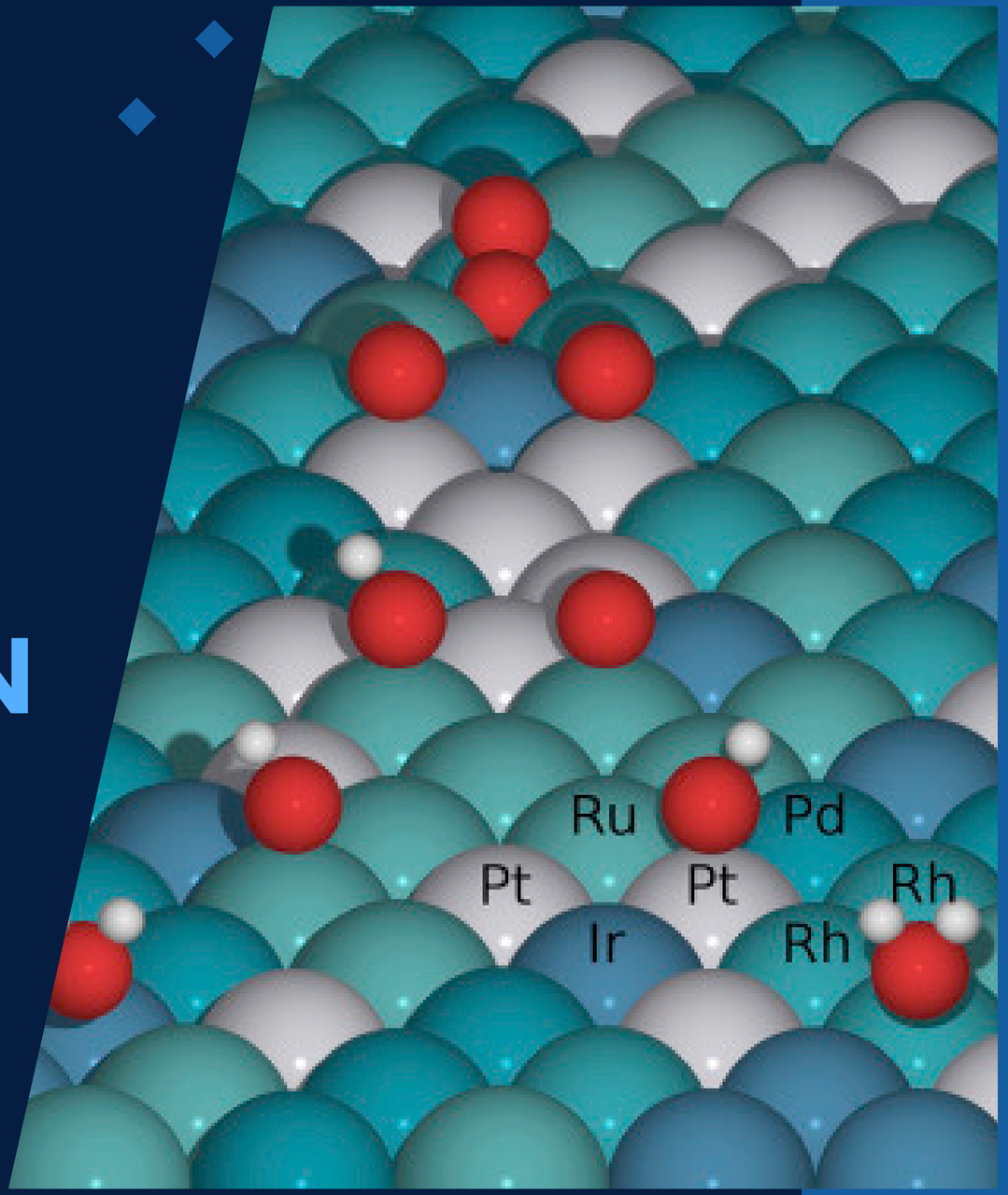
Department of
Mechanical Engineering

Indian Institute of Technology Bombay

ME793

PROPERTY PREDICTION OF HEAs USING ML

Presented by: Archit Mundada [22B2259]
Mokshit Naidu [200100104]
Shiv Modi [19D100011]



OBJECTIVES AND DELIVERABLES

1	2	3	4
Utilizing multiple regressor models to predict various properties of HEAs like Yield strength, Melting temperature, Bulk Modulus, etc.	**Obtaining correlations among various input variables and arriving at a refined choice for the selection that affects the output variables the most**	**Testing other algorithms like SVM[2], GP[3], KNN[4], NN[5] to predict properties and explain the performance difference**	**If time avails, creating a model that predicts possible HEA compositions for a user defined property.**

Our final deliverable will be a program that will utilize regressor models and neural networks, incorporating multiple features that will be able to:

1. Predict parameters such as Yield Strength, UTS, and Melting Point to name a few
2. Compare against other algorithms for testing
3. Create possible HEA compositions and combinations for any pre-defined property

DATA CLEANING

NaN values

We found a dataset of ~1100 entries with features like Melting Temp., D_VEC, ΔH, ΔS, etc. We filtered cases of NaN/empty values.

Data Types

There was a discrepancy in numerical datatypes- INT or FLOAT. We converted all numbers to FLOATs.

Encoding

Two features had categorical data- S_Phase and Phase comprising of FCC, BCC, IM, AM, etc.- we use one-hot encoding to split data into multiple features with binary values.

Outliers

After encoding, we removed the features that had low incidence in the data so that our output won't get skewed

Normalize

We normalized the features to enhance model performance by ensuring all features contribute properly, stabilizing training and improving interpretability.

DATA SOURCE: https://zenodo.org/records/5155150

Final data with 1103 entries and 105 properties

We are also looking into another dataset:

https://www.nature.com/articles/s41524-022-00779-7 (in supplementary material)

200100104

Mokshit Naidu

DATA CLEANING

	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA
1	IMsp	IMp	AMsp	AMp	SSsp	FCCp	B2p	BCCp	HCPp	2BCCp	SSp	L12p	2FCCp
2	1	1	0	0	0	0	0	0	0	0	0	0	0
3	1	1	0	0	0	0	0	0	0	0	0	0	0
4	1	1	0	0	0	0	0	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0	0	0	0	0	0
6	1	1	0	0	0	0	0	0	0	0	0	0	0
7	1	1	0	0	0	0	0	0	0	0	0	0	0
8	1	1	0	0	0	0	0	0	0	0	0	0	0
9	1	1	0	0	0	0	0	0	0	0	0	0	0
10	1	1	0	0	0	0	0	0	0	0	0	0	0
11	1	1	0	0	0	0	0	0	0	0	0	0	0
12	1	1	0	0	0	0	0	0	0	0	0	0	0
13	1	1	0	0	0	0	0	0	0	0	0	0	0
14	1	1	0	0	0	0	0	0	0	0	0	0	0
15	1	1	0	0	0	0	0	0	0	0	0	0	0
16	1	1	0	0	0	0	0	0	0	0	0	0	0
17	1	1	0	0	0	0	0	0	0	0	0	0	0

One-hot encoded data (after removal of outliers)

	C	D
1	**S_Phase**	**Phase**
2	IM	IM
3	IM	IM
4	IM	IM
5	IM	IM
6	IM	IM
7	IM	IM
8	IM	IM
9	IM	IM
10	IM	IM
11	IM	IM
12	IM	IM
13	IM	IM
14	IM	IM
15	IM	IM

Original data with phase columns

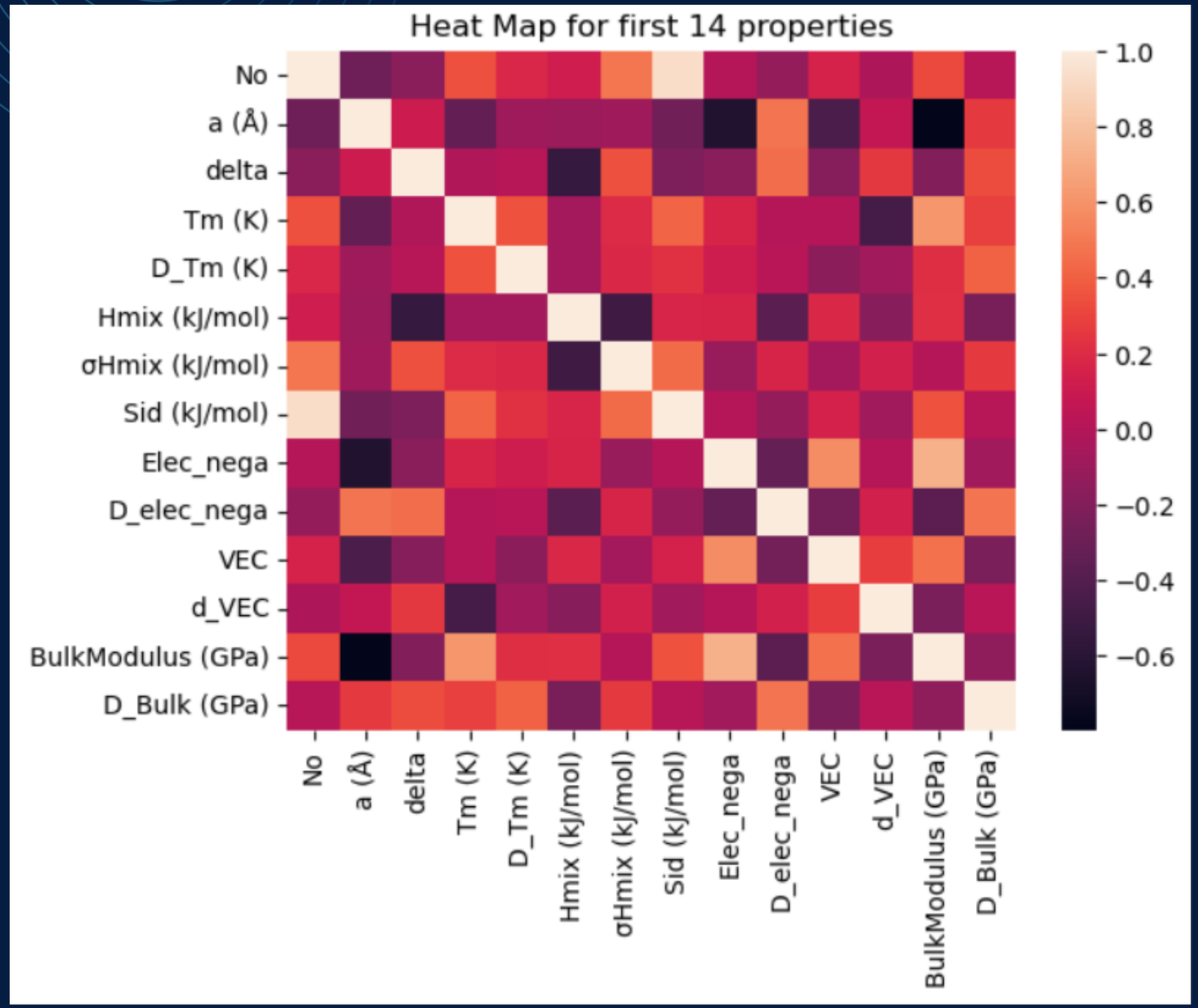
1033	ZrHfTiAlCuNi	6	SS+IM	SS+IM
1034	Al0.5B0.2CoCrCuFeNi	7		
1035	Al0.5B0.6CoCrCuFeNi	7		
1036	Al0.5BCoCrCuFeNi	7		
1037	AlCoCrCuFeMnNi	7	SS+IM	FCC+BCC+IM
1038	AlCoCrCuFeMo0.2Ni	7	SS	BCC+FCC
1039	AlCoCrCuFeMo0.4Ni	7		
1040	AlCoCrCuFeMo0.6Ni	7		
1041	AlCoCrCuFeMo0.8Ni	7		
1042	AlCoCrCuFeMoNi	7		
1043	Al0.2CoCrCu0.8FeNiSi0.2	7	IM	IM

Empty data entries- rows like these were removed

200100104
Mokshit Naidu

Data Analysis

- The first analysis that we did with the data was with heat maps- to find correlation among variables
- Our target properties (or Y) with the current dataset are ‘Bulk_Modulus’ and ‘D_Bulk’.
- Since there are 105 total properties, displaying a heat map for all of them is not useful- it is not interpretable.
- Thus, we have shown the heat map for the first 14 properties, last two of which are our target properties.
- It is clear that Bulk Modulus has a high correlation with the atomic radius of atoms.
- Clearly, some features are more important compared to others and we will see this in our upcoming analyses.



22B2259

Archit Mundada

Data Analysis

- Next up, we calculated the F-values and P-Values of all (103) features with respect to Bulk Modulus and D Bulk.
- These values were in good agreement with the results obtained using the previous method (Pearson Correlation coefficient)
- The result was that there are about 50 key properties correlated with Bulk Modulus, and about 30 key properties correlated with D_Bulk.
- Rest of the properties had P values more than 10^-3
- The bottom values commmonly corresponded to elements

	features	f_values	p_values
0	a (Å)	1885.752281	7.714943e-241
1	Elec_nega	1235.266947	4.480027e-182
2	Tm (K)	651.885366	2.586338e-113
3	VEC	305.140318	1.680523e-60
4	SSsp	243.778620	8.544190e-50

	features	f_values	p_values
0	D_elec_nega	320.565804	4.061945e-63
1	D_Tm (K)	216.025241	8.687131e-45
2	Zr	199.002622	1.158671e-41
3	FCCp	187.965045	1.296955e-39
4	delta	133.417913	3.272282e-29

	features	f_values	p_values
93	K	0.0	1.0
94	Ne	0.0	1.0
95	Ns	0.0	1.0
96	O	0.0	1.0
97	Os	0.0	1.0
98	Rb	0.0	1.0
99	S	0.0	1.0
100	Se	0.0	1.0
101	Te	0.0	1.0
102	Tl	0.0	1.0

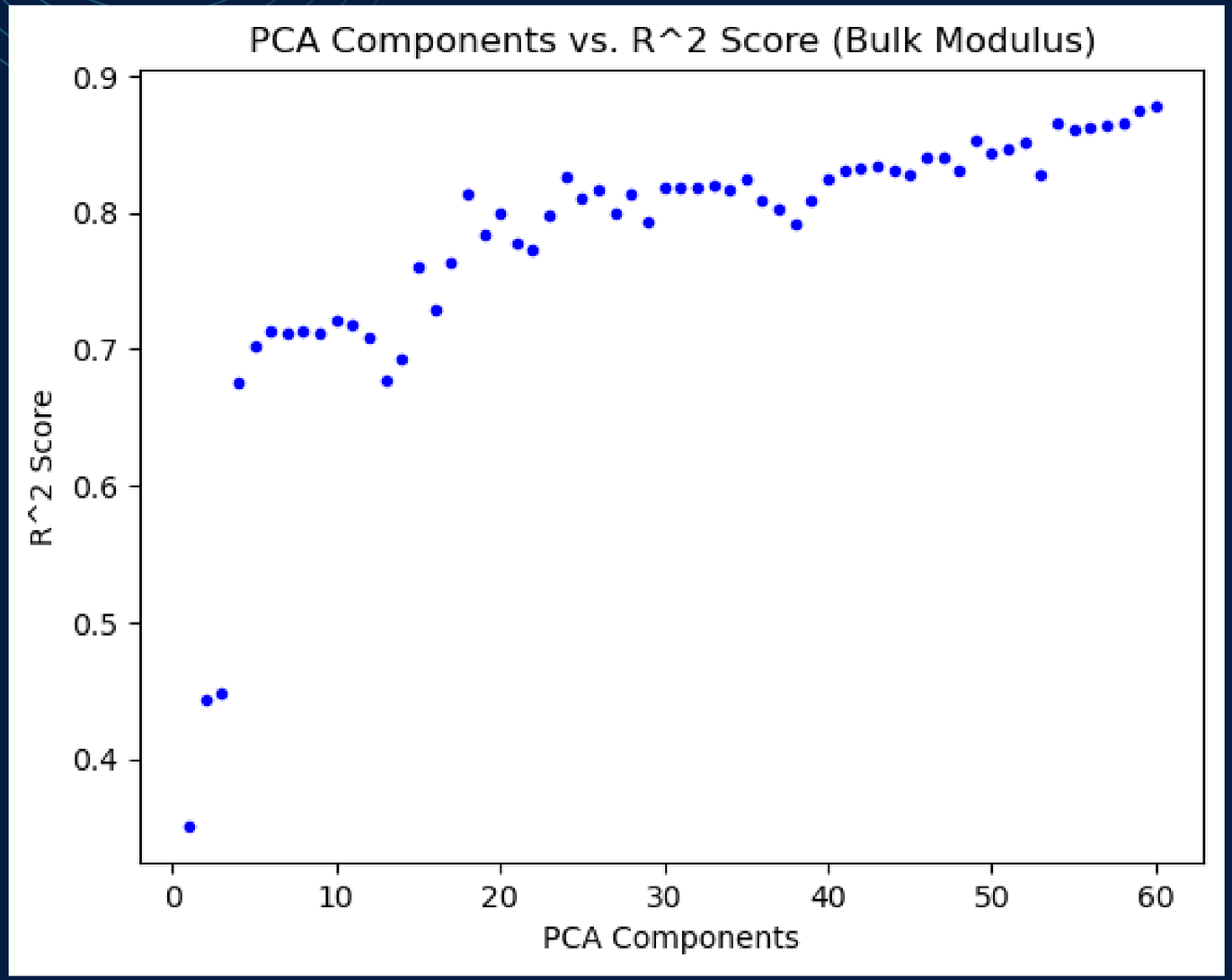
	features	f_values	p_values
93	K	0.0	1.0
94	Ne	0.0	1.0
95	Ns	0.0	1.0
96	O	0.0	1.0
97	Os	0.0	1.0
98	Rb	0.0	1.0
99	S	0.0	1.0
100	Se	0.0	1.0
101	Te	0.0	1.0
102	Tl	0.0	1.0

Top 5 and Bottom 10 values for Bulk Modulus (left) D_Bulk (right)

19D100011
Shiv Modi

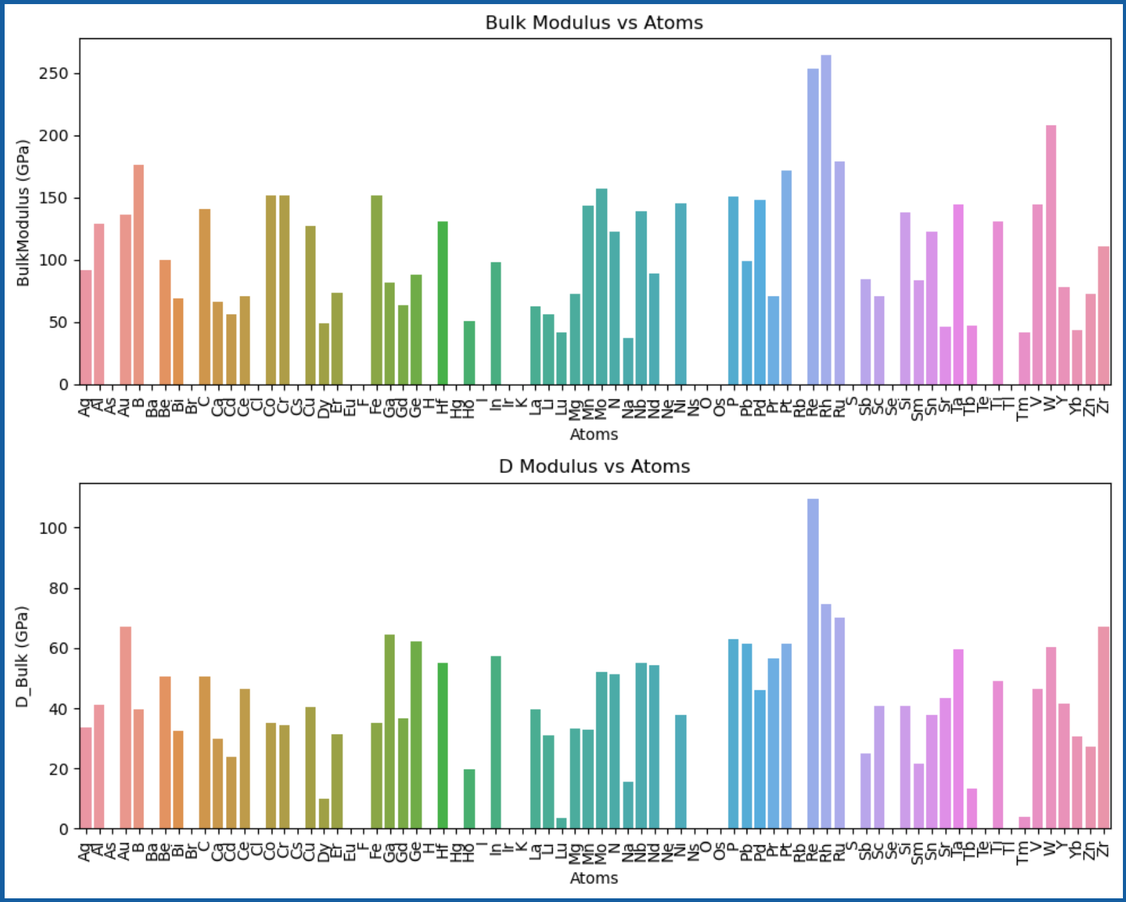
Data Analysis

- We also did the Principal Component Analysis of the data, and fit the linear regressor for Bulk Modulus on the principal components.
- We observed that the R^2 score did not change much after about 20 principal components, suggesting again that there are few main features that we need to target.
- There is some variability in the output of PCA depending on the seed of the randomizer.
- As we will see later, a R^2 score of 0.8 is close to the best R^2 score of 0.9 that we obtained using all the features



22B2259
Archit Mundada

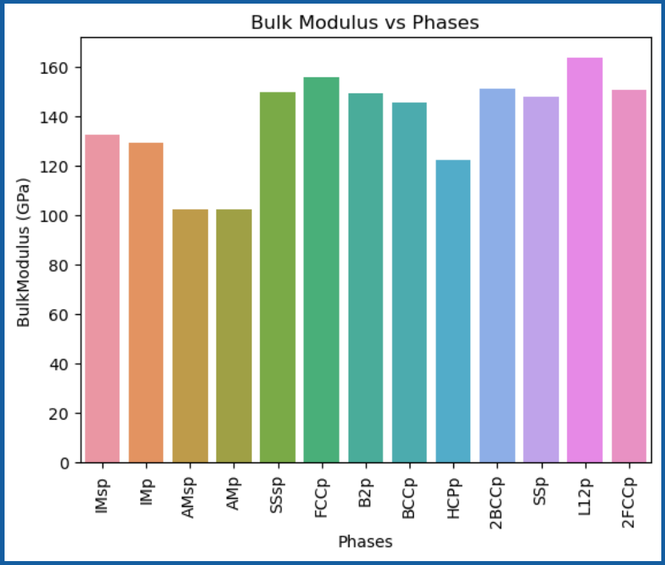
PLOTS (EDA)

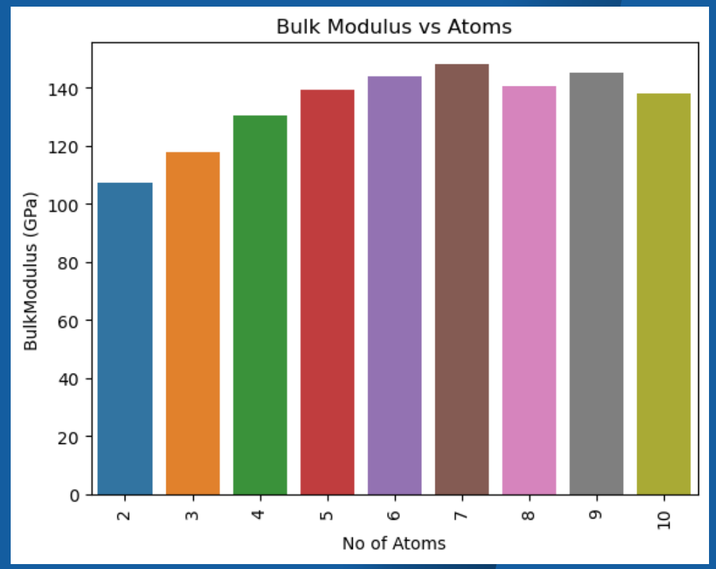


- In the coming slides, we have plotted various trends in the data
- On the left, there are two bar plots- the top plot displays the average bulk modulus of HEAs that have a particular element, say Ag present in them
- Similarly, the bottom plot corresponds to the D_bulk parameter.
- The gaps in the plot are because of presence of no HEAs of the given elements, and this has been taken care of while cleaning the data
- This plot gives a clear direction about which elements to choose for a particular range of bulk modulus

200100104
Mokshit Naidu

PLOTS (EDA)



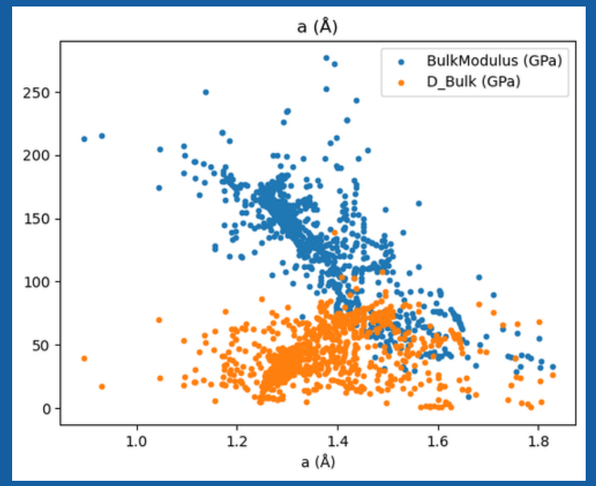


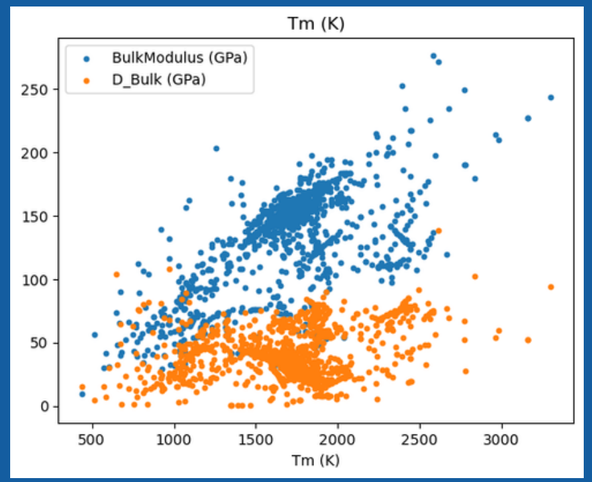
- Next, we have plotted average Bulk Modulus for a given phase or number of atoms in the alloy
- These plots again give us a direction on our choices for a certain range of the property value
- Since our final goal is to create a ‘HEA designer’ model, this data indicates that a tree-like model might be a good choice for it.

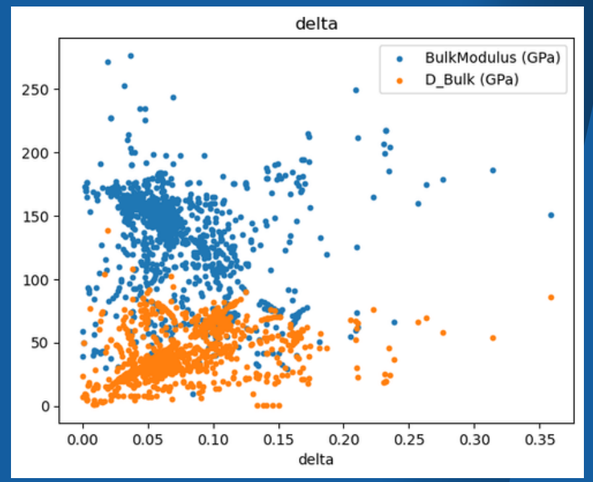
22B2259
Archit Mundada

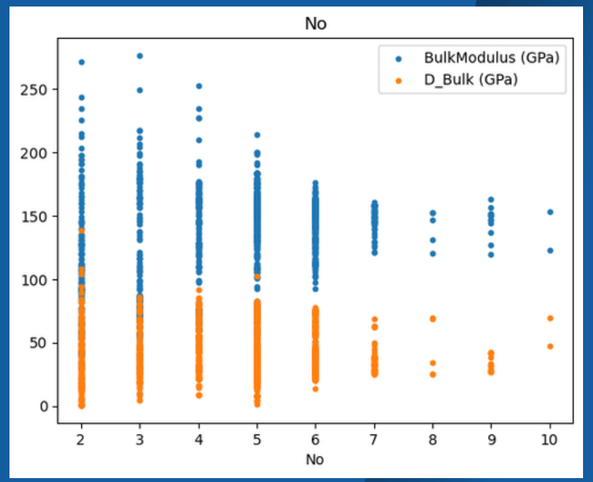
PLOTS (EDA)

- These plots indicate the trends for the Bulk Modulus and its standard deviation with other property variables.
- Often, we can observe linear relationships for Bulk Modulus , but the D_Bulk (standard deviation) parameter often displays non-linear behavior
- Thus, it is logical to use a multilinear regressor for the Bulk Modulus model. We don’t expect equally good results for D_bulk.



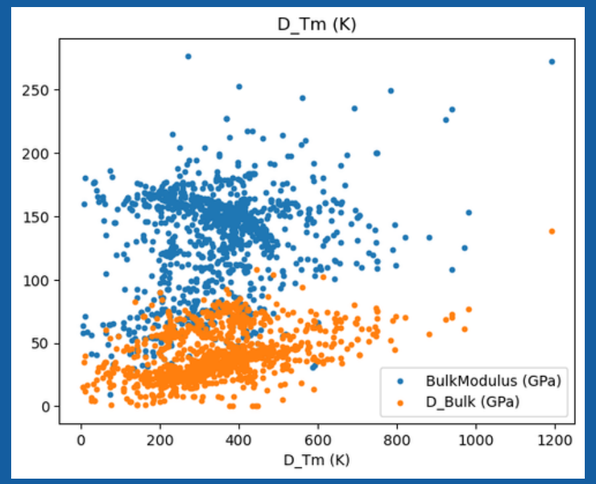


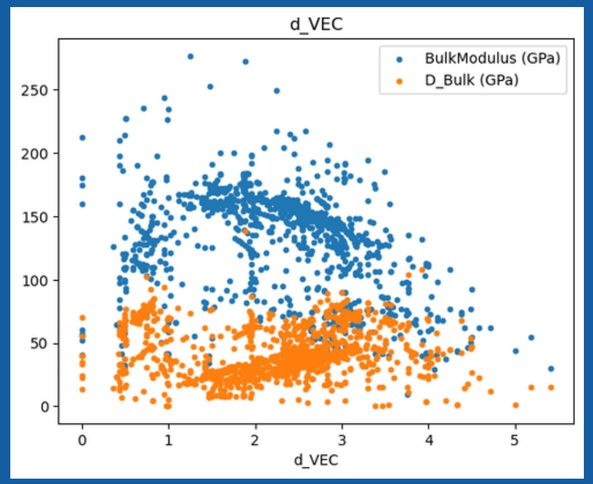


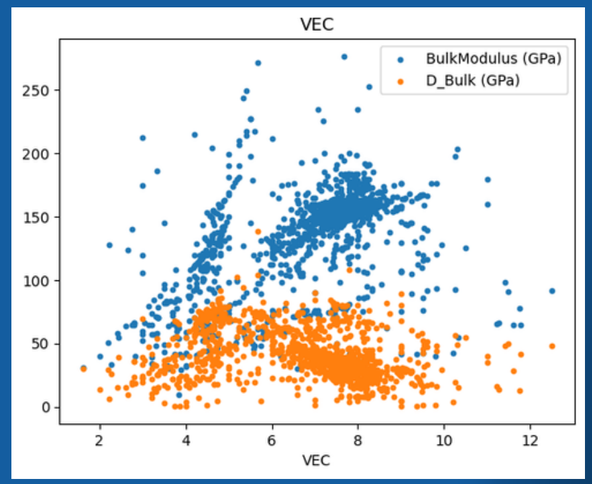


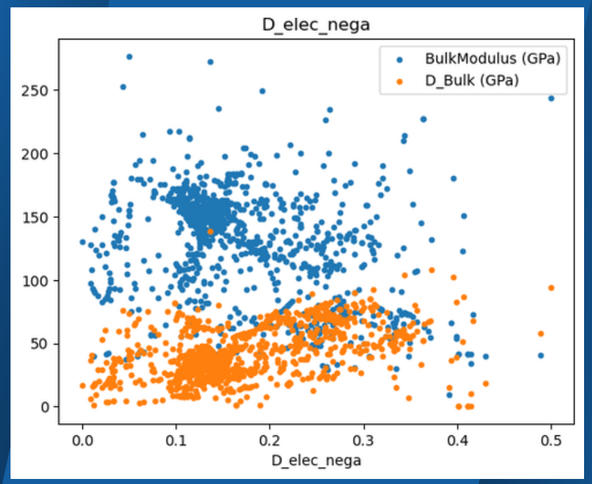
19D100011
Shiv Modi

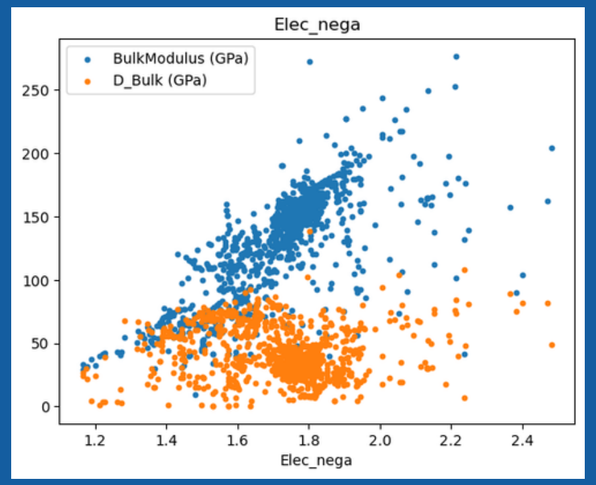
PLOTS (EDA)

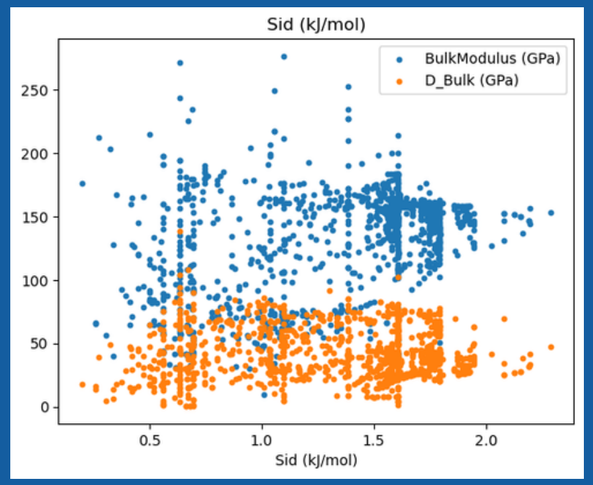


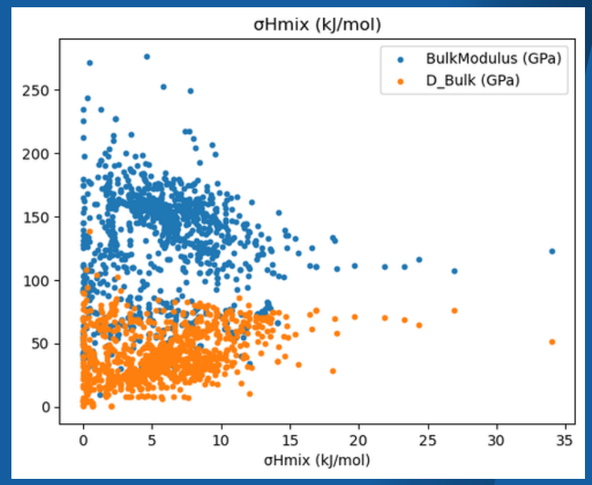


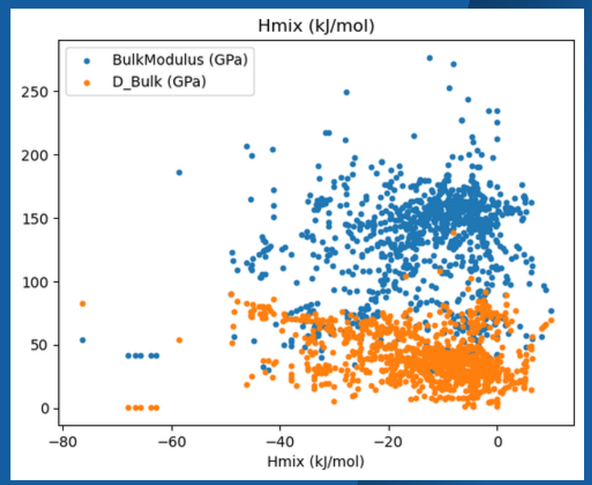












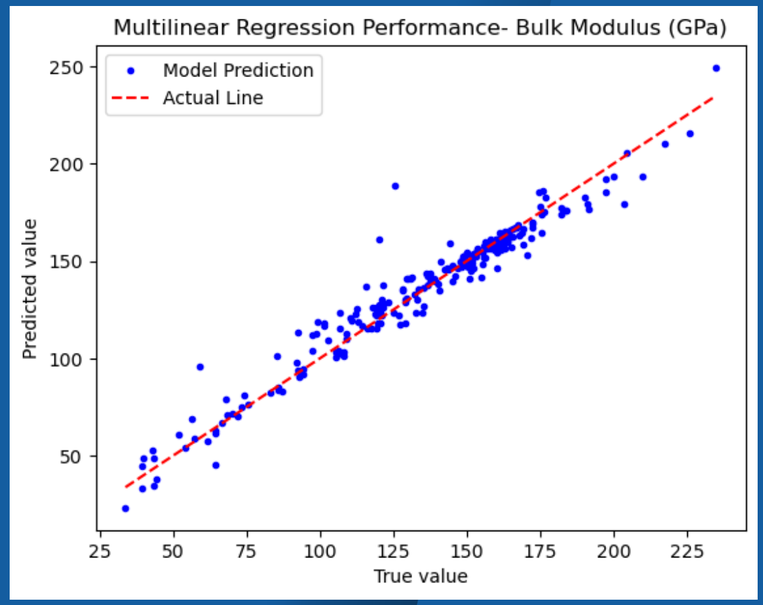
19D100011
Shiv Modi

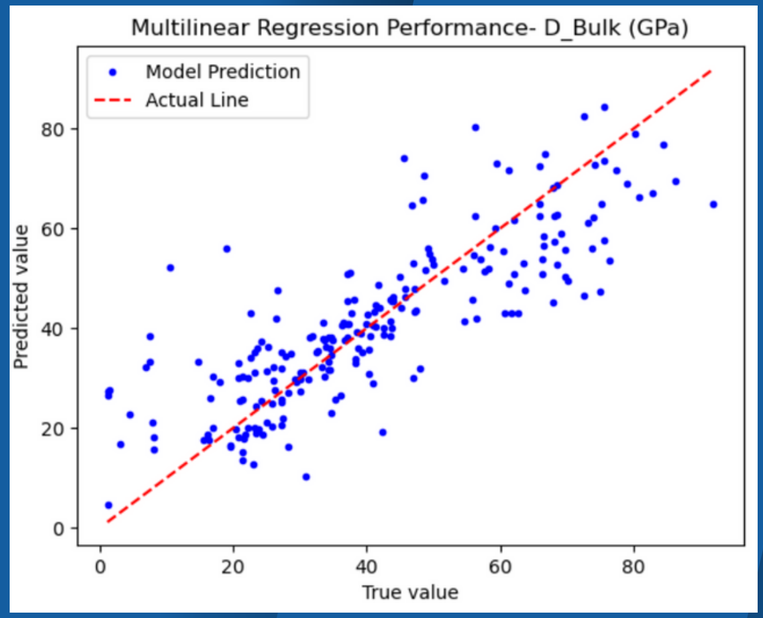
MULTILINEAR REGRESSOR

METHODOLOGY

We employed a relatively simple model- multilinear regressor. The results were good- **r^2 value of 0.94 for a train-test split of 4:1**.
Performance for D_bulk was worse (as expected)- a r^2 score of only 0.69

Still, proper feature selection needs to be done in this aspect as the model performance dipped to as low as r^2 = 0.84 when we tried to restrict features.





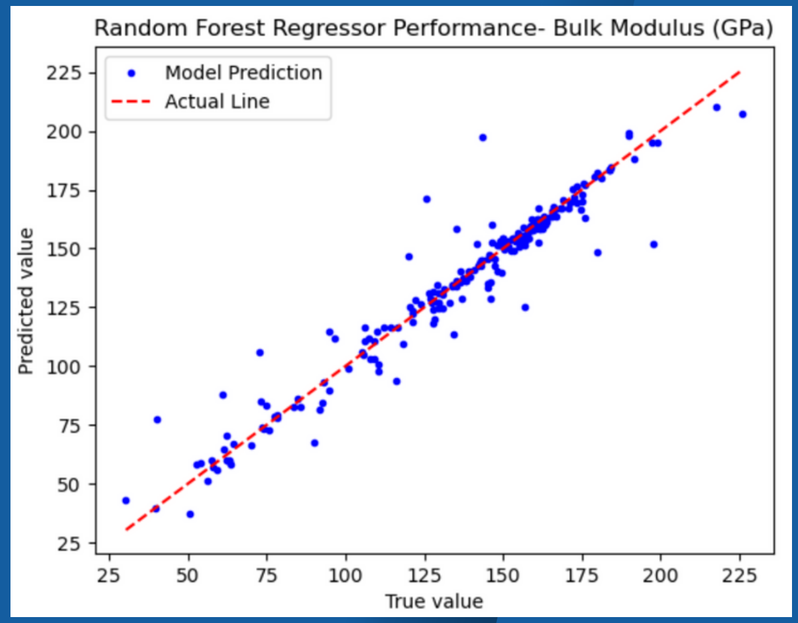
22B2259
Archit Mundada

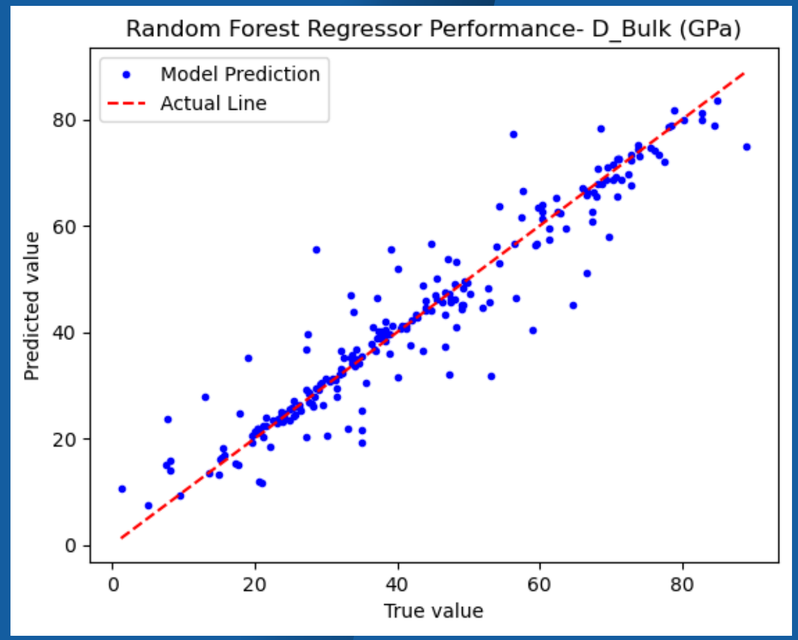
RANDOM FOREST REGRESSOR

METHODOLOGY

Employing this regressor was one of our initial goals with this project. The results obtained were roughly, an OOB score of 0.72 and r^2 score of 0.93 for Bulk Modulus, and OOB score of 0.72 and r^2 of 0.90 for D_bulk. A non linear model performed better for D_Bulk, as expected.

An observation was that the r^2 values did not change much on reducing the feature selections.



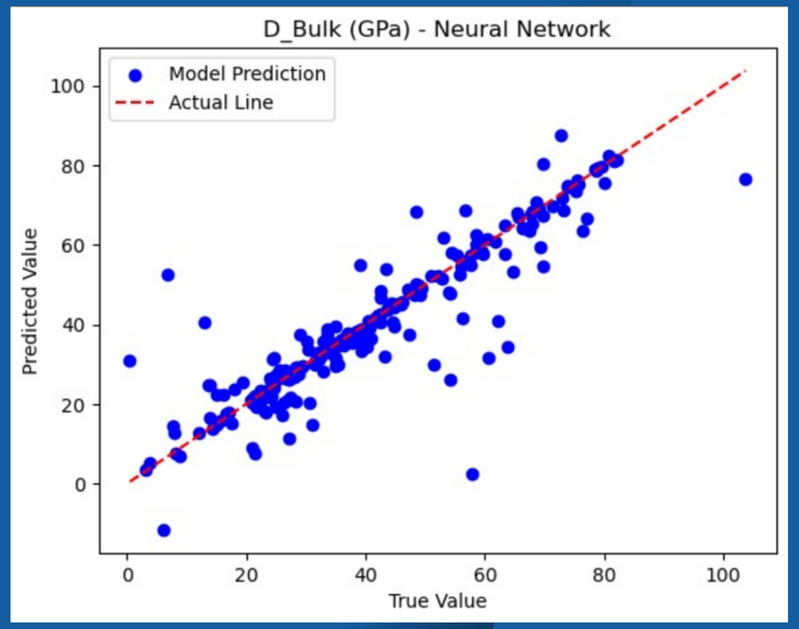


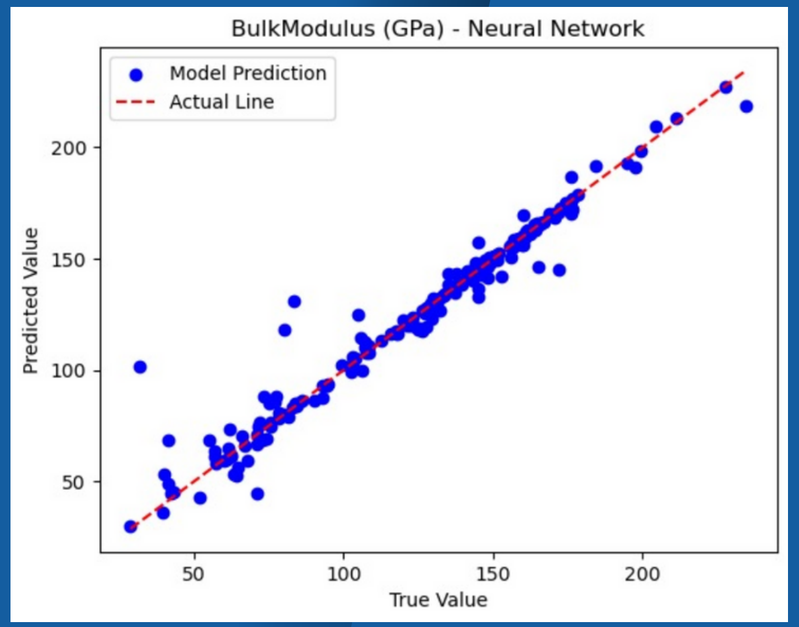
200100104
Mokshit Naidu

NEURAL NETWORK

METHODOLOGY

- The important takeaway from this model is the verification of the non-linear relationship of variables.
- We used the MLP Regressor, whose default activation function is the ReLu function.
- The test accuracy for D_Bulk was 0.88, significantly higher compared to the multilinear regressor model.





19D100011
Shiv Modi

CONCLUSION

Next up: refining the current models

Although we achieved decent performance across our models, we can still tweak their parameters to get better performance.

Interpretation of model differences

Observation of plots for model choice and correlation for feature selection has made the explanation of model differences possible.

Observations:

There are about 30 important parameters to work with according to our analyses for Bulk Modulus, and some of these are trivial- related to the stoichiometric ratio of the compound

Model choices

Upon our observation of the data, we will go ahead with tree methods to create the final deliverable. The EDA method has proven good for model selections.