Prediction of the ERR for mode I (DCB) for a composite laminate with arbitrary properties using ML

**Introduction**

Failures in Structures are common issues in the word composite materials   
When the composite material is produced the manufacturer can not guarantee the quality of the properties of the material because it is man-made, and this creates a deviation of + -15% to the amount of energy required to cause a fracture,  
 it is a very important subject because there are a lot of accidents that happen because of Failures in Structures (Titanic, Challenger, etc), and getting a machine learning to understand what is the energy that leads to fracture can even save a life if it will be very accurate and make composite Structures safer to use and design.  
we used machine learning models with given data about the layers of the material such as :

"UD:E\_A",UD:E\_T",UD:G\_A",

" UD:G\_T"," UD:nu\_A"," UD:nu\_T"," weave 0/90:E\_{11} = E\_{33}",

" weave 0/90:E\_{22}"," weave 0/90:G\_{13}",

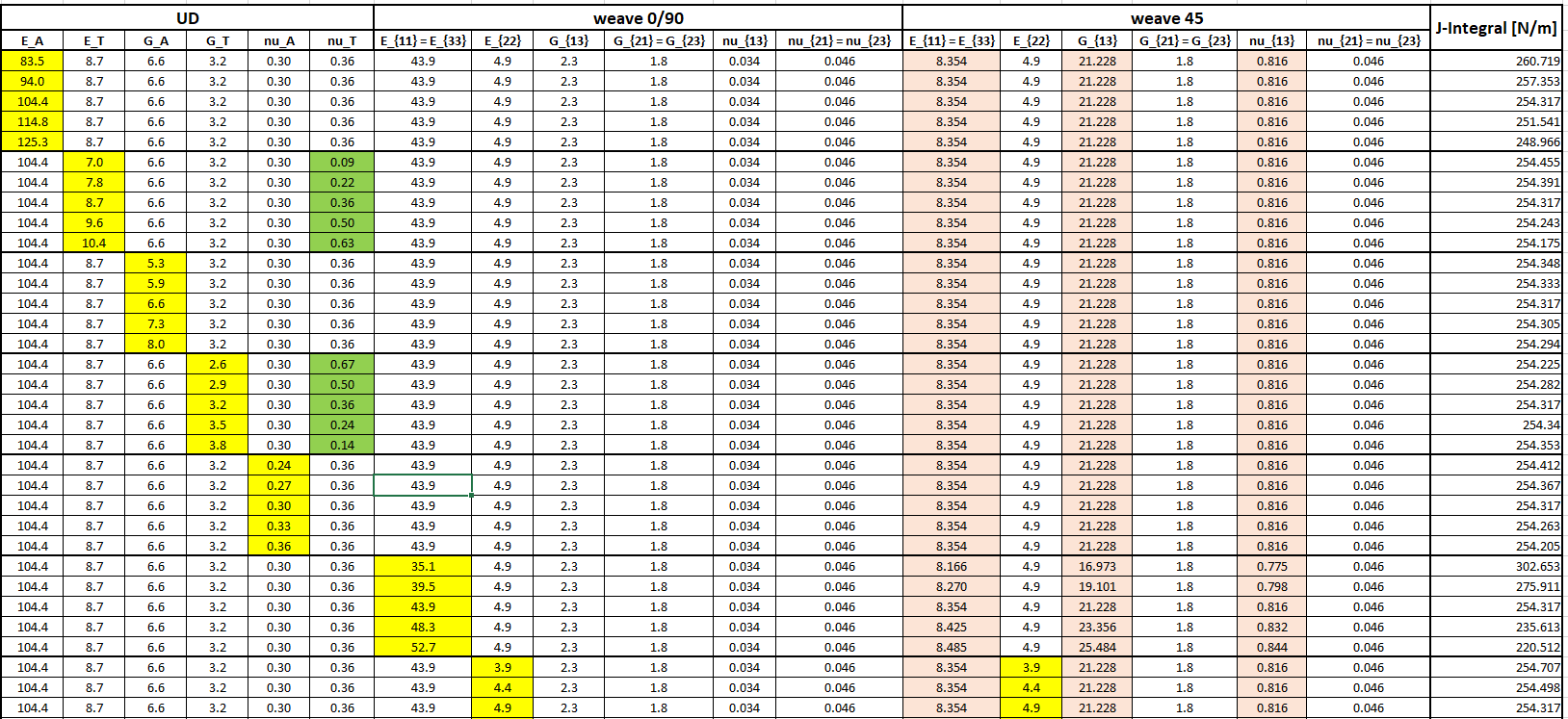
" weave 0/90:G\_{21} = G\_{23}"," weave 0/90:nu\_{13}"," weave 0/90:nu\_{21} = nu\_{23}",

" weave 45:E\_{11} = E\_{33}"," weave 45:E\_{22}"," weave 45:G\_{13}",

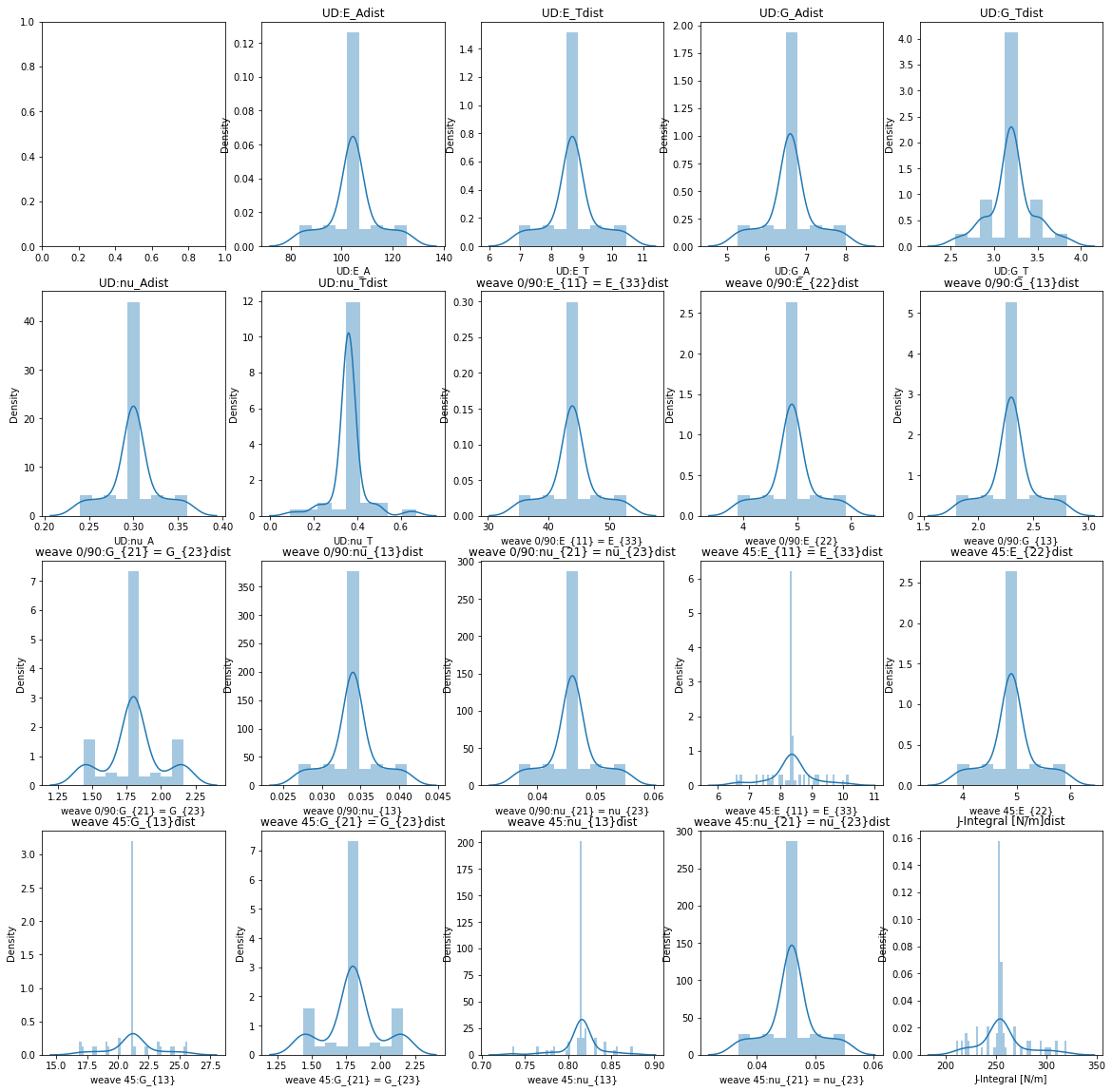
" weave 45:G\_{21} = G\_{23}"," weave 45:nu\_{13}"," weave 45:nu\_{21} = nu\_{23}",

and the target value was the amount of energy it will take to break it (J-integral [N/m])

**Dataset and Features**

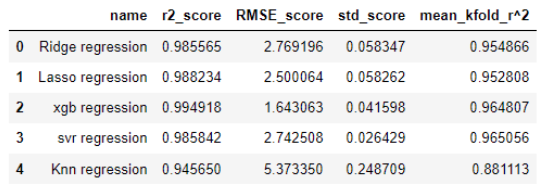
The data set has 90 rows and 17 features that in each row one of the material Properties has changed (the yellow tabs in the table below)  
 and in 50 from the 90 all the values changed at once

And the distribution of the features can be seen below

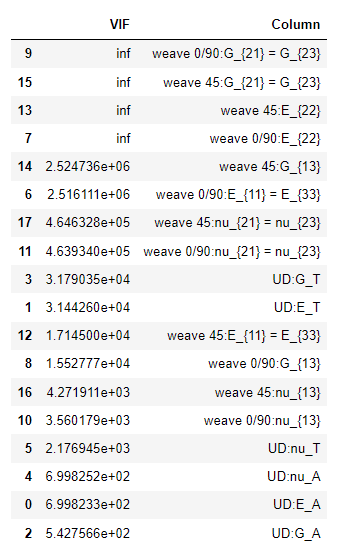
****

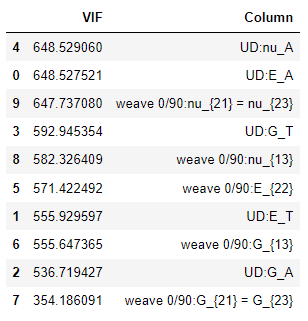
**Method and results**

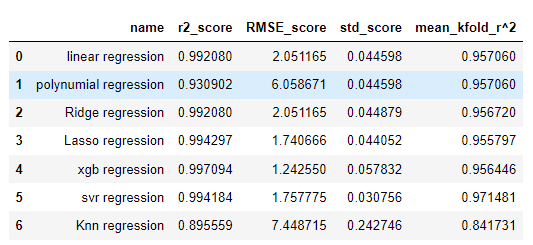
In the beginning, we tested Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression, K Neighbors Regressor, Random Forest Regressor, SVM, and Xgboost  
and the leading value that we used to optimize the model is the R^2 score,   
after getting more data we have used the following algorithms (also can see the scores they are getting in the table )

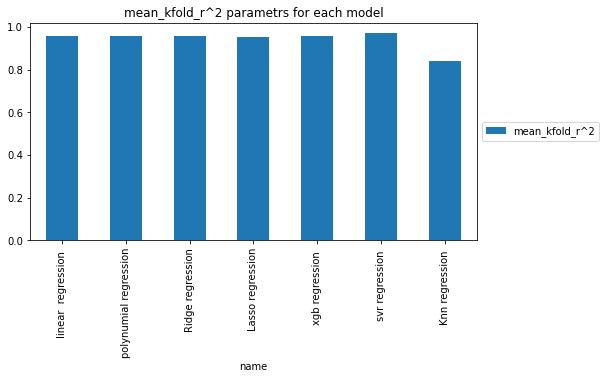


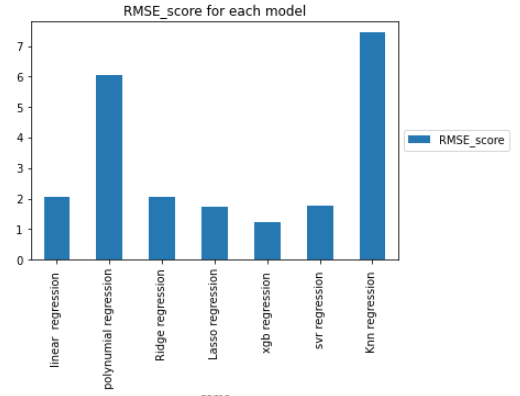
after checking these algorithms and after we get more data from Mor, we checked the VIF to detect multicollinearity because linear regression, Ridge, and lasso are not like tree-based algorithms (and also SVM) and they affected by multicollinearity

the outcomes of the multicollinearity test were very high as can see in the table below on the left side, and on the right side can see what happens if we will remove the weave 45 feature (because they have a connection to the other features because it is the same power but just from a 45-degree angle as Mor explained)



so, the decision we make with Mor is to stick to tree-based and SVM algorithms (xgboost and SVM) because they influenced by multicellularity the results for all the models after getting all data set can be seen in the table and the RMSE and mean of CV R^2 scores are in the charts below: 

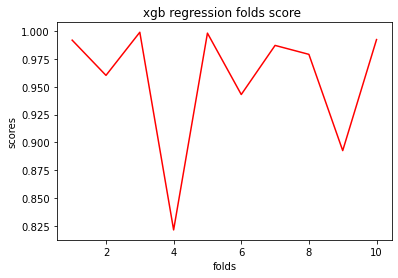


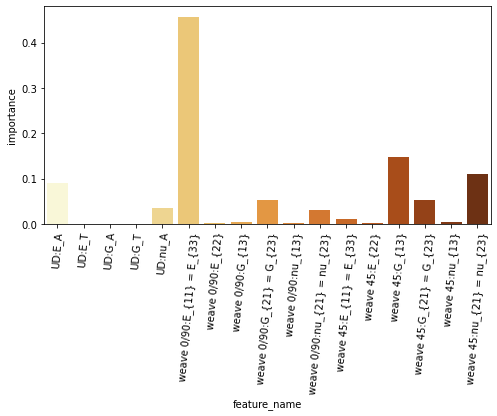
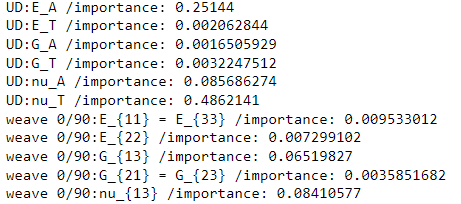


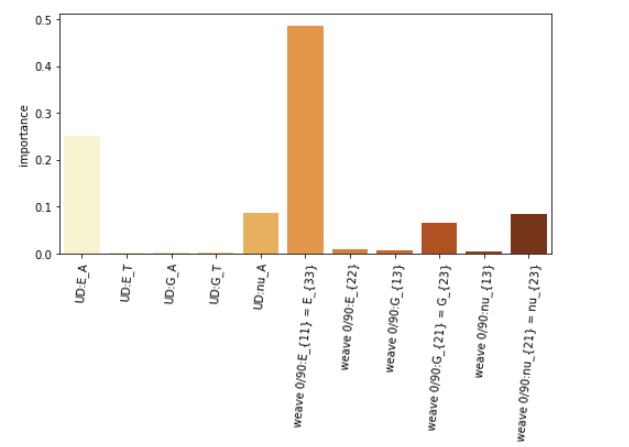
And because of the good outcomes and that Xgboost doesn’t have a multicollinearity effect we choose Xgboost as the model for this data (the CV outcomes can be seen in the chart below and also the R^2 and RMSE scores and also the hyperparameters of the models can see below)

the best hyperparameters for xgboost

base score=0.5, booster='gbtree', colsample\_bylevel=1,colsample\_bynode=1, colsample\_bytree=0.4, enable\_categorical=False,gamma=0.3, gpu\_id=-1, importance\_type=None,interaction\_constraints='', learning\_rate=0.1, max\_delta\_step=0,max\_depth=6, min\_child\_weight=1, missing=nan,monotone\_constraints='()', n\_estimators=100, n\_jobs=8,num\_parallel\_tree=1, predictor='auto', random\_state=0, reg\_alpha=0,reg\_lambda=1, scale\_pos\_weight=1, subsample=1, tree\_method='exact',validate\_parameters=1, verbosity=None)

xgboost model chart:  
After that, we make feature importance of the Xgboost model, and the results were compared to the knowledge that Mor has about the importance of the features that were dropped (0 importance) and the features that have relatively high importance   
(as can see in the chart and the list below)

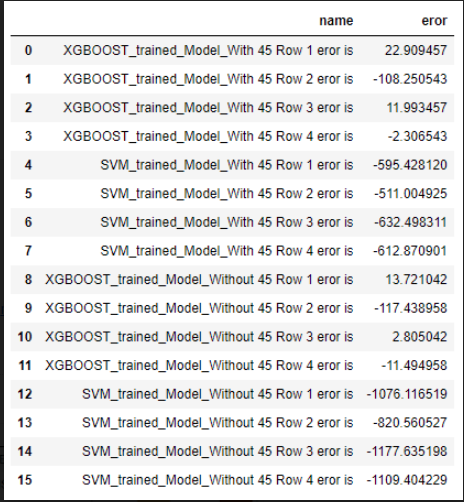
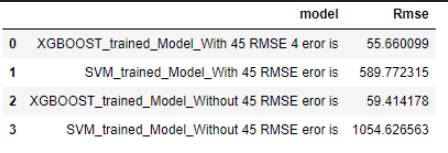


After evaluating the basic model by the ask of Mor we made the same process after removing from the model the “Weave 45” related features and got outcomes that are not so close and the cv scores are not so good for the Xgboost (have high std in the cross-validation that is more than 2 times from the original model) and the feature importance distribution remains the same as the original model



**Conclusion**

the algorithm with the highest performance was Xgboost, he had the lowest RMSE and the highest R^2 and had great scores in the cross-validation mean R^2 & cross-validation R^2 std   
the most important thing that made us choose the xgboost is that he doesn’t affect by high multicollinearity so although the results as seen before have a high VIF score we don’t need to “throw “ features that are in this case very essential to the model construction because its purpose is to replace the original numeric model the mechanical engineers are using and let the model get the exact same parameters that the numeric model is getting.  
it the end we delivered Mor the Xgboost and SVM 2 models and tested it with “real” data of known material and checked how far the results of the model are from the scientifically known values (values that don’t have a 15% deviation ), and the outcomes can be seen in the tables below :



**Future work**The future work on this project needs to be about getting much more data so the model can reflect the world because around 100 samples are not enough for the model to simulate the world and all the relations between the features and the target value,   
and our advice is that Mor needs to focus on the feature with the high importance as can see in the feature importance chart.  
after testing the real-world results in my opinion the new train data should be more similar to the second row of real data we got from Mor which is a common value for materials.  
in conclusion, the model needs at least 10 times more data(1000 rows ) with different values that have small intervals between each change in the independent variables of the models.

**Link to Github**

<https://github.com/itayzaada/DCB_prediction>