IDENTIFYING HIGH PERFORMANCE STEEL ALLOYS BY PREDICTING MECHANICAL PROPERTIES USING MACHINE LEARNING

AJIT PONNAPPAN APPAN
G2302892L

Ajit P.

Introduction

Materials discovery of metal alloys with exceptional mechanical properties is a difficult task. The creation of these alloys involves using exotic metals and high temperatures and pressures. Due to these difficult processing conditions, the scope for trial and error in high performance alloy development is limited. The search space for alloys is also vast, for an alloy consisting of just 5 elements there are 160 x 10^6 possible alloy combinations.[1]

In recent years, due to the availability of large experimental and computational datasets, improved algorithms, and the growth of computer processing power, machine learning has emerged as an efficient tool for reducing the search space and identifying candidate alloys that may have exceptional mechanical properties like high hardness and high strength. This project uses **regression** models to predict **material properties** from their corresponding compositions and **classification** models to identify groups of high strength alloy **compositions**.

Methodology

Machine learning models are first trained on the dataset of actual elemental compositions and properties. A search space of possible combinations of various elements and compositions present in the theoretical alloy is then created. Regression models are used to predict the Yield Strength of alloy compositions present in the search space and new compositions with the highest predicted values of strength are identified. Classification models use a labels of 'low strength' or 'high strength' to classify the compositions present in the search space. A CSV of 'high strength' compositions is written for further experimental evaluation.

Datasets used

The dataset[2] consists of more than 900 compositions of steel alloy used in welding applications taken from the book Metallurgy of Basic Weld Metal by G. M. Evans and N. Bailey.[3] Each Steel composition includes individual ranges of 12 (C, Si, Mn, P, S, Cu, Ni, Cr, Mo, V, Ti, Al) alloy element additions to Fe as columns. The material properties present are Yield strength 'YS' and Ultimate tensile strength 'UTS'. Values for %Elongation and %Reduction in Area as well as the temperature at which the tensile test is performed are also present. Based on Yield strength, the alloys are classified as 'low strength' (<750 N/mm^2) and 'high strength' (>750 N/mm^2). The dataset is present in the CSV file named as "SteelDB2.csv".

The maximum and minimum values of yield strength and ultimate strength are shown below.

	Max	Min
Yield Strength (YS) (N/mm^2)	1026	310
Ultimate Tensile Strength (UTS) (N/mm^2)	1123	345

The maximum and minimum compositions of each of the elements present in the dataset is shown below.

Alloying Element	Max (%)	Min (%)
Carbon	0.152	0.035
Manganese	2.1	0.23
Silicon	1.11	0
Sulfur	0.046	0.003
Phosphorus	0.04	0.003
Aluminium	0.068	0.0001
Titanium	0.077	0.0001
Chromium	3.5	0.03
Nickel	5.48	0.03
Molybdenum	1.16	0.005
Vanadium	2.873	0.003
Copper	2.04	0.02

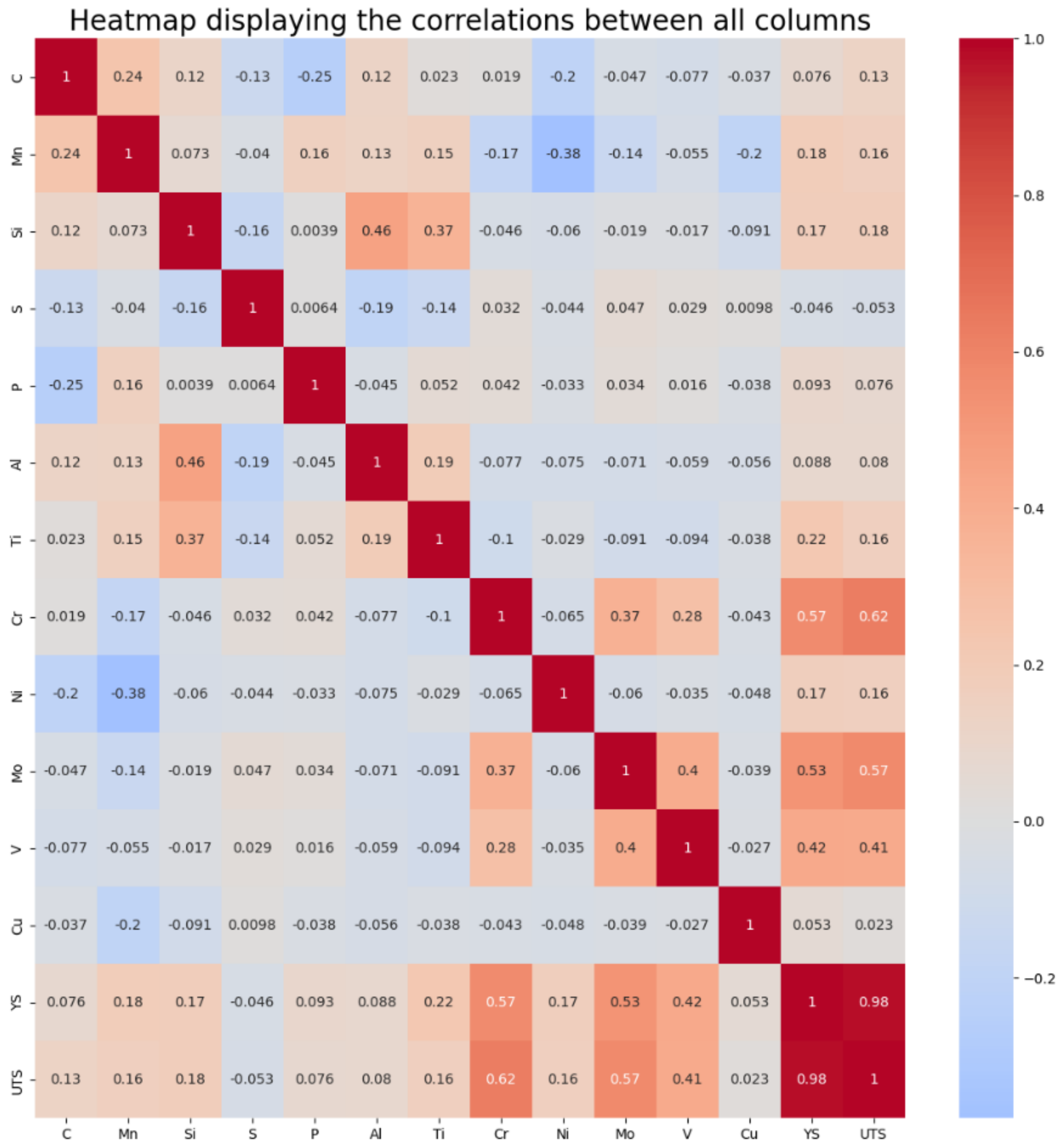
The composition having maximum Yield strength in the dataset is shown below

C	Mn	Si	S	P	Al	Ti	Cr	Ni
0.077	1.06	0.39	0.006	0.008	0.0005	0.0039	2.91	0.03

Mo	V	Cu	YS	UTS	% Elongation	RA	Temp	Strength Classification
1.14	2.817	0.03	1026	1114	17.3	60.1	36	High

We see that few elements (Mn, Cr, Mo, V) have values close to the maximum composition in the dataset while the rest of the elements have values close to their minimum composition. This distribution is used when creating the search space of theoretical alloys.

The heatmap below highlights correlations between the columns of the dataset



Heatmap displaying the correlations between all columns

	C	Mn	Si	S	P	Al	Ti	Cr	Ni	Mo	V	Cu	YS	UTS
C	1	0.24	0.12	-0.13	-0.25	0.12	0.023	0.019	-0.2	-0.047	-0.077	-0.037	0.076	0.13
Mn	0.24	1	0.073	-0.04	0.16	0.13	0.15	-0.17	-0.38	-0.14	-0.055	-0.2	0.18	0.16
Si	0.12	0.073	1	-0.16	0.0039	0.46	0.37	-0.046	-0.06	-0.019	-0.017	-0.091	0.17	0.18
S	-0.13	-0.04	-0.16	1	0.0064	-0.19	-0.14	0.032	-0.044	0.047	0.029	0.0098	-0.046	-0.053
P	-0.25	0.16	0.0039	0.0064	1	-0.045	0.052	0.042	-0.033	0.034	0.016	-0.038	0.093	0.076
Al	0.12	0.13	0.46	-0.19	-0.045	1	0.19	-0.077	-0.075	-0.071	-0.059	-0.056	0.088	0.08
Ti	0.023	0.15	0.37	-0.14	0.052	0.19	1	-0.1	-0.029	-0.091	-0.094	-0.038	0.22	0.16
Cr	0.019	-0.17	-0.046	0.032	0.042	-0.077	-0.1	1	-0.065	0.37	0.28	-0.043	0.57	0.62
Ni	-0.2	-0.38	-0.06	-0.044	-0.033	-0.075	-0.029	-0.065	1	-0.06	-0.035	-0.048	0.17	0.16
Mo	-0.047	-0.14	-0.019	0.047	0.034	-0.071	-0.091	0.37	-0.06	1	0.4	-0.039	0.53	0.57
V	-0.077	-0.055	-0.017	0.029	0.016	-0.059	-0.094	0.28	-0.035	0.4	1	-0.027	0.42	0.41
Cu	-0.037	-0.2	-0.091	0.0098	-0.038	-0.056	-0.038	-0.043	-0.048	-0.039	-0.027	1	0.053	0.023
YS	0.076	0.18	0.17	-0.046	0.093	0.088	0.22	0.57	0.17	0.53	0.42	0.053	1	0.98
UTS	0.13	0.16	0.18	-0.053	0.076	0.08	0.16	0.62	0.16	0.57	0.41	0.023	0.98	1

We see that Elements like Chromium, Molybdenum and Vanadium have a strong positive correlation with Yield Strength and Ultimate Tensile Strength while Sulfur has a negative correlation with Strength.

Yield strength and Ultimate Tensile strength are very strongly correlated so we focus on just Yield Strength as the property used for training and prediction.

Machine learning models used

- **Linear regression**
  Linear models consist of sums of parameters or weights multiplied with the independent variables. As seen in the correlation map, it may be possible to represent Yield Strength as a sum of weights multiplied with the compositions of each element. While linear regression is a simple model, as seen in Ref [5] simple regression methods perform better than complex or non-linear models when predicting higher values of material properties than what is present in the training data. As we are looking for alloys with exceptional Yield strength, linear regression is a useful model.

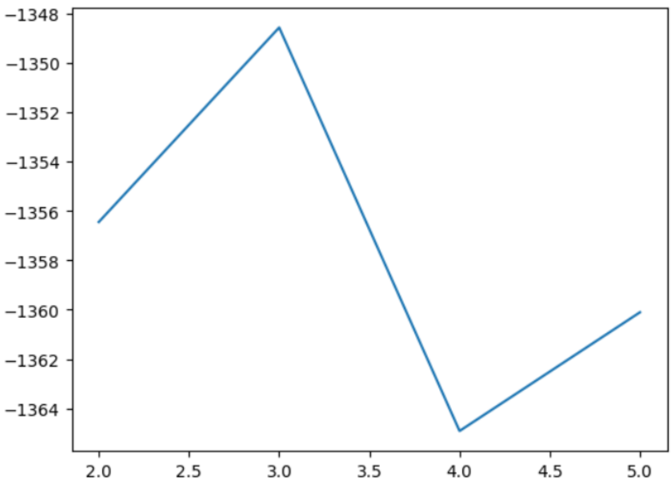
$$A_{r3}(^{\circ}C) = 735.6 + 180.1(C + Cr) + 1206.9(S + P) - 10.9(Si + Mn + Ni + Cu + Mo) + 755.3(Al + N) - 328.8(V + Nb + Ti)$$

(Example of linear equation to find phase transformation temperature from Ref [3])

- **Random Forest regressor**
  Random forest regressor is a decision tree based model that takes the average of multiple tree predictions to predict values. It shows high accuracy and resistance to overfitting.

- **K-Nearest Neighbours regressor**
  Humans approaching the task of finding new alloy materials will naturally identify extraordinary materials based on composition and operate under the assumption that similar chemistries might have similar properties. K-Nearest neighbours model uses a similar approach by assigning properties to alloy compositions based on similar neighbouring compositions.

  To identify the optimal number of neighbours to use in the algorithm parameter tuning is performed with GridSearchCV along with Kfolds cross validation. Grid Search is used with different values of neighbours from 2 to 6 and Kfold split of the training data into 5 groups is performed for scoring. The optimal number of neighbours is found to be 3.



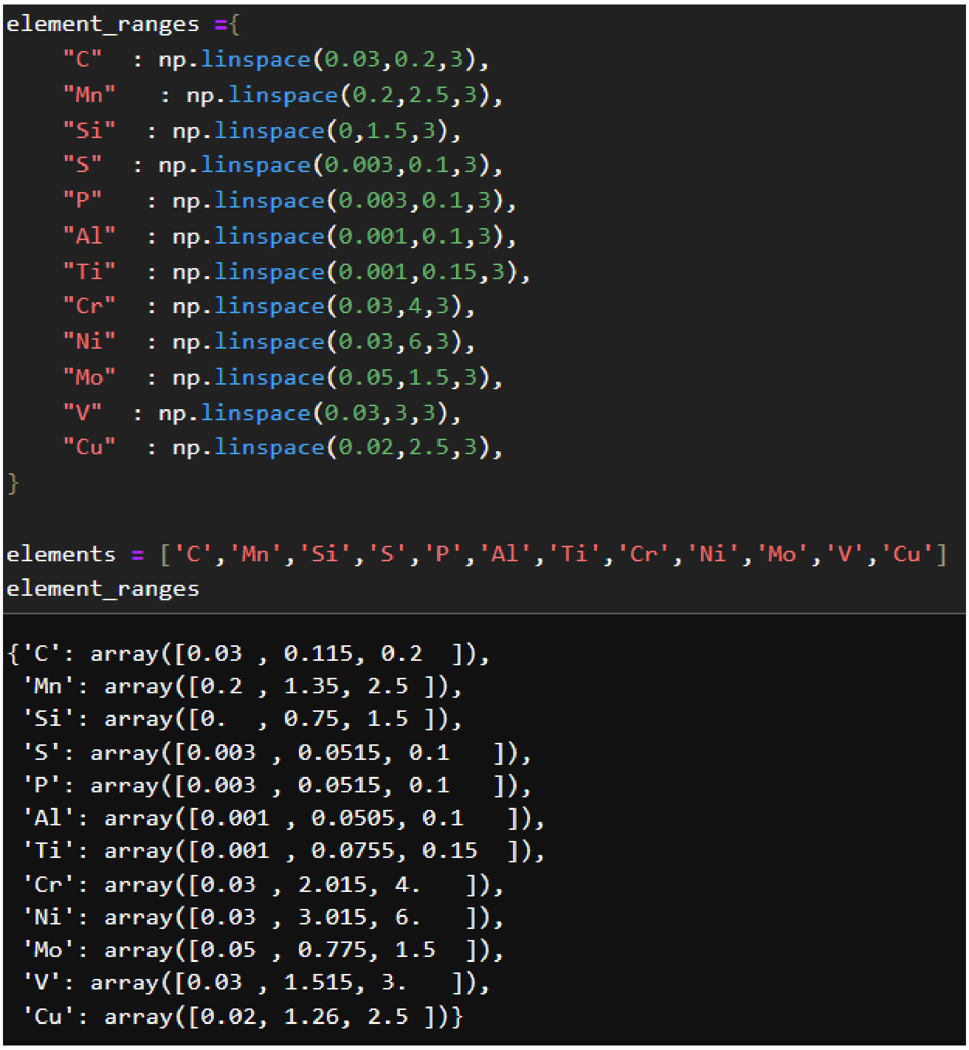
(Plot of mean test score from GridsearchCV vs number of neighbours)

It is seen that n_neighbors = 3 gives the least negative score value

- **Support Vector Classifier**
  Classification models cannot predict properties but as shown in Ref [5] classification models show consistently higher precision and equivalent recall compared to regression models when identifying groups of alloys with exceptional properties. A Support Vector Classifier is used with a radial basis function (rbf) kernel. It performs nonlinear mapping of an input vector into a high-dimensional feature space and then constructs an optimal hyperplane for separating the features discovered.

Search Space

A search space of theoretical alloys is created for the purpose of finding new alloys with exceptional properties. To create this search space a dictionary of element compositions as numpy arrays of 3 compositions is created based on the maximum and minimum values present in the original dataset. The theoretical maximum composition is taken to be slightly higher than the maximum present in the dataset (eg. Chromium has a maximum composition of 3.5% in the dataset but the search space uses a maximum of 4%)

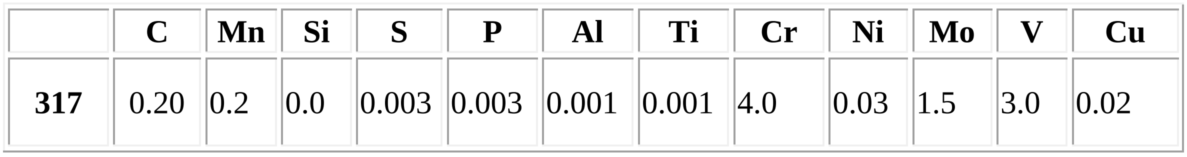


```
element_ranges ={
    "C"   : np.linspace(0.03,0.2,3),
    "Mn"    : np.linspace(0.2,2.5,3),
    "Si"   : np.linspace(0,1.5,3),
    "S"   : np.linspace(0.003,0.1,3),
    "P"    : np.linspace(0.003,0.1,3),
    "Al"   : np.linspace(0.001,0.1,3),
    "Ti"   : np.linspace(0.001,0.15,3),
    "Cr"   : np.linspace(0.03,4,3),
    "Ni"   : np.linspace(0.03,6,3),
    "Mo"   : np.linspace(0.05,1.5,3),
    "V"   : np.linspace(0.03,3,3),
    "Cu"   : np.linspace(0.02,2.5,3),
}

elements = ['C','Mn','Si','S','P','Al','Ti','Cr','Ni','Mo','V','Cu']
element_ranges
```

```
{'C': array([0.03 , 0.115, 0.2  ]),
 'Mn': array([0.2 , 1.35, 2.5 ]),
 'Si': array([0.  , 0.75, 1.5 ]),
 'S': array([0.003 , 0.0515, 0.1   ]),
 'P': array([0.003 , 0.0515, 0.1   ]),
 'Al': array([0.001 , 0.0505, 0.1   ]),
 'Ti': array([0.001 , 0.0755, 0.15  ]),
 'Cr': array([0.03 , 2.015, 4.   ]),
 'Ni': array([0.03 , 3.015, 6.   ]),
 'Mo': array([0.05 , 0.775, 1.5  ]),
 'V': array([0.03 , 1.515, 3.   ]),
 'Cu': array([0.02, 1.26, 2.5 ])}
```

When creating a theoretical alloy composition, 4 element combinations are chosen from a list of all elements using itertools.combinations function. The composition of these 4 elements uses the medium and maximum compositions present in the array (x[1] and x[2]) while the rest of the elements use the minimum composition in the array (x[0]). For example in the theoretical composition below, C, Cr, Mo and V have their maximum % values while the rest of the elements have their minimum values.



	C	Mn	Si	S	P	Al	Ti	Cr	Ni	Mo	V	Cu
317	0.20	0.2	0.0	0.003	0.003	0.001	0.001	4.0	0.03	1.5	3.0	0.02

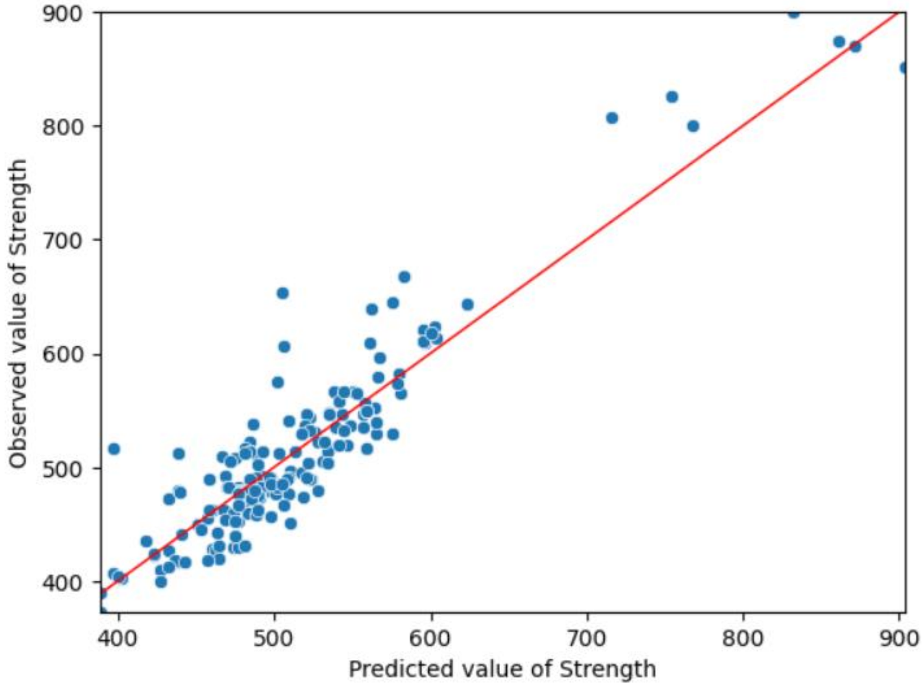
Performance of models

I. TRAIN-TEST

- **Linear Regression**

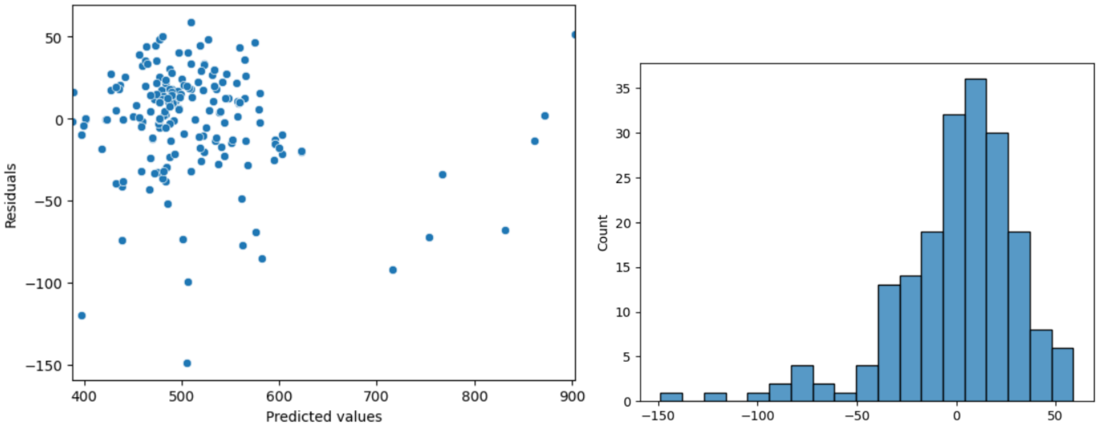
R^2 score: 0.867950267675643
Mean square error: 1000.4074178785373

Actual vs predicted test values:



Most of the points fall close to the red line and on either side of the line.

Residuals:

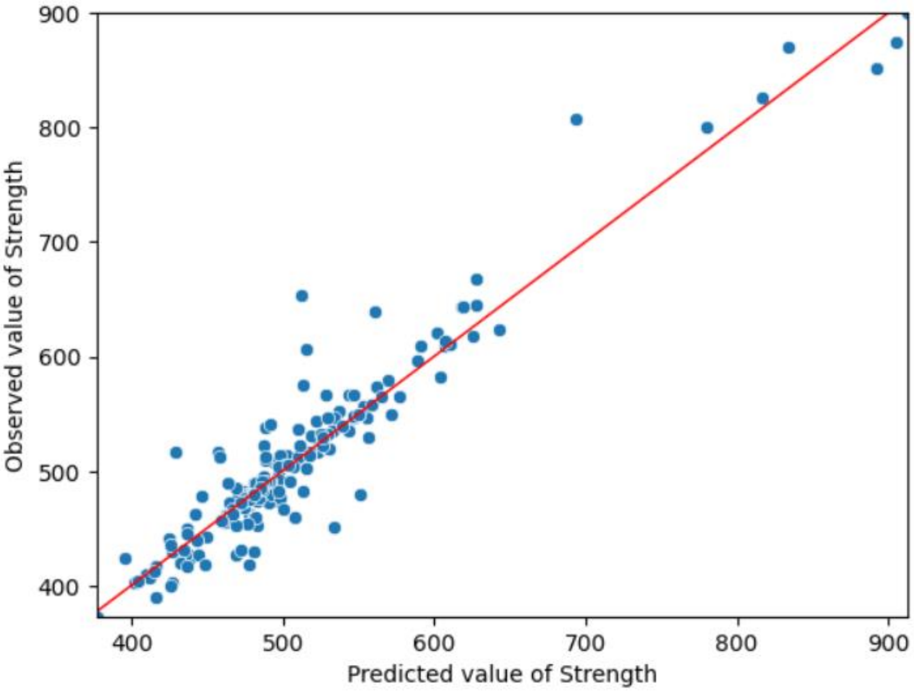


Residual = predicted – actual test value
Residuals are randomly distributed and shows a normal distribution around 0

- **Random Forest Regression**

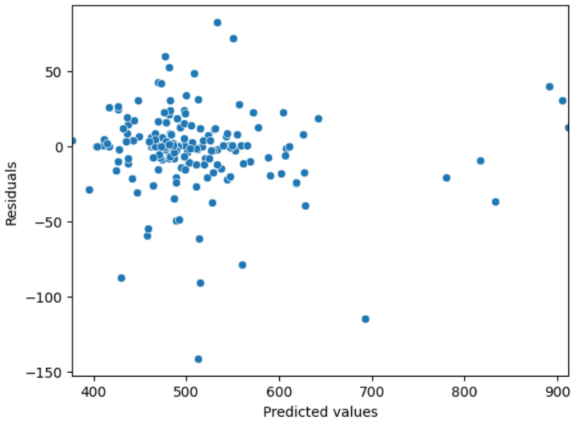
R^2 score: 0.9080216466474634
Mean square error: 696.827061732394

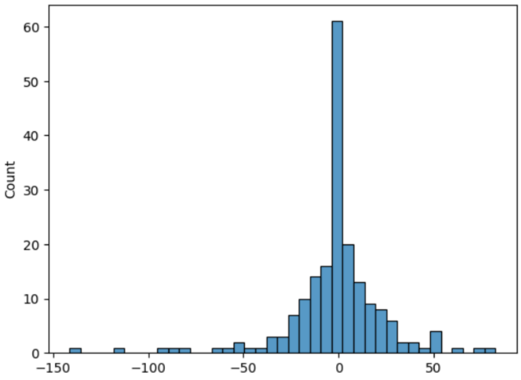
Actual vs predicted test values:



Most of the points fall close to the red line and on either side of the line.

Residuals:



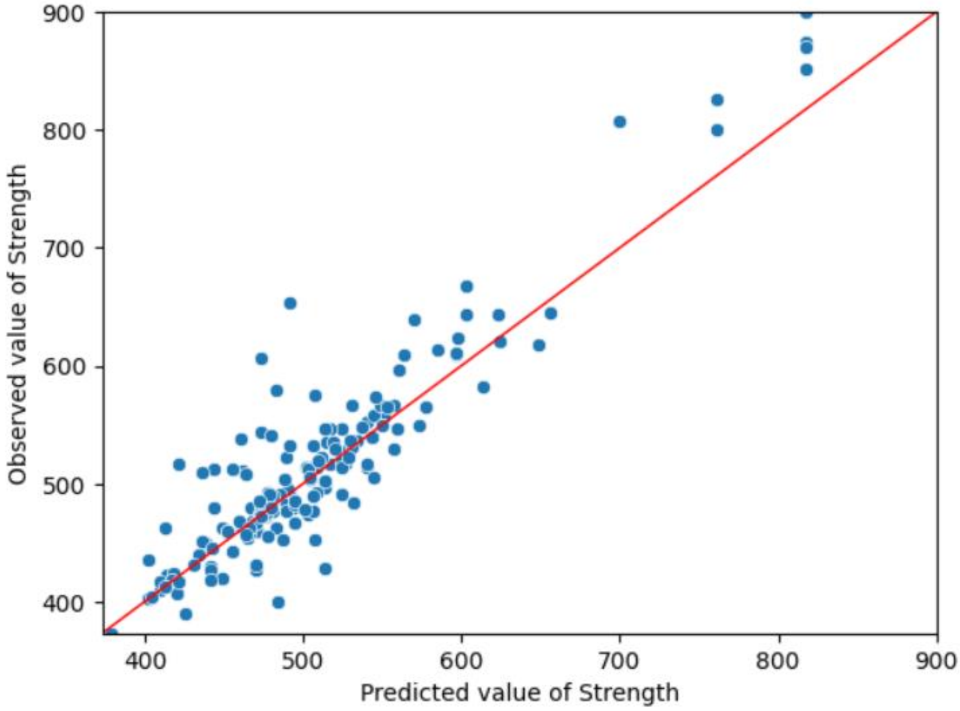


Residual = predicted – actual test value
Residuals are randomly distributed and shows a normal distribution around 0

- **K Nearest Neighbours Regression**

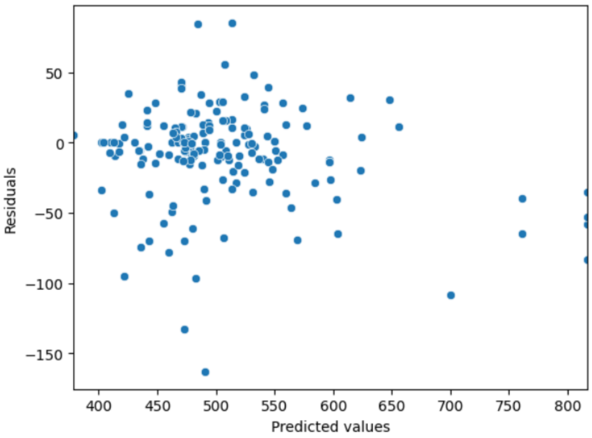
  R^2 score: 0.8559251614746025
  Mean square error: 1091.509499136442

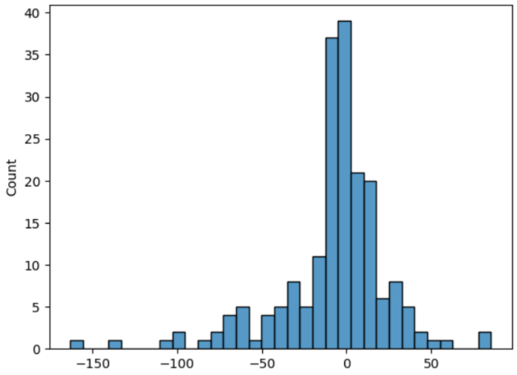
  Actual vs predicted test values:



It is seen that for higher values, KNN consistently predicts a lower value than the actual strength.

Residuals:



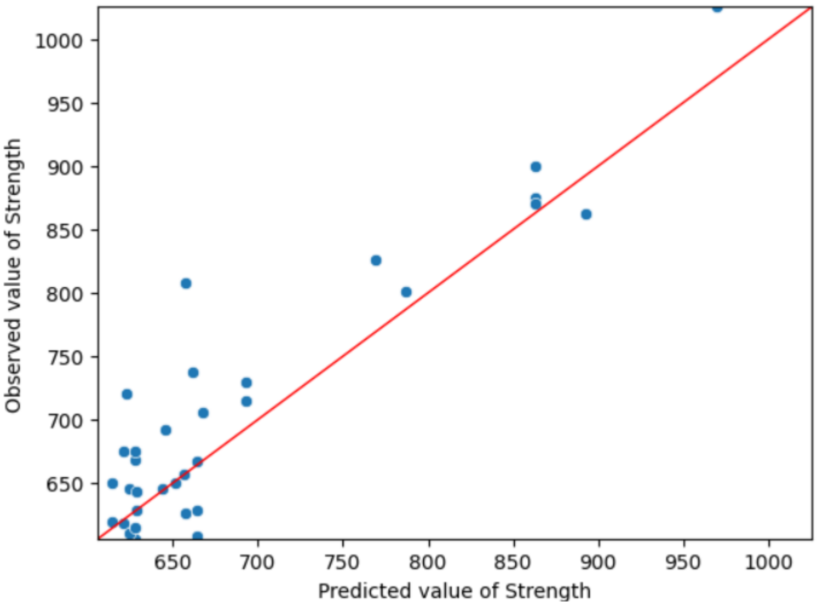


Residual = predicted – actual test value
Residuals are randomly distributed and shows a normal distribution around 0. However there is a small bias in the distribution towards the negative residuals.

- **K Nearest Neighbours Regression (limited dataset)**

  A limited dataset with only compositions having Yield strength greater than 600 N/mm^2 is also used so that the model gives a larger predicted yield strength. However lower R^2 score and larger Mean square error are seen.

  R^2 score: 0.8073332016168258
  Mean square error: 2052.4006734006743

  The problem of lower predicted values than observed is still seen.

- **Support Vector Classifier**

  Confusion Matrix

	Predicted High	Predicted Low
True High	4	1
True Low	0	284

  1 high strength composition was predicted as low strength.

	Precision	Recall	F1-score
High	1	0.8	0.89
Low	1	1	1

II. SEARCH SPACE PREDICTION

A search space 989 theoretical compositions is fed into the trained models. The maximum value of Yield strength is noted for each model used.

- **Linear Regression**
  Max Yield strength: 1007.288 N/mm^2
  Composition:

	C	Mn	Si	S	P	Al	Ti	Cr	Ni	Mo	V	Cu
553	0.03	2.5	0.0	0.003	0.003	0.001	0.001	4.0	6.0	0.05	3.0	0.02

- **Random Forest Regression**
  Max Yield strength: 684.25 N/mm^2
  Composition:

	C	Mn	Si	S	P	Al	Ti	Cr	Ni	Mo	V	Cu
351	0.03	2.5	1.5	0.003	0.1	0.001	0.001	4.0	0.03	0.05	0.03	0.02

- **K Nearest Neighbours Regression**
  Max Yield strength: 682.667 N/mm^2
  Composition:

	C	Mn	Si	S	P	Al	Ti	Cr	Ni	Mo	V	Cu
387	0.03	2.5	1.5	0.003	0.003	0.001	0.001	4.0	0.03	0.05	3.0	0.02

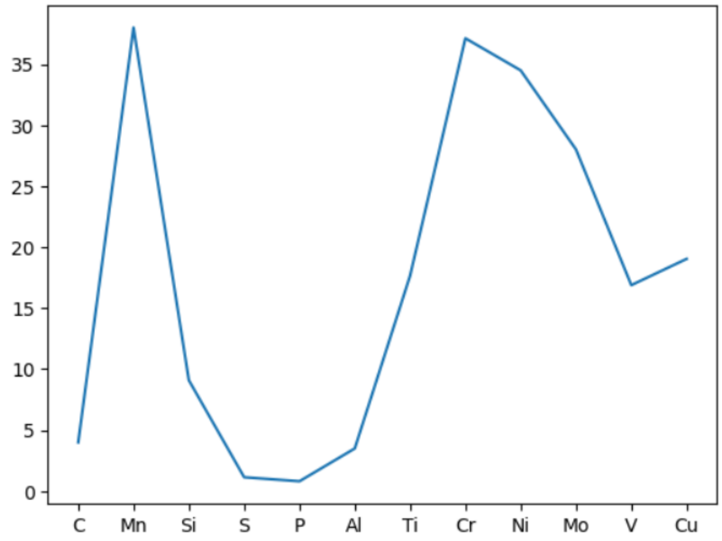
- **K Nearest Neighbours Regression (limited dataset)**
  Max Yield strength: 969.667 N/mm^2
  Composition:

	C	Mn	Si	S	P	Al	Ti	Cr	Ni	Mo	V	Cu
75	0.2	2.5	0.0	0.003	0.003	0.001	0.001	4.0	0.03	0.05	3.0	0.02

- **Support Vector Classifier**
  170 high strength (>750 N/mm^2) compositions are identified.

Discussion

Among the regression methods, Random Forest regressor shows the best performance from the R^2 and Mean Square Error scores. However, the value of Yield Strength predicted by it as also relatively low. Random Forest Regressor takes the average of values predicted by many decision trees. As it takes the average, it is not capable of predicting values that are at the extremes of the dataset and hence it is not a useful model for predicting high performance alloys.



Regression Coefficients obtained from Linear regression model show that Manganese, Chromium, Molybdenum and Nickel compositions have the greatest effect on strength.

Linear regression predicts the highest value of Yield Strength among the models used. The predicted composition has maximum values for Mn, Cr, Ni and V. The predicted value is also relatively close to the value of maximum yield strength in the original dataset. Similar results are seen in [5] where a ridge regressor is sued.

KNN regressor predicts a max yield strength value close to that of Random forest. The KNN algorithm takes the average of only 3 neighbouring compositions however the predicted value is still relatively small. Performance of KNN regressor is also seen to be the worst of the 3 models.

KNN regressor trained on a limited dataset predicts a much higher value for Max Yield Strength. The composition selected by model trained on both full and limited datasets have max % values for Mn, Cr and V. The fourth element with higher percentage is Silicon for the full dataset and Carbon for the limited dataset.

One of the High strength alloys was misclassified as low strength by the Support Vector Classifier. However no low strength alloys were classified as high strength. Therefore the 170 compositions identified may be accurately classified as high strength compositions.

Conclusion

Both classification and regression models can be used to identify high strength steel alloy compositions.

Linear regression is the best model among the regression models tested as it has good performance and predicts a high value for the desired property of Yield Strength.

Random Forest and K Nearest Neighbour regressor models have difficulty predicting property values at extremes of the dataset. They are less useful for identifying High performance materials.

3 different regression models suggest that 2.5% Manganese, 4% Chromium and 3% Vanadium should be used in the high strength steel alloy.

References

1. Mohanty, T., Chandran, K. S., & Sparks, T. D. (2023). Machine learning guided optimal composition selection of niobium alloys for high temperature applications. *APL Machine Learning*, *1*(3).
2. Evans, G. M. 2015. Database — Weld metal composition and properties. DOI: 10.13140/RG.2.1.3628.1764
3. Evans, G. M., and Bailey, N. 1999. Metallurgy of Basic Weld Metal. Abington, UK: Abington Publishing.
4. Salganik, V. M., Chikishev, D. N., Pozhidaeva, E. B., and Nabat-chikov, D. G. 2016. Analysis of structural and phase transformations in low-alloy steels based on dilatometric studies. Metallurgist 59(9):766–773. DOI: 10.1007/s11015-016-0172-3
5. Varadarajan, R., & Sampath, K. (2023). Application of Machine Learning to Regression Analysis of a Large SMA Weld Metal Database. *Welding Journal*, *102*(3), 31S-52S.
6. Kauwe, S. K., Graser, J., Murdock, R., & Sparks, T. D. (2020). Can machine learning find extraordinary materials?. *Computational Materials Science*, *174*, 109498.