceeding to the reverse-modeling stage, the random forest model is selected for thermal conductivity. The random forests model for thermal conductivity predicitions is used for reverse modeling.

Final selection: RANDOM FOREST REGRESSION MODEL

Please refer to the **appendices:D** to have a look at all the graphs of the regression models used for thermal conductivity.

3.4) Freezing and Reverse modeling:

3.4.1) Freezing the model:

The final models for thermal conductivity and the thermal strength have been selected from the previous step based on the Cross-validation(CV) scores. As the C.V score is above 70 percent for both the values, it is good enough to proceed to the reverse modeling step. This intuition based on the information received by the supervisor; he said if the c.v score is 70 percent(0.7), then the model is good to proceed to the next step.
With his instructions, we have frozen two models that were selected to proceed to the next stage; further, more changes will not be made to the models selected.

3.4.2)Reverse modeling:

Before this stage, the model is used to predict thermal conductivity or tensile strength based on the input features of chemical composition and strength. The project's main objective is to provide the chemical composition as output using the tensile strength as an input.
This step is used to create a reverse model that takes in either thermal conductivity and tensile strength as input and provides the chemical composition as output.

3.4.3)Methodology of the reverse model:

- **Step: 1** Collection of best samples from the database based on the desired input property:

In the step-1 of the reverse modeling process, the input value is taken from the user, which can be either thermal conductivity or tensile strength.
For example, if the input given by the user is 100ksi for tensile strength, this step will collect all the chemical compositions data from the original dataset that consists of tensile strength values near 100KSI(input). All the collected chemical composition data points will have tensile strength near to the input value given by the user are collected at this stage. These chemical compositions of alloys that were collected are the best samples.

- **Step: 2** Generating synthetic data of composition from the best samples:s

With the best samples collected from the dataset, for example, ten chemical composition data points collected from step 1. All these ten compositions are used to create synthetic data. Assume that these ten compositions have ten columns listing all the alloy elements, therefore

100 cells, the synthetic data consists of all the permutations of 100 cells that are 100! This will build an enormous synthetic data set.

- **Step3**: Creation of a data frame(one for each)

The synthetic data is consists of only chemical composition, which could only be used to predict the thermal conductivity, so this data can be directly used to predict Thermal conductivity.

The synthetic data will have add-ons such as temper code and form to the synthetic dataset to have all the variables required to predict the tensile strength.

- **Step 4**: Training the synthetic data to predict the tensile strength/thermal conductivity

These two synthetic data sets for thermal conductivity and tensile strength are trained with the two models that were selected, one for thermal conductivity and one for tensile strength from section 3.3.

- **Step 5**: Comparing the predicted values with the input from the user

There will be predicted values from each of the tensile strength and thermal conductivity algorithms in the training stage. These predicted values are then compared to the input given by the user. Let us use the same example used in step 1, and the input user property is 100 ksi(tensile.strength), the predicted values that have a value near to 1000ksi will be selected. The chemical compositions data points of these predicted values that are selected will be given as output to the user.

- **Step 6** : Recommendation of top 10 results

This final step will take the chemical compositions from step 5 and provide it to the user. This recommendation step will also include the chemical compositions from the original data set if the user input property value matches exactly with the property values present in the original data set.

The interface of this will be discussed in the results section along with the top 10 results.

3.5) Creation of user-interface script and end of the project:

with the completion of reverse-modeliing, this entire model is frozen and converted into a single script. We have a user-interface script that takes input as thermal conductivity/ tensile strength and provides chemical compositions as the output. This is the end stage of the project.

Chapter 4 Results and analysis

NOTE**:** In section 3.3.1, 3.3.2, the analysis is already done on the results to select the best performing models based on the Cross-validation score. This section only discusses the results, limitations of the results. A comparison of the final output results with the research question is discussed in the conclusion section**.**

4.1) Thermal conductivity predictions:

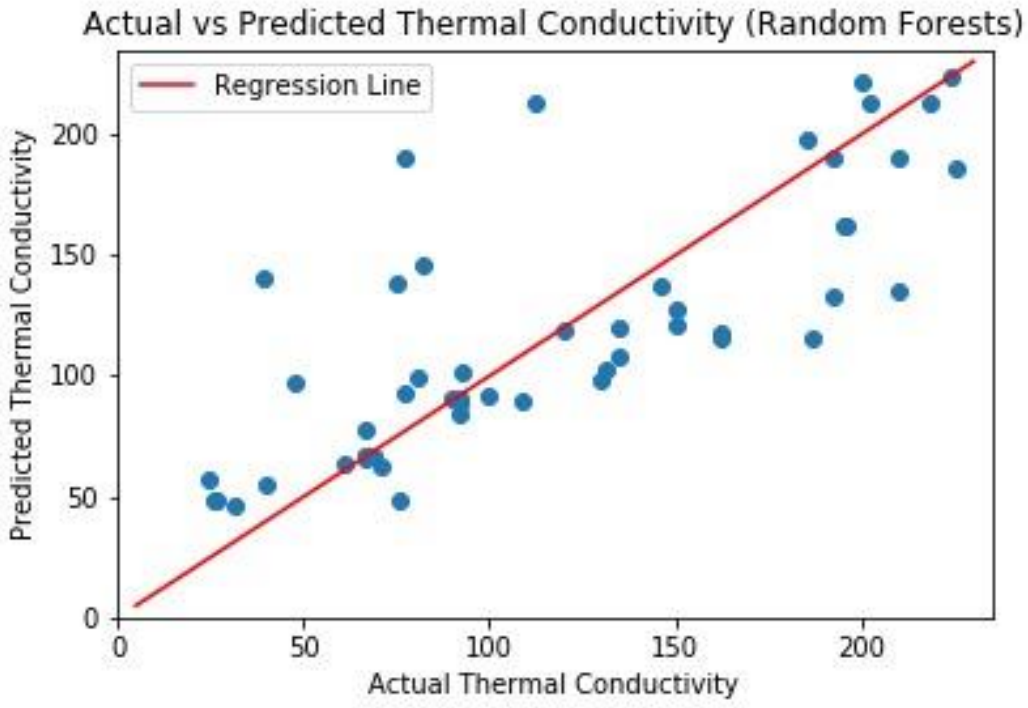

FIG: 15 Predicted vs actual values graph for Random forests (thermal conductivity).

Discussion: From the graph, it can be seen that most of the predicted values(blue points) are near the red line representing the actual values. But many predicted points are far away from the line. What does that mean? If you refer to the Cross-validation score for thermal conductivity discussed in the methodology section, it's 76.9 percent, which is not highly accurate but good enough to use the model to predict the outcomes.

Limitations:

Accuracy of the model is limited to 76.9 percent using the random forests algorithm. In contrast, other models like Multiple linear regression, SVR, Decision trees have not even shown 50 percent accuracy in predicting the data. This might be due to the insufficient data points for thermal conductivity modeling.

Future work: The only relationship that is used to identify the thermal conductivity is chemical composition. Future work aims to find if thermal conductivity is related to any other variable other than chemical composition and iterate the process again to produce the best results.

4.2 Tensile strength predictions:

The random forest is used for predicting the tensile strength of the material using chemical composition and temper as inputs. The image below shows the predicted and actual values of the entire data set after training the data set using the random forest algorithm. The blue line represents the real values, and the red points represent the predicted outcomes. The results are highly likely as most of the points fit the real values.

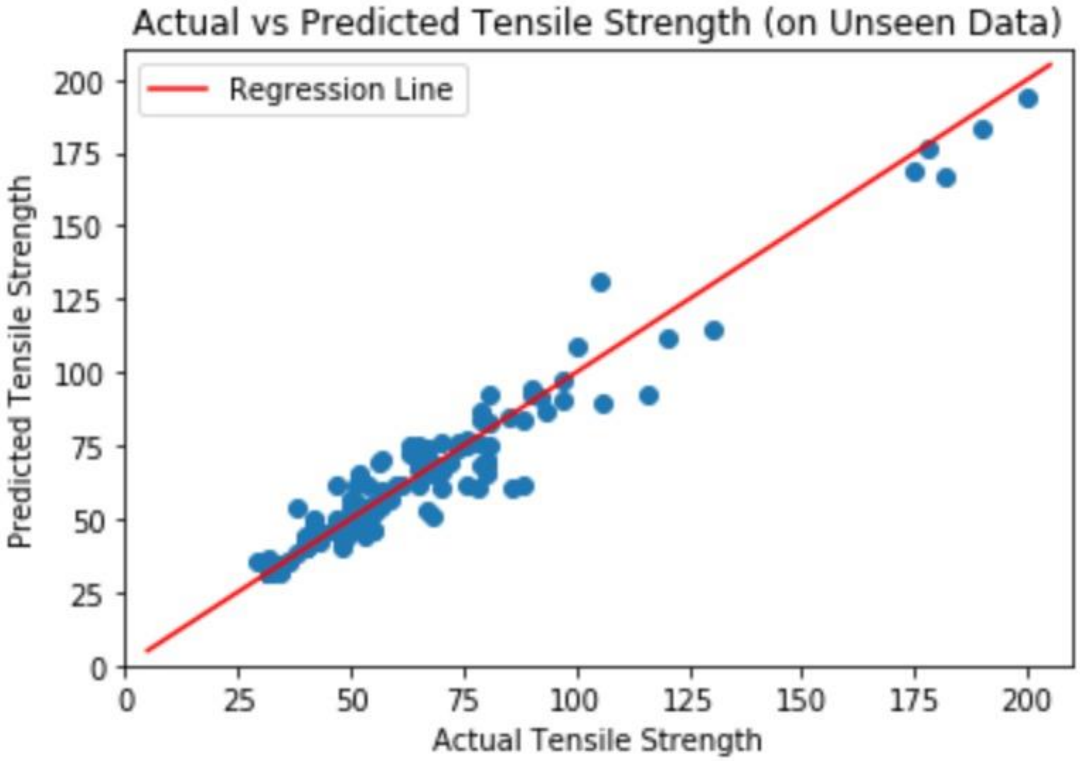

FIG: 16 Predicted vs actual values graph for Random forests for tensile strength

Accuracy of the results:

For tensile strength, we have a huge number of distinct data points almost 1900, so there is no chance of overfitting due to insufficient data. The cross-validation score for this random forest

model is 0.89, which is almost 90 percent, this is very high. No further work needs to be done in case of thermal conductivity.

4.3) Results from reverse modelling :

The image below represents the reverse modeling interface, which is the final output of this project. The method represents property, either tensile strength or thermal conductivity. N represents the number of outcomes required by the user. The value represents the input value from the user for the specific property. From the image below, method 2:thermal conductivity is selected, with a thermal conductivity value of 56. When this code is run, the output is generated with the confidence percentage and all the chemical compositions that are required to make the alloy

Final output: If you refer to the image, the final out put is the top 10 combinations of the chemical compositions that can be used to make an alloy with thermal conductivity 56.

[23]:
```
# User Input
METHOD = 2  # method = 1 : Tensile Strength, method = 2 : Thermal Conductivity
VALUE = 56
N = 10

# Program functionality

# Global Variables
DATA = pd.read_csv("../../Copper Dataset (version 3.0).csv")

# Load model corresponding to the METHOD
model = select_appropriate_model(method = METHOD)

# Generate Synthetic Data for given Tensile Strength
synth_data = generate_synthetic_datapoints(desired_value = VALUE, data = DATA, method = METHOD, n = 10)

# Obtain most suitable alloy compositions with confidence percentage
results = composition_from_value(desired_value = VALUE,
                                 synth_data = synth_data,
                                 model = model,
                                 method = METHOD,
                                 n = N).round(2)
results
```

[23]:

	Confidence %	cu	pb	p	ni	sn	co	be
0	73.8	95.13	0.38	0.19	0.46	0.92	0.00	2.92
1	73.8	94.89	0.40	0.20	0.48	0.97	0.00	3.06
2	73.6	95.31	0.38	0.00	0.46	0.92	0.00	2.92
3	73.6	95.08	0.40	0.00	0.48	0.97	0.00	3.06
4	73.5	94.84	0.38	0.19	0.46	0.92	0.31	2.91
5	73.4	95.02	0.38	0.00	0.46	0.92	0.31	2.91
6	73.1	94.59	0.40	0.20	0.48	0.96	0.32	3.05
7	73.1	96.01	0.39	0.19	0.46	0.00	0.00	2.94
8	73.1	95.82	0.41	0.20	0.49	0.00	0.00	3.09
9	73.0	94.78	0.40	0.00	0.48	0.96	0.32	3.05

Figure: Reverse modelling(Pandey,2020)

Limitations :

- The outputs are not 100 percent accurate, as the random forest models used to produce the outputs are not 100 percent accurate.
- Separate model for thermal conductivity and tensile strength.

- No experimentation and design stage: In order to design and test the alloys from the outputs generated from the digital model, there needs to be experimentation and design stage, where the experiments are performed using the chemical compositions provided by the digital model to create an alloy. The tests are performed on the designed alloy to find if the alloy designed matches the properties of the predicted alloys. All these steps stopped due to COVID-19 circumstances.

Chapter 5 Conclusions and outlook

5.1 Response to the research question

This project has become successful in developing a model that predicts an alloy's chemical composition using the input properties of tensile strength or thermal conductivity.
The project can impact the alloy manufacturing industries in many ways, mainly by reducing the cost and the time taken in the laboratory to manufacture the alloy. This final model can help the researcher's in developing new copper alloys easily. For example, if a researcher wants to manufacture a copper alloy with tensile strength 150ksi, the value can be used as input for this model developed in the project and get chemical composition required to make the compound as the object. So, this will cut the time spent by the researchers in the trial and error method to come up with compositions required to have the desired property.
The compositions generated by our method is not completely accurate as our models are not 100 percent accurate, but this provides with the estimations of the chemical composition and alloying elements needed to make an alloy. This will result in low designing costs and high productivity because of the low time, which will force the research and development departments of manufacturing companies to adapt to the machine learning methodology to design new alloys.

5.2) What worked well and what did not work well is research?

The data collection stage worked well for the collection of data points for thermal strength as we have collected around 1900 data points, which helped us to get the best results using the machine learning algorithm. The more the data, the better will be the output.
The data collection stage did not work well for thermal conductivity, as there are more repetitive data points in the data set totaling 1900 data points and only 210 distinct points. This caused a major problem in the machine learning process while making predictions.

5.3) Limitations of the approach and future work:

- Data set: The dataset size is very less for thermal conductivity; more data needs to be added in the future; therefore, improving the performance of the thermal conductivity regression model.

- Separate models for Thermal conductivity and tensile strength: At the initial stage of the project, the main focus was on creating a model that takes in both thermal conductivity and tensile strength of the materials. By the end of the project, we ended up with two separate models for thermal conductivity and tensile strength. This is one of the major limitations of the project.
- Regression models used for this project are inbuilt and are provided by the sci-kit learn library. Neural networks were not even considered for this project as the project team lacks the knowledge; in the future, I strongly recommend using neural networks to understand the mathematical relationship between the data and to predict the outcomes more accurately.
- In regression algorithms used, the polynomial regression model is not used for predicting because of the dimensionality of the dataset, which makes it very hard to find the relations between the data using polynomial features. We did not further continue using this regression model as we assumed it would not work for this project. We have come to this agreement without even testing it. These works need to be done in the future, which could find the accurate mathematical relationship between the dataset.

5.4) Final statement:

This project shows us the advantage of machine learning in the materials engineering field, which tells us the importance of interactions in engineering fields to develop new ideas and methods to solve a problem

References

- Doitpoms.ac.uk. (2019). DoITPoMS - TLP Library Introduction to thermal and electrical conductivity . [online] Available at: https://www.doitpoms.ac.uk/tlplib/thermal_electrical/printall.php [Accessed 25 Oct. 2019].
- Neutrium.net. 2020. *Thermal Conductivity Of Metals And Alloys – Neutrium*. [online] Available at: <https://neutrium.net/heat_transfer/thermal-conductivity-of-metals-and-alloys/> [Accessed 9 June 2020].
- 2020. [online] Available at: <https://www.udemy.com/course/machinelearning/learn/lecture/5732730#overview> [Accessed 9 June 2020].
- Medium. 2020. *Decision Tree Regression In Machine Learning*. [online] Available at: <https://medium.com/@chughkashish12/decision-tree-regression-in-machine-learning-4f117158cdcc> [Accessed 9 June 2020].
- Medium. 2020. *A Brief Introduction To Supervised Learning*. [online] Available at: <https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590> [Accessed 9 June 2020].
- Medium. 2020. *Random Forest And Its Implementation*. [online] Available at: <https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f> [Accessed 9 June 2020].
- Medium. 2020. *Cross-Validation In Machine Learning*. [online] Available at: <https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f> [Accessed 9 June 2020].
- Copper.org. (2019). [online] Available at: https://www.copper.org/publications/pub_list/pdf/introduction-to-copper-and-copper-al loys.pdf [Accessed 25 Oct. 2019].
- Pandey, A., 2020. *Machine Learning Algorithms And Reverse Modeling*. Canberra: Anaconda Jupytr.

Appendix A

For all the finalised codes, explanation of codes and the read me file for our script , please use this link

https://drive.google.com/drive/folders/1ScYsEOkM-J0k3ZdOjOvDMYvwvxOCO3vR

As it is complex to present code in the appendices, provided our repository link, which consists of the readme file, codes used to generate the output.

Appendix B

-

Appendix C

Hyperparameter tuning using Randomized search for random forests.

Code:

```
from sklearn.model_selection import RandomizedSearchCV

# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 10, stop = 2000, num = 20)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4, 6]
# Method of selecting samples for training each tree
bootstrap = [True, False]
# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
print(random_grid)

# First create the base model to tune
regressor = RandomForestRegressor()
# Random search of parameters, using 3 fold cross validation,
# search across 100 different combinations, and use all available cores
regressor_random = RandomizedSearchCV(estimator = regressor, param_distributions = random_grid, n_iter = 100, cv = 3, verbose=2, random_state=42, n_jobs = -1)
# Fit the random search model
regressor_random.fit(X, y)

L=regressor_random.best_params_
```

The code is easily understandable with the comments, the final row 79 “L” gives us the output of the best parameters to be used for the random forests model.

The image belows shows the random forests regression syntax, where the values in the brackets of randomforestregressor(row 2) shows us the hyperparameters that used in my random forests model. These were the hyperparameters that were generated after performing hyperparameter tuning method which is shown in the code above.

```
from sklearn.ensemble import RandomForestRegressor
regressor = RandomForestRegressor(n_estimators = 533, min_samples_split = 2, min_samples_leaf = 1, max_features = "sqrt", max_depth= 110, bootstrap = False, random_state = 0)
regressor.fit(X, np.ravel(y))
y_pred = regressor.predict(X)
```

Appendix D

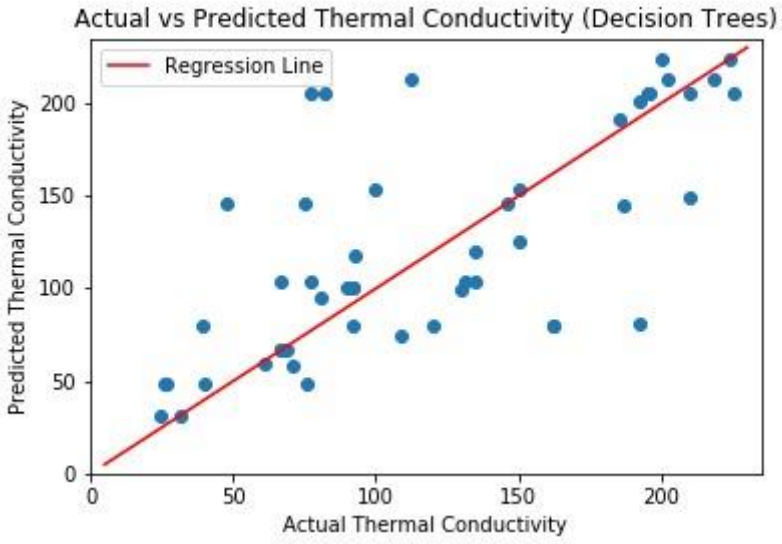

Fig: 17

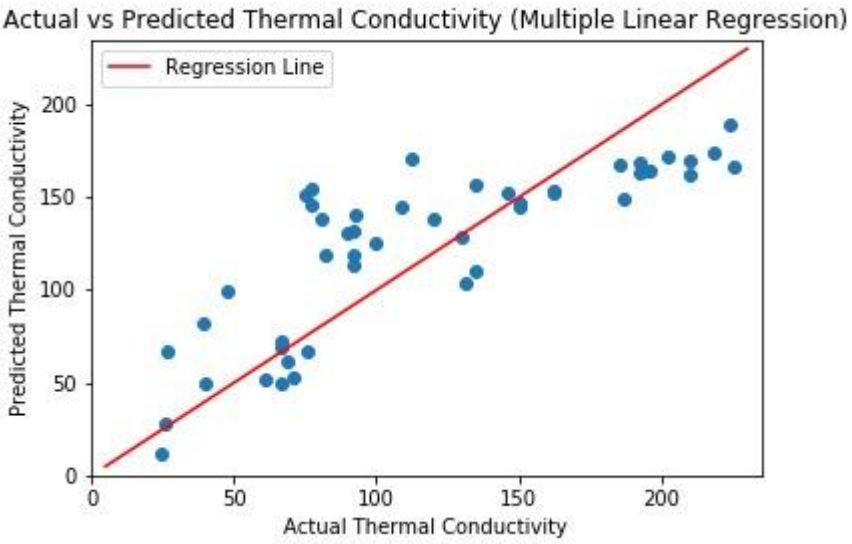

Fig: 18

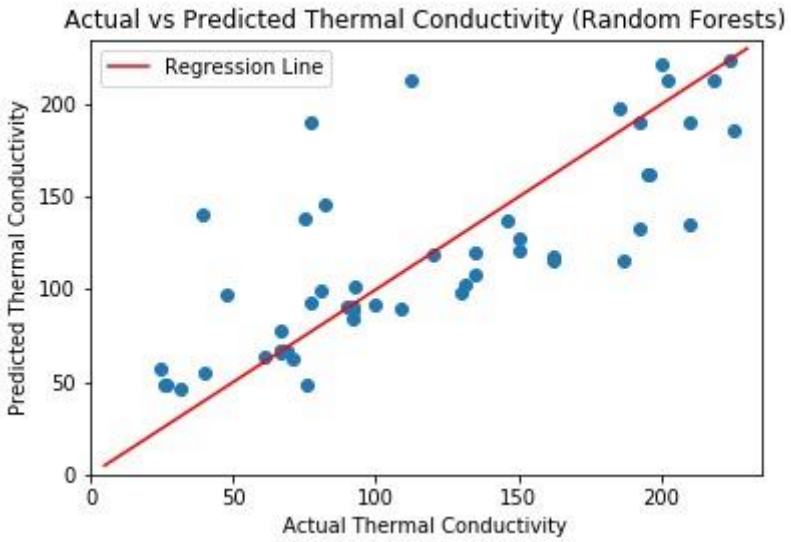
Actual vs Predicted Thermal Conductivity (Random Forests)
Regression Line
Predicted Thermal Conductivity
Actual Thermal Conductivity
0
50
100
150
200

Fig: 19

D