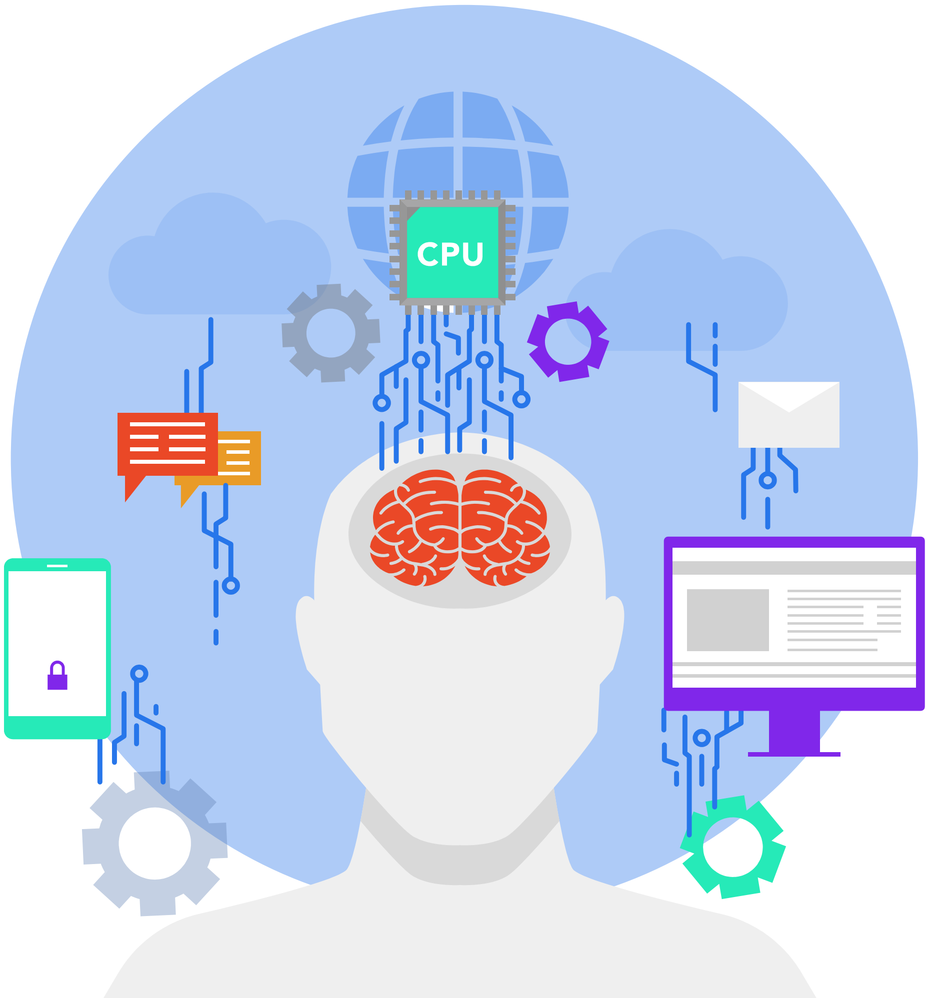


MSE643

Property Prediction and Model Comparison

Team Members

Harsh Kumar Singh
Mrigank Kumar
Abhisekh Prakash
Sudhanshu
Kamal Raj Rawat
Manish Kumar

Instructor: Prof. Krishanu Biswas

Project Work Flow

Collected Steel alloy Composition data

Prepared the data

Data Cleaning, Data Splitting And Data Transformation

Training of models

SV, DT, KNN(10), Ridge regression, Random Forest, Linear regression and ANN models

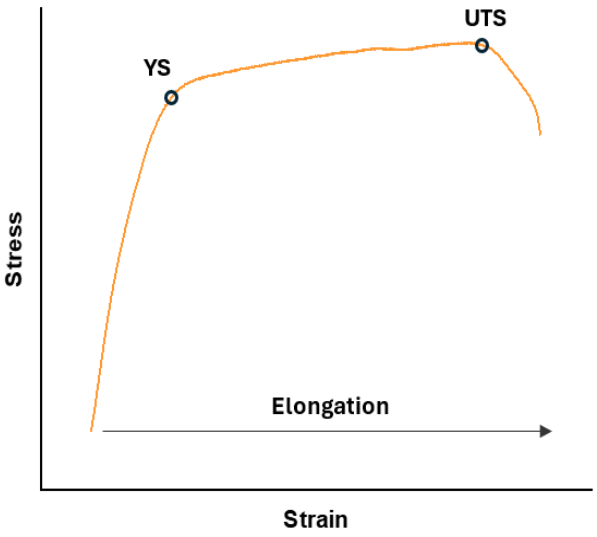
Evaluating models

Comparing their performance based on error analysis

Conclusion

Comparison with ANN

Problem Definition and Objectives



Broad Objectives

Prediction

01 Yield strength, Ultimate tensile Strength and Elongation of Different composition of steel alloys

Comparison

02 Finding the best model based on errors.

About Data and Data Preperation

01 Data

We have the initial Data of steel strength, we have 17 columns , of all elements such as Fe, Mg, Cr, etc content as a feature and then yield strength ,tensile strength and elongation as labels.

02 Data pre-processing

Removing the first column which contained the Formula of Steel

03 Standardization

Converting data into uniform format using the Minimum and Maximum formula given b below

$$X' = \frac{X - Xmin}{Xmax - Xmin}$$

04 Missing data

We had 9 Null data points for elongation, so we filled them with average Values from all the data points we have

```
RangeIndex: 312 entries, 0 to 311
Data columns (total 17 columns):
 #   Column            Non-Null Count
---  ------            --------------
 0   Fe                312 non-null
 1   c                 312 non-null
 2   mn                312 non-null
 3   si                312 non-null
 4   cr                312 non-null
 5   ni                312 non-null
 6   mo                312 non-null
 7   v                 312 non-null
 8   n                 312 non-null
 9   nb                312 non-null
 10  co                312 non-null
 11  w                 312 non-null
 12  al                312 non-null
 13  ti                312 non-null
 14  yield strength    312 non-null
 15  tensile strength  312 non-null
 16  elongation        303 non-null
dtypes: float64(17)
memory usage: 41.6 KB
```

Data preparation continues

01 Train

[(224, 14), (224, 3)]

Remaining = (280 - 560 = 224

02 Validate

[(56, 14), (56, 3)]

20% of (312-32) ≈ 56

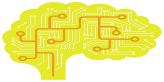
03 Test

[(32, 14), (32, 3)]

10% of 312 = 31.2 ≈ 32

```
X_new,X_test,y_new,y_test = train_test_split(X_scaled,y_scaled,test_size=0.1,random_state=1);
X_train,X_valid,y_train,y_valid = train_test_split(X_new,y_new,test_size=0.2,random_state=1)
```

Methods used for comparison



01 R2 Score

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

02 RMSE

$$\sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \hat{x}_i\right)^2}{N}}$$

03 MAE

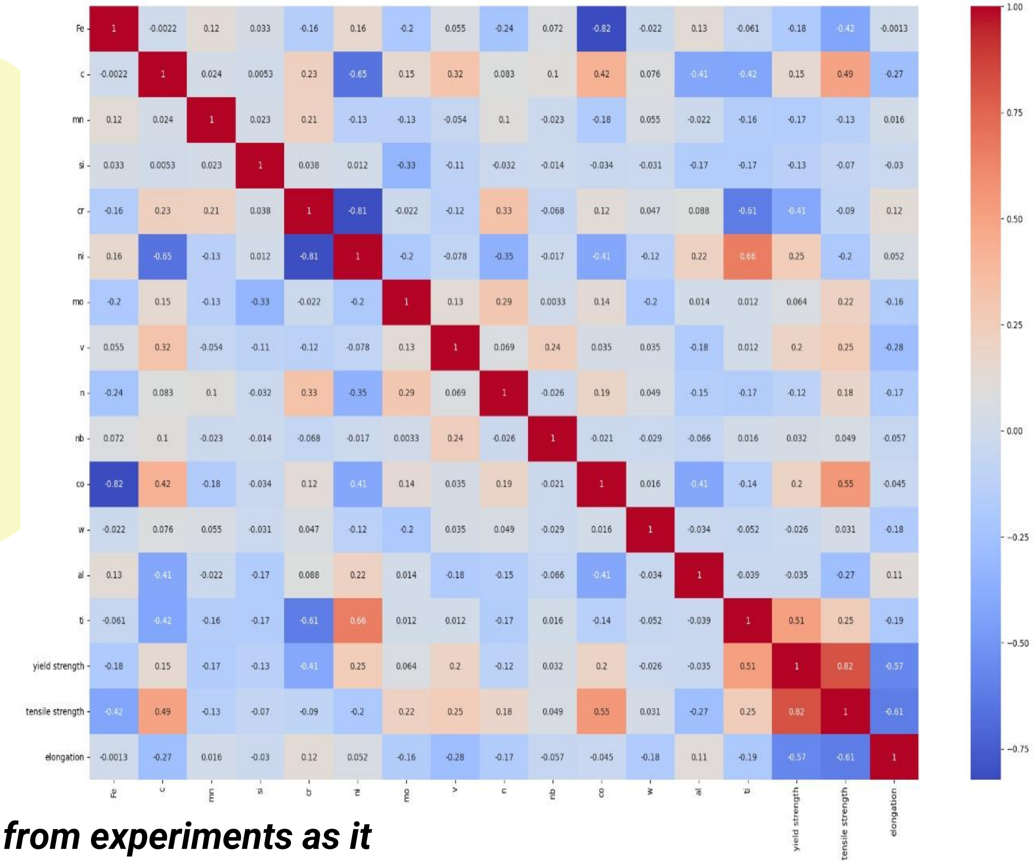
$$\frac{\sum_{i=1}^{n}\left|y_i - x_i\right|}{n}$$

Co-relation data

Observation

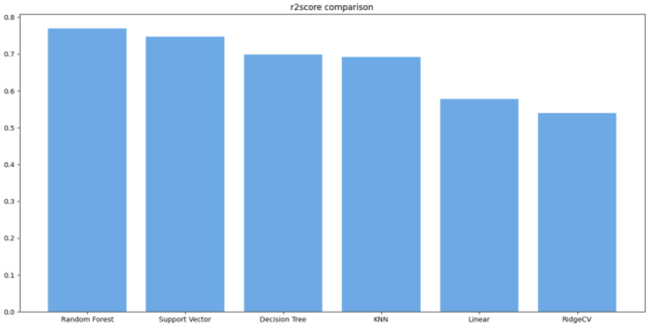
- Carbon correlation on Elongation (-0.27) indicating increased brittleness
- Tensile Strength and Yield Strength are strongly correlated (0.82).
- Elongation is negatively correlated with both Yield Strength (-0.57) and Tensile Strength (-0.61) — indicating a trade-off between strength and ductility.

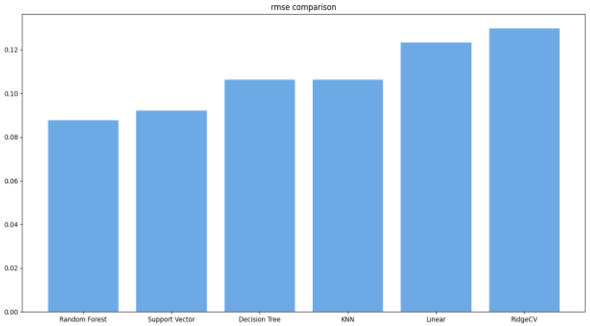
$$\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

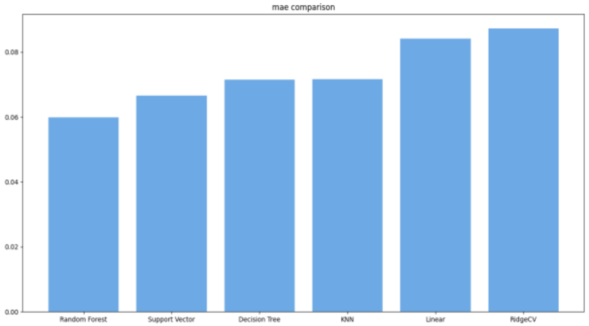


This reflects that our data are correctly taken from experiments as it matches with the theoretical correlations and studies.

Error charts

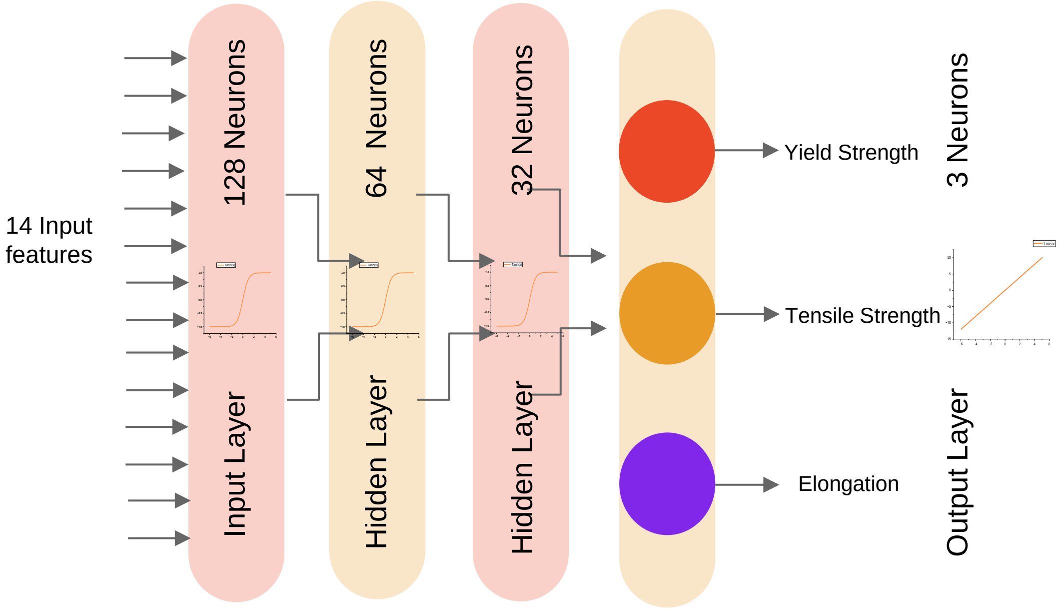






- **Random Forest** consistently performs the best with the **lowest RMSE (~0.088)** and **lowest MAE (~0.06)**, indicating high prediction accuracy and minimal average error.
- **Support Vector** follows closely, with a slightly higher RMSE (~0.092) and MAE (~0.067), still performing better than most other models.
- **Decision Tree and KNN** models show comparable error levels with **RMSE around 0.107** and **MAE ~0.073**, indicating moderate performance.
- **Linear and RidgeCV** models perform the worst, with **RidgeCV having the highest RMSE (~0.13)** and **MAE (~0.088)**, suggesting they are less suited for this regression task.

Artificial Neural Network



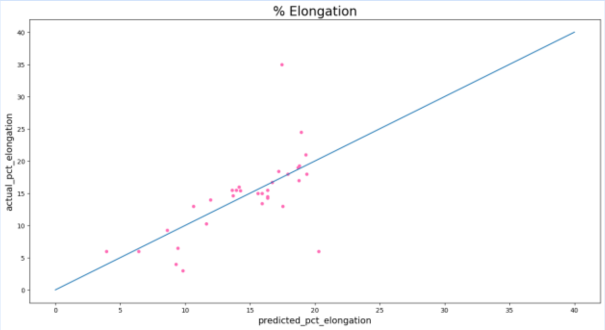
Activation Function used

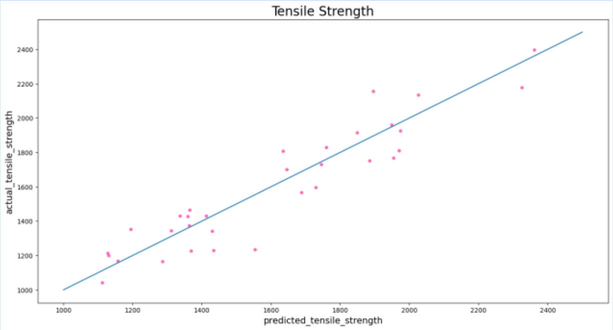
$$Tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

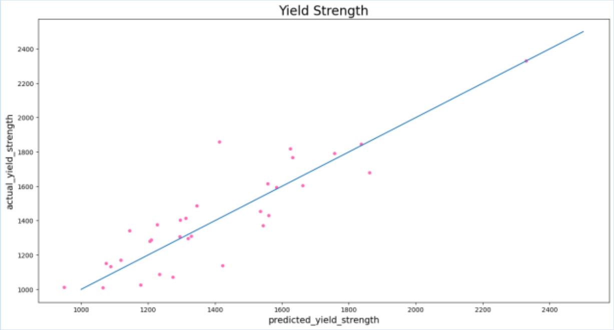
EPOCH used 500
Batch size- 128

```
history = model.fit(X_train, y_train, batch_size = 128, shuffle=True, epochs=500)
```

Comparision with all models and conclusion







	R2 Score	RMSE	MAE
Random Forest	.769904	0.087720	0.0986
ANN	0.7790	0.0823	0.0566

From the analysis ML and DL models(ANN),we found out that ANN would be the best model for prediction of YS, UTS and Elongation through the composition of 14 elements in the steels