IE 6318 – Project Assignment

Name – Ankit Ranjan

Student ID – 1001832327

Guided by – Dr. Shouyi Wang

Title: Predicting multiple molecular properties for identification of new compounds using data mining techniques

Abstract:

The ability to identify new compounds is important because it allows researchers to determine whether an organism has a genetic disease. A genetic disease is often caused by a mutant version molecule, which has a composition slightly different from that of the original biological molecule it replaces. To that end, a number of druglike properties like molecular weight, topological polar surface area and lipophilicity are commonly used tool for studying molecules. In this project, data mining techniques for multi output regression have been examined to predict the values for these properties based on the compound's fingerprints.

Background:

Biological compounds (like proteins) are present in all living things where they play a central role in the chemical processes essential for life. They are made up of strings of amino acids they fold up in an infinite number of ways into elaborate shapes that hold key to how they carry out vital functions.

Even tiny rearrangement of these vital molecules can have catastrophic effects on our health, so one of the most efficient ways to understand the disease and find new treatments is to study the molecules involved.

An improved understanding of shapes of biomolecules could play a pivotal role in the development of novel drugs to treat diseases. Being able to investigate the shape of biomolecules quickly and accurately has the potential to revolutionize life sciences.

In 1972, Christian Anfinsen was awarded a Nobel prize for his work showing that it should be possible to determine the shape of biomolecules based on the sequence of their amino acid building blocks.

There are 10,000's of human biomolecules and many billions in other species, including bacteria and viruses, but working out the shape of just one requires expensive equipment and can take years.

One of the recent problems to solve was to research the Covid-19 vaccine, where scientists studied how the spike biomolecules on the surface of the Sars-Cov-2 virus interacts with receptors in human cells.

Understanding how a biomolecules sequence folds up into three dimensions is one of the fundamental questions.

By knowing the 3D structures of the biomolecules, we can help to design drugs and intervene with health problems whether those be infections or inherited disease.

A better understanding of biomolecules structures and the ability to predict them using a computer means a better understanding of life, evolution, and, of course, human health and disease.

Introduction:

Druglikeness is a qualitative concept used in drug design for how "druglike" a substance is with respect to factors like bioavailability. It is estimated from the molecular structure before the substance is even synthesized and tested. A druglike molecule has properties such as:

Molecular weight (ExactMolWt in RDkit): The smaller the better, because diffusion is directly affected. The great majority of drugs on the market have molecular weights between 200 and 600 Daltons, and particularly <500; they belong to the group of small molecules.
Proteins are made up of amino acid residues which are bound together by peptide bonds between the amino nitrogen (N) and the carboxyl group. Just 21 distinct amino acids exist, but they can bind together in such a variety of three-dimensional (3D) polypeptide chains that almost limitless potential protein chain sequences exist.

This leads to millions of intricate, distinct, potential protein structures. The diversity of protein structures is due to subtle chemical variations which come from differences in the amino acids charge, shape, functional group composition, and size.

Protein molecular weight can be accurately predicted based on the known molecular weights of the amino acids if the sequence and composition of amino acids in a linear chain are known.

Protein’s structure is mainly defined by:

Primary Structures: The underlying amino acid chain within the polypeptide.

Secondary Structures: The main substructures that form from these bound amino acid chains.

Tertiary Structures: The formation of the final 3D structure made up of secondary structures, such as α-helical secondary phases which are held together by several mechanisms.

As the primary structure is the most foundational level of protein structure, protein molecular weight is a key parameter to confirm. Having a robust understanding of the unmodified protein molecular weight can help in initial assessments of the biomolecule’s functionality. This could include:

- Metabolic regulation
- Binding and transportation of small molecular species
- Enzyme catalysis
- Immunological responses
- Gene regulation

Lipophilicity: The role of lipophilicity (MolLogP in RDkit) in determining the overall quality of candidate drug molecules is of paramount importance. Recent developments suggest that, as well as determining pre-clinical ADMET (absorption, distribution, metabolism, elimination and toxicology) properties, compounds of optimal lipophilicity might have increased chances of success in development.

A negative value for logP means the compound has a higher affinity for the aqueous phase (it is more hydrophilic); when logP = 0 the compound is equally partitioned between the lipid and aqueous phases; a positive value for logP denotes a higher concentration in the lipid phase (i.e., the compound is more lipophilic). LogP = 1 means there is a 10:1 partitioning in Organic : Aqueous phases.

Topological Polar Surface Area: The **polar surface area** (**PSA**) or **topological polar surface area** (**TPSA**) (CalcTPSA in RDkit) of a molecule is defined as the surface sum over all polar atoms or molecules, primarily oxygen and nitrogen, also including their attached hydrogen atoms.

PSA is a commonly used medicinal chemistry metric for the optimization of a drug's ability to permeate cells. Molecules with a polar surface area of greater than 140 angstroms squared tend to be poor at permeating cell membranes. For molecules to penetrate the blood–brain barrier (and thus act on receptors in the central nervous system), a PSA less than 90 angstroms squared is usually needed.

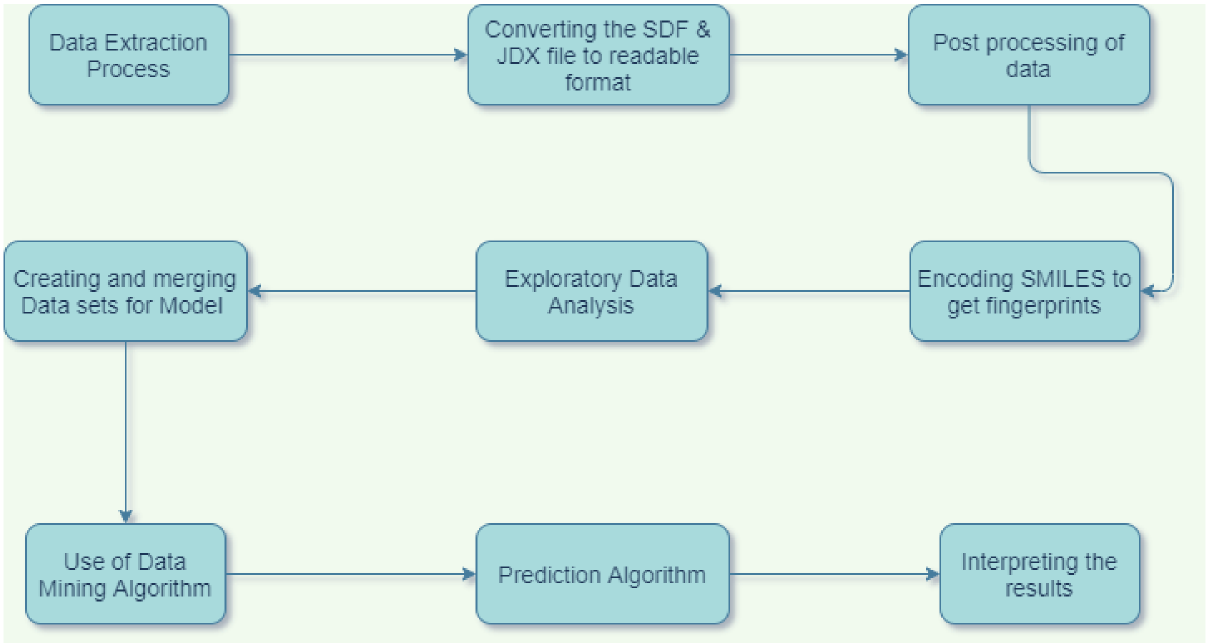
SMILES, which was proposed by Weininger, is currently widely recognized and used as a standard representation of compounds for modern chemical information processing. SMILES provides a linear notation method to represent chemical compounds in a unique way in the form of strings over a fixed alphabet. SMILES uses specific grammar and characters to describe all the atoms and structure of a chemical compound. SMILES can strictly express structural differences including the chirality of compounds.

A “fingerprint” is a vector that represents a property of a chemical compound. Many methods for creating fingerprints have been reported. The launch pad we normally use for all fingerprints is 2D fingerprint to indicate what kind of partial structure the compound possesses. In this regard, the most commonly used algorithm is the extended-connectivity fingerprint (ECFP, also known as the circular fingerprint or Morgan fingerprint).

The notion of *chemical similarity* (or *molecular similarity*) is one of the most important concepts in chemoinformatics.[1][2] It plays an important role in modern approaches to predicting the properties of chemical compounds, designing chemicals with a predefined set of properties and, especially, in conducting drug design studies by screening large databases containing structures of available (or potentially available) chemicals. These studies are based on the similar property principle of Johnson and Maggiora, which states: *similar compounds have similar properties*.

The most popular **similarity** measure for comparing chemical structures represented by means of fingerprints is the **Tanimoto** (or Jaccard) coefficient T. Two structures are usually considered similar if T > 0.85 (for Daylight fingerprints).

Flow Chart:


Data Extraction Process
Converting the SDF & JDX file to readable format
Post processing of data
Encoding SMILES to get fingerprints
Exploratory Data Analysis
Creating and merging Data sets for Model
Use of Data Mining Algorithm
Prediction Algorithm
Interpreting the results

Exploratory Data Analysis:

Parameter Name	Unit	Description
Setting/Input Parameters:		
Fingerprints	Bit vector	It is a vector that represents a property of a chemical compound.
Output Parameters:		
Exact Molecular weight	Mols	The molecular weight of the compound in mols

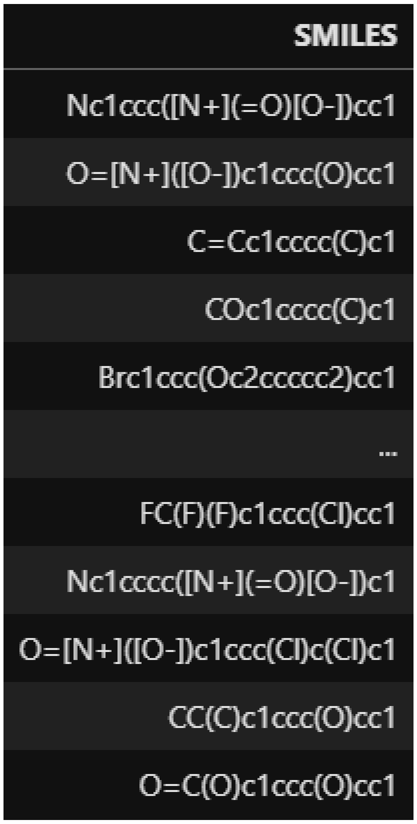
Lipophilicity	LogP	LogP, this is the partition coefficient of a molecule between an aqueous and lipophilic phase, usually octanol and water.
Topological Polar Surface Area	Å^2	The surface sum over all polar atoms or molecules, primarily oxygen and nitrogen, also including their attached hydrogen atoms.

The 3-D structure of the provided compounds was extracted from the website pubchem.com in the file format .SDF (Spatial Data File), whereas the mass spectra of the compounds were extracted from the website, webbook.nist.gov/chemistry in .JDX format (JCAMP-DX). This JDX format is to be used for a future study to predict the mass spectra of compounds from their fingerprints. The data must be processed appropriately in a Python data frame to enable us to apply the relevant machine learning algorithms.

Preprocessing:

RDkit was used to read the SDF format into the data frame and JDX format was read with JCAMP-DX. The common files in the respective folders of the aforementioned file formats were extracted so as to prepare the data for a future study on mass spectra prediction as well.

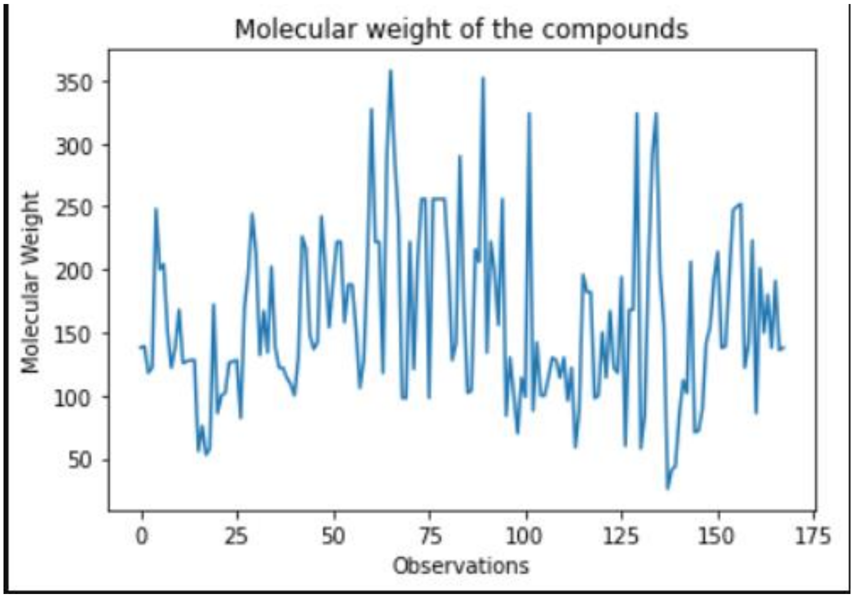
Using RDkit, the SMILES were extracted from the SDF files and beyond that, the chemical fingerprints of each compound was extracted from these SMILES. Each fingerprint is a bit vector of binary data representing the properties of the compound. This is an extremely sparse vector and has a large size of more than 8 million features. Thus, a PCA was performed on the training and testing dataset so as to extract the 50 most important features. This was done to enable the processing of this computationally expensive data. Thus, the desired array for the input data for the machine learning models is obtained.



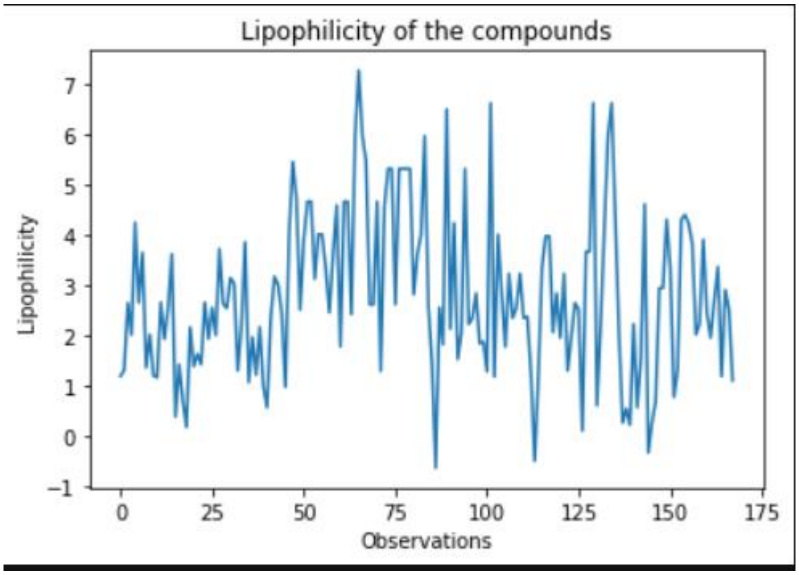
SMILES
Nc1ccc([N+](=O)[O-])cc1
O=[N+]([O-])c1ccc(O)cc1
C=Cc1cccc(C)c1
COc1cccc(C)c1
Brc1ccc(Oc2ccccc2)cc1
...
FC(F)(F)c1ccc(Cl)cc1
Nc1cccc([N+](=O)[O-])c1
O=[N+]([O-])c1ccc(Cl)c(Cl)c1
CC(C)c1ccc(O)cc1
O=C(O)c1ccc(O)cc1

The three properties to be predicted were the exact molecular weight, Lipophilicity and Topological Polar Surface Area (using the in-built functions from RDkit: ExactMolWt (mols), MolLogP (LogP), CalcTPSA(Å^2)).

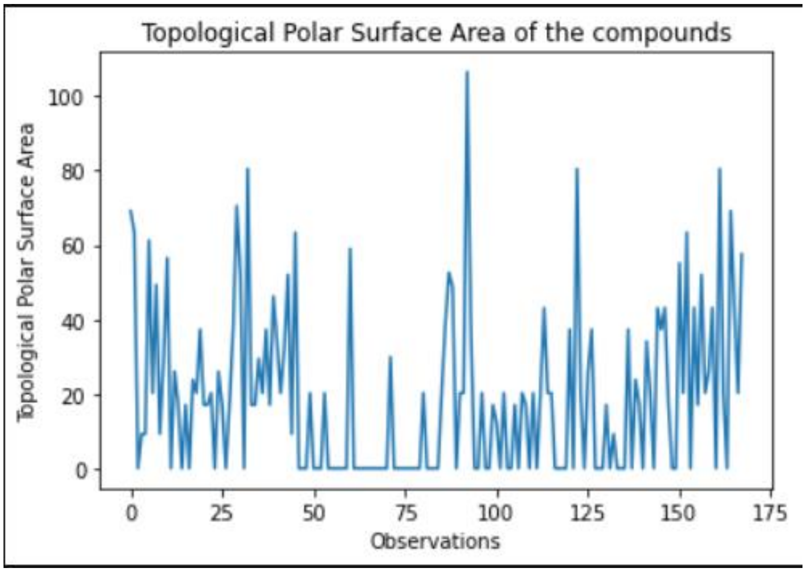
Data visualization:



From the above plot we observe that for the collection of compounds in this dataset, the molecular weights vary vastly, from 50 to 350.

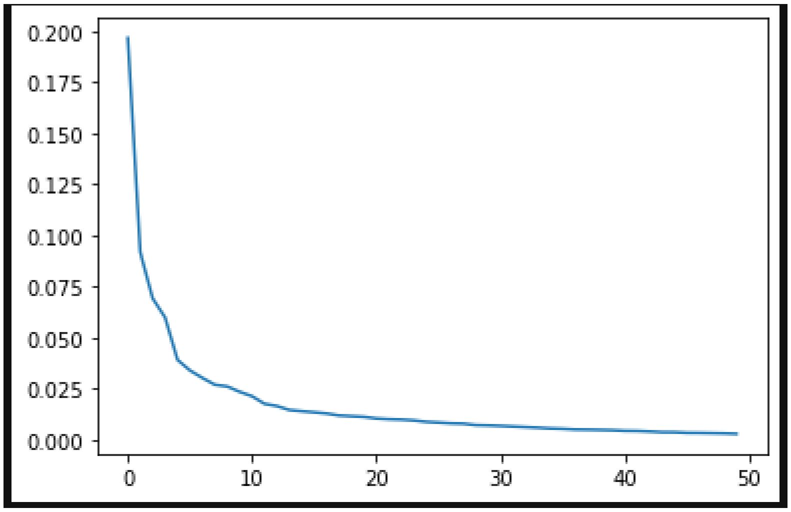


The lipophilicity of the compounds is more evenly spread in a range of a small negative value to 7.

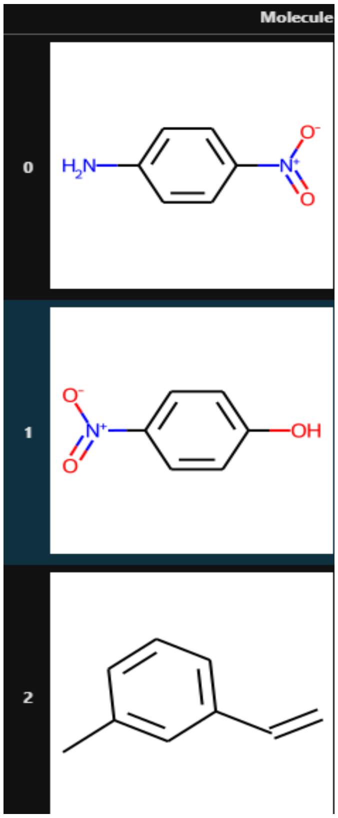


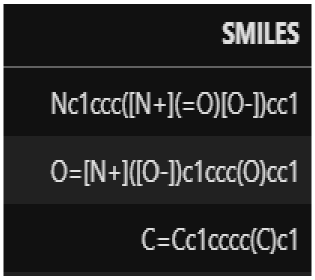
From the above plot we observe that for the collection of compounds in this dataset, the TPSA vary vastly, from 0 to 100.

Using RDkit, the molecular structures for some molecules are:

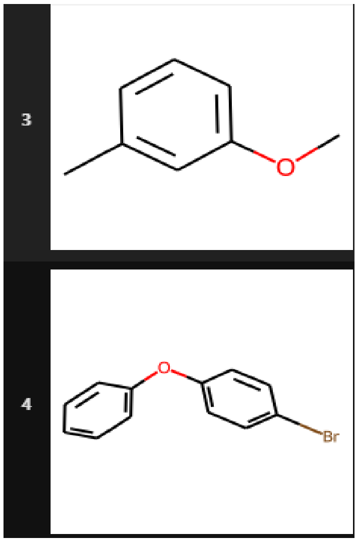


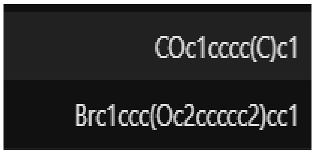
From the PCA it is evident that for 20% of the variance, the most variance is explained within 5 features.





SMILES
Nc1ccc([N+](=O)[O-])cc1
O=[N+]([O-])c1ccc(O)cc1
C=Cc1cccc(C)c1





COc1cccc(C)c1
Brc1ccc(Oc2ccccc2)cc1

Prediction methods and Results:

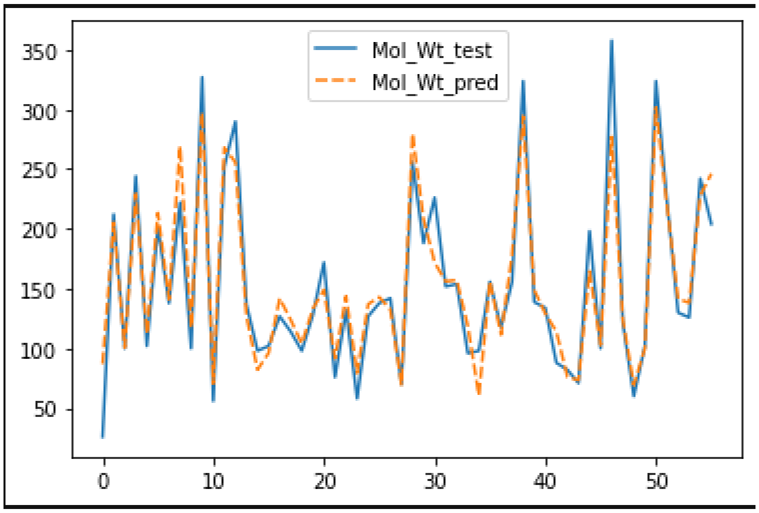
Algorithms	Description
K-nearest Neighbors Regression	KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood. In K-Nearest Neighbors Regression the output is the property value for the object. There is no model other than the raw training dataset and the only computation performed is the querying of the training dataset when a prediction is requested. It is a simple algorithm, but one that does not assume very much about the problem other than that the distance between data instances is meaningful in making predictions. As such, it often achieves very good performance. When making predictions on regression problems, KNN will take the mean of the k most similar instances in the training dataset.
Linear Regression	Multioutput regression are regression problems that involve predicting two or more numerical values given an input example. An example might be to predict a coordinate given an input, for example predicting x and y values. Another example would be multi-step time series forecasting that involves predicting multiple future time series of a given variable. Many machine learning algorithms are designed for predicting a single numeric value, referred to simply as regression. Some algorithms do support multioutput regression inherently, such as linear regression and decision trees. There are

	also special workaround models that can be used to wrap and use those algorithms that do not natively support predicting multiple outputs.
Random Forest Multioutput Regression	Random Forest regression refers to ensembles of regression trees where a set of n tree un-pruned regression trees are generated based on bootstrap sampling from the original training data. For each node, the optimal feature for node splitting is selected from a random set of m feature from the total N features. The selection of the feature for node splitting from a random set of features decreases the correlation between different trees and thus the average prediction of multiple regression trees is expected to have lower variance than individual regression trees. A random forest regressor is used, which supports multi-output regression natively, so the results can be compared. The random forest regressor will only ever predict values within the range of observations or closer to zero for each of the targets. As a result, the predictions are biased towards the center of the circle. Using a single underlying feature, the model learns both the x and y coordinate as output.
Gradient Boosting for Regression	GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function. Gradient Boosting Regressor supports a number of different loss functions for regression which can be specified via the argument loss; the default loss function for regression is least squares ('ls').

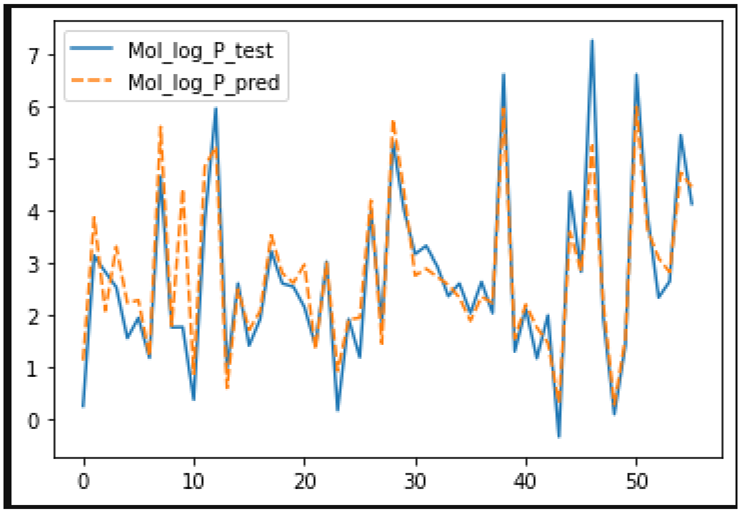
The total dataset chosen was 178 compounds and testing set was 33% of that.

Linear Regression:

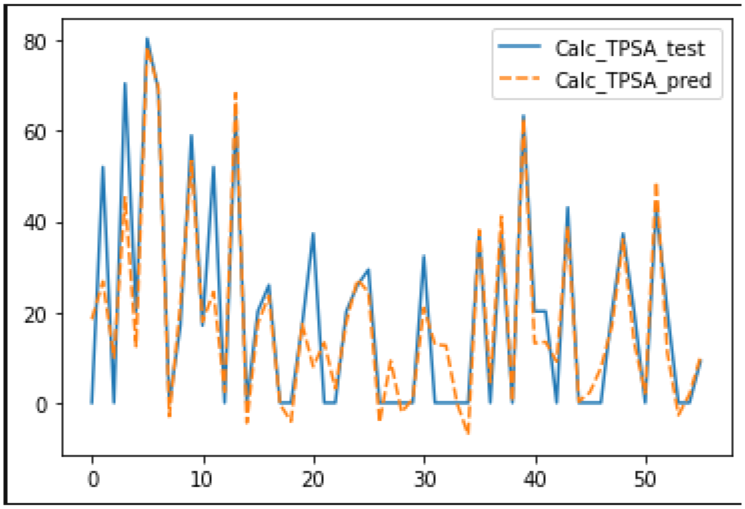
Molecular weight (X – Observations vs Y – Predicted and True values):



Lipophilicity (X – Observations vs Y – Predicted and True values):



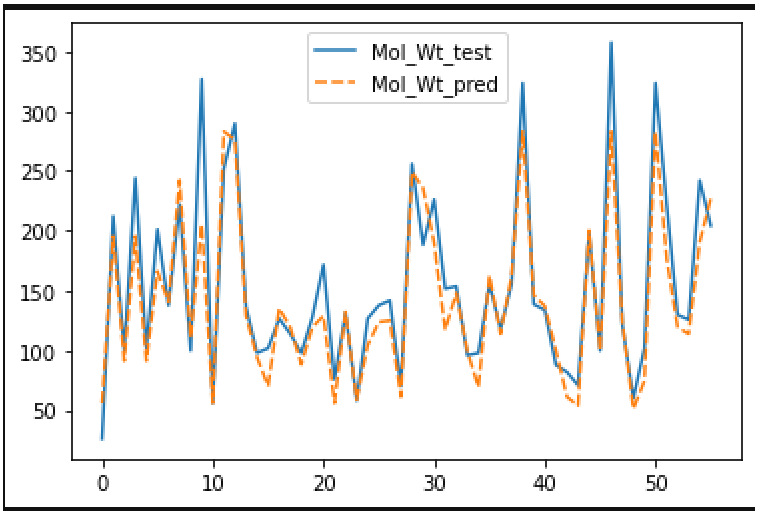
TPSA (X – Observations vs Y – Predicted and True values):



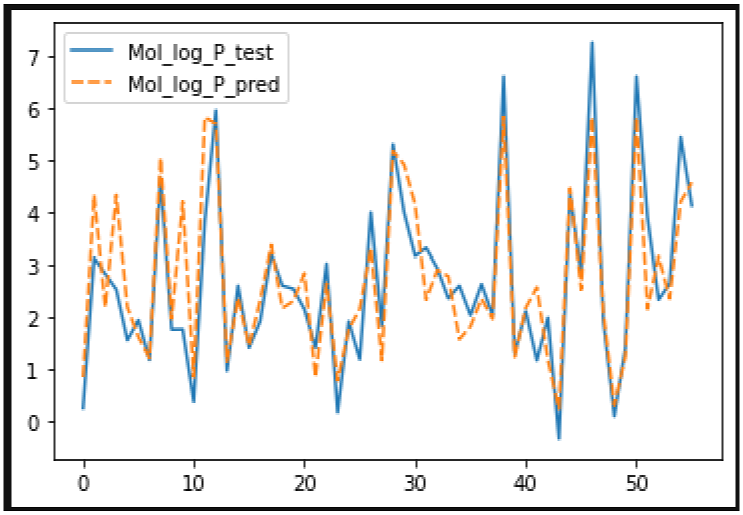
For Linear Regression from the above graphs, the predicted values for all three properties were close to the actual values.

KNN Regressor:

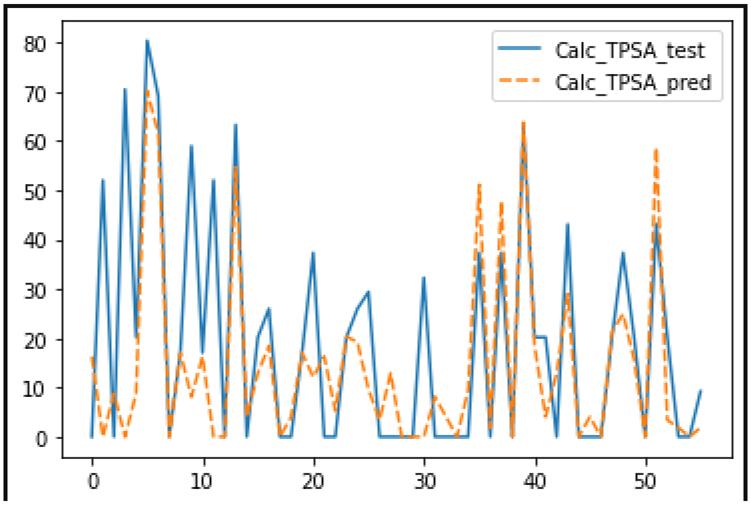
Molecular weight (X – Observations vs Y – Predicted and True values):



Lipophilicity (X – Observations vs Y – Predicted and True values):



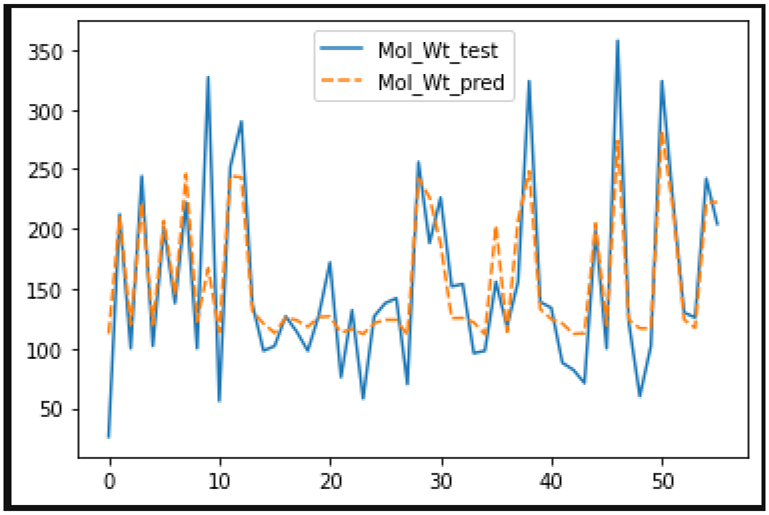
TPSA (X – Observations vs Y – Predicted and True values):



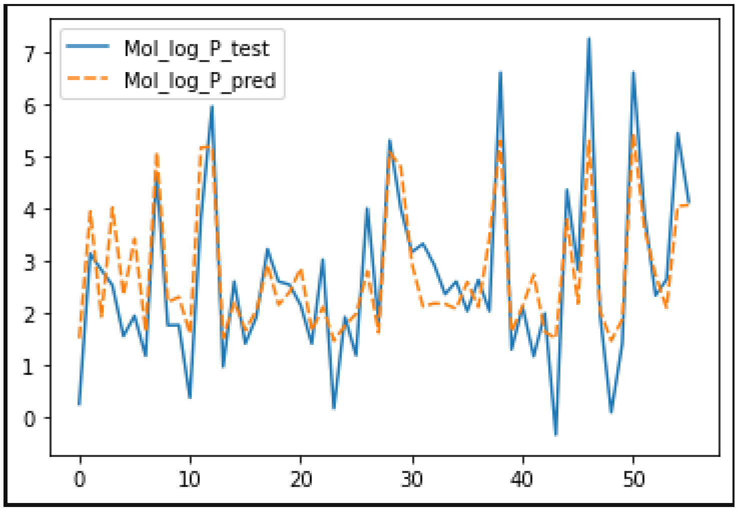
For KNN regressor from the above graphs, the predicted values for TPSA were far from the actual but it was closer for Lipophilicity and Molecular weight.

Random Forest Regressor:

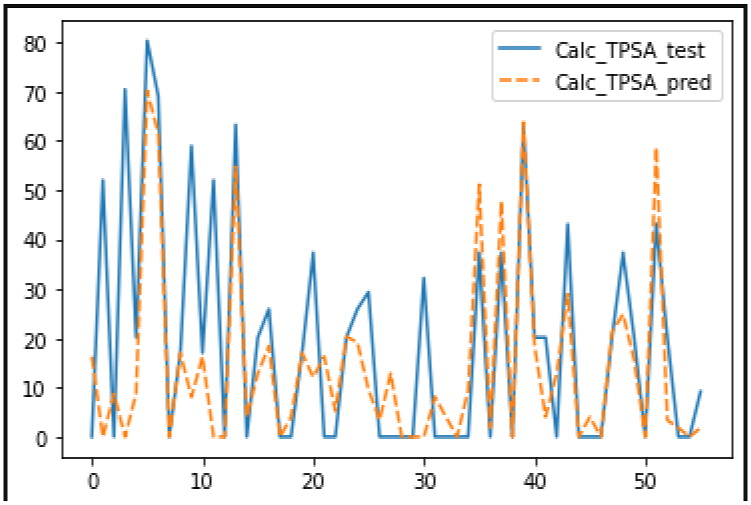
Molecular weight (X – Observations vs Y – Predicted and True values):



Lipophilicity (X – Observations vs Y – Predicted and True values):



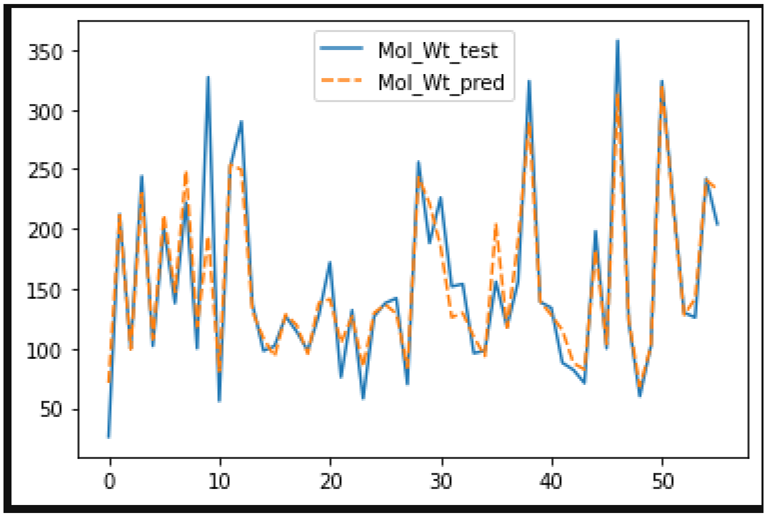
TPSA (X – Observations vs Y – Predicted and True values):



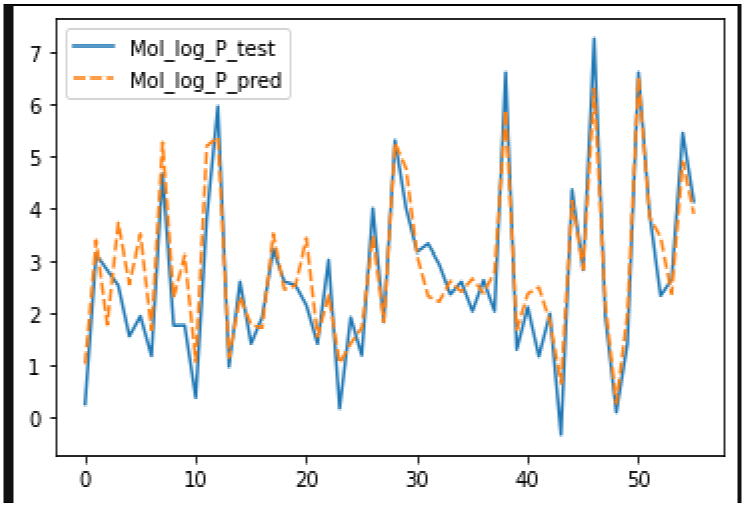
For Random Forest regressor from the above graphs, the predicted values for TPSA were slightly far from the actual but it was closer for Lipophilicity and Molecular weight.

Gradient Boosting:

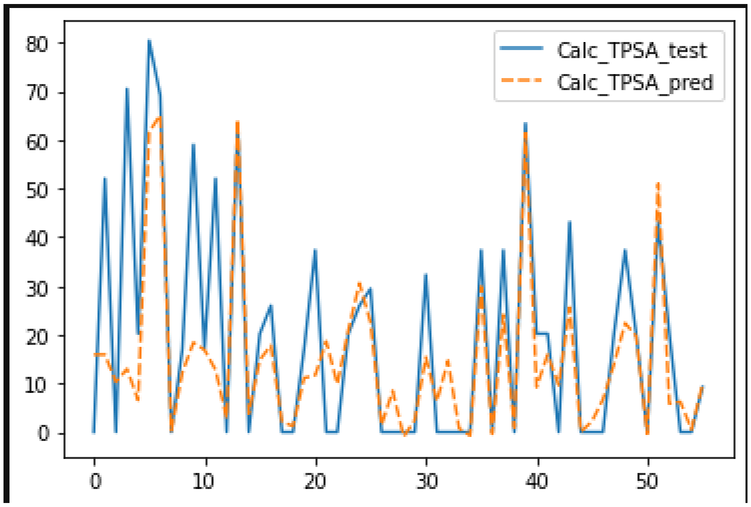
Molecular weight (X – Observations vs Y – Predicted and True values):



Lipophilicity (X – Observations vs Y – Predicted and True values):



TPSA (X – Observations vs Y – Predicted and True values):



For Gradient Boosting from the above graphs, the predicted values for Lipophilicity were slightly far from the actual but it was closer for TPSA and Molecular weight.

Evaluation metrics and distance/similarity measures:

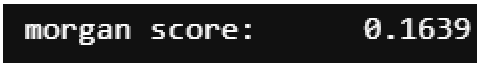
Similarity measure:

Jaccard Coefficient: The most popular similarity measure for comparing chemical structures represented by means of fingerprints is the Tanimoto (or Jaccard) coefficient T. Two structures are usually considered similar if T > 0.85 (for Daylight fingerprints). [5]. The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for gauging the similarity and diversity of sample sets.

Jaccard Coefficients

J = number of 11 matches / number of not-both-zero attributes' values = (M11) / (M01 + M10 + M11)

From the Morgan Fingerprint function, the chemical similarity was found for two chemical SMILES: 'O=[N+]([O-])c1ccc(O)cc1' and 'Cc1cc(-c2ccc(N)c(C)c2)ccc1N' was found to be:



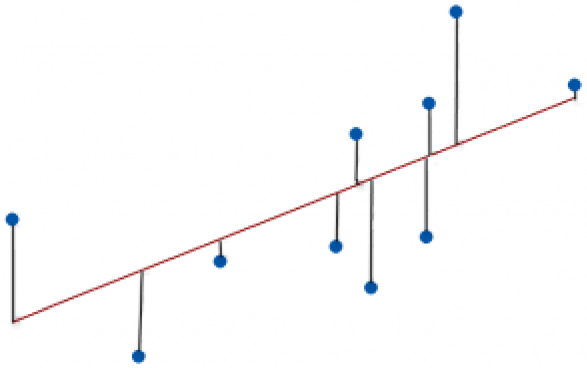
```
morgan score:      0.1639
```

This is another way of comparing the similarity of two compounds than the bit vector representation from the Jaccard coefficients. This similarity is used to compare the likeness of two compounds in drug discovery as well.

R^2 Score

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale. After fitting a linear regression model, you need to determine how well the model fits the data. Does it do a good job of explaining changes in the dependent variable? There are several key goodness-of-fit statistics for regression analysis. In this post, we'll examine R-squared (R^2), highlight some of its limitations, and discover some surprises. For instance, small R-squared values are not always a problem, and high R-squared values are not necessarily good.

Assessing Goodness-of-Fit in a Regression Model- Residuals are the distance between the observed value and the fitted value. Linear regression identifies the equation that produces the smallest difference between all the observed values and their fitted values. To be precise, linear regression finds the smallest sum of squared residuals that is possible for the dataset.



R-squared evaluates the scatter of the data points around the fitted regression line. It is also called the coefficient of determination, or the coefficient of multiple determination for multiple regression. For the same data set, higher R-squared values represent smaller differences between the observed data and the fitted values. R-squared is the percentage of the dependent variable variation that a linear model explains.

R-squared is always between 0 and 100%. 0% represents a model that does not explain any of the variation in the response variable around its mean. The mean of the dependent variable predicts the dependent variable as well as the regression model. 100% represents a model that explains all the variation in the response variable around its mean. Usually, the larger the R^2, the better the regression model fits your observations. However, this guideline has important caveats that I'll discuss in both this post and the next post.

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

Adjusted R-squared

The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors. Suppose you compare a five-predictor model with a higher R-squared to a one-predictor model. Does the five-predictor model have a higher R-squared because it's better? Or is the R-squared higher because it has more predictors? Simply compare the adjusted R-squared values to find out! The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it is usually not. It is always lower than the R-squared.

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

Root Mean Squared Error

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. Root mean square error is commonly used in climatology, forecasting, and regression analysis to verify experimental results.

$$RMSE = \sqrt{(f-o)^2}$$

Mean Absolute Error

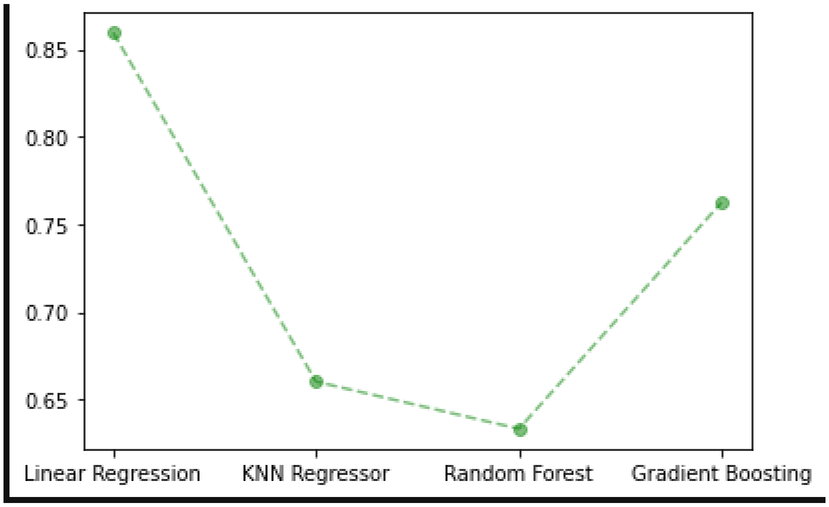
The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. It measures *accuracy* for continuous variables. The equation is given in the library references. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

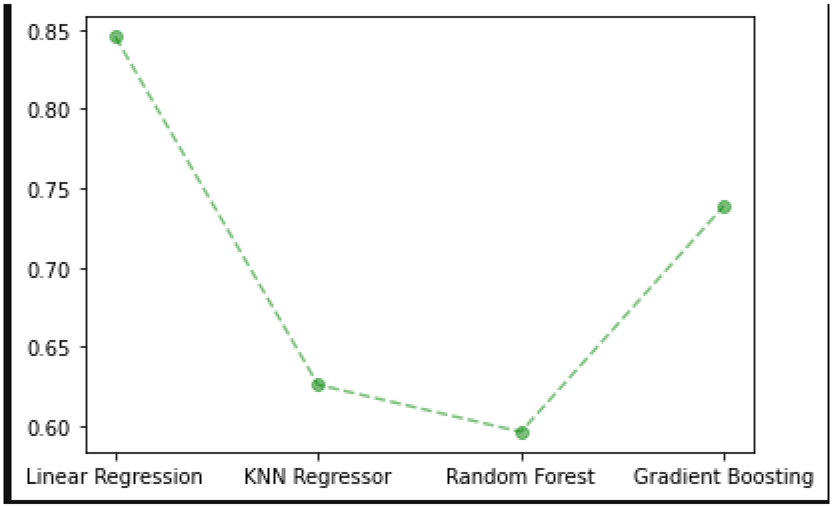
The MAE is also the most intuitive of the metrics since we are just looking at the absolute difference between the data and the model's predictions. Because we use the absolute value of the residual, the MAE does not indicate **underperformance** or **overperformance** of the model (whether the model under or overshoots actual data). Each residual contributes proportionally to the total amount of error, meaning that larger errors will contribute linearly to the overall error. Like we have said above, a small MAE suggests the model is great at prediction, while a large MAE suggests that your model may have trouble in certain areas. A MAE of 0 means that your model is a **perfect** predictor of the outputs (but this will almost never happen). While the MAE is easily interpretable, using the absolute value of the residual often is not as desirable as **squaring** this difference. Depending on how you want your model to treat **outliers**, or extreme values, in your data, you may want to bring more attention to these outliers or downplay them. The issue of outliers can play a major role in which error metric you use.

Evaluation metrics:

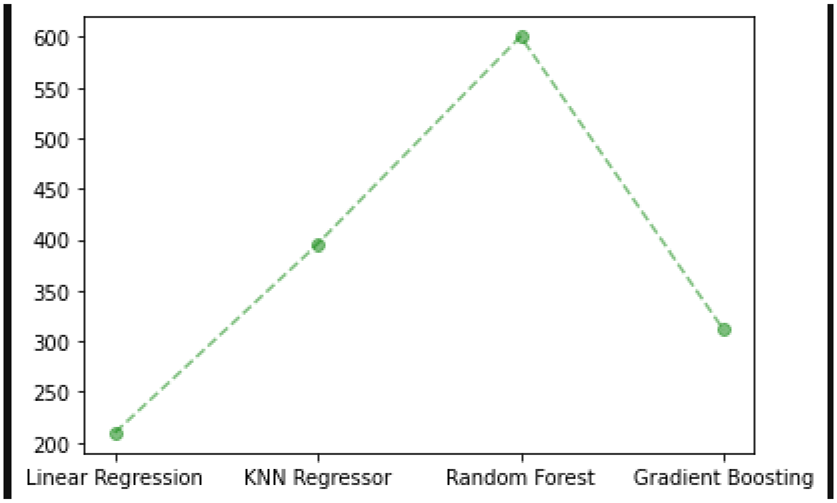
R-2 score:



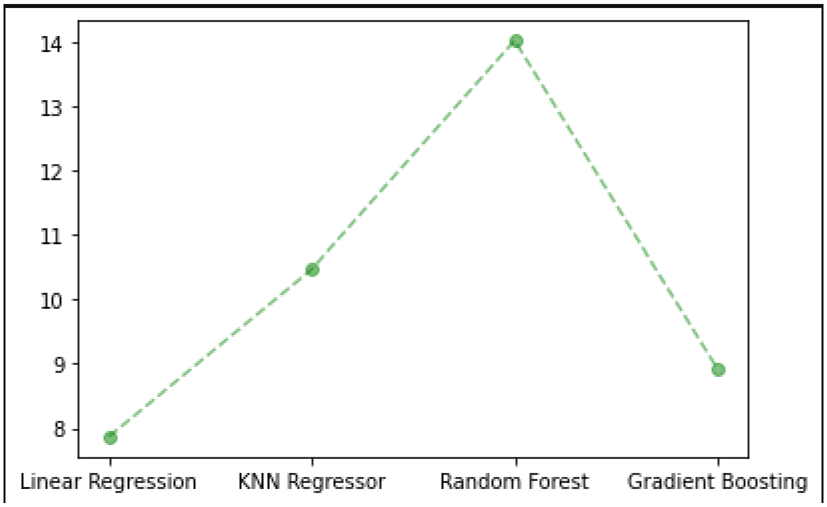
Adjusted R-2 score:

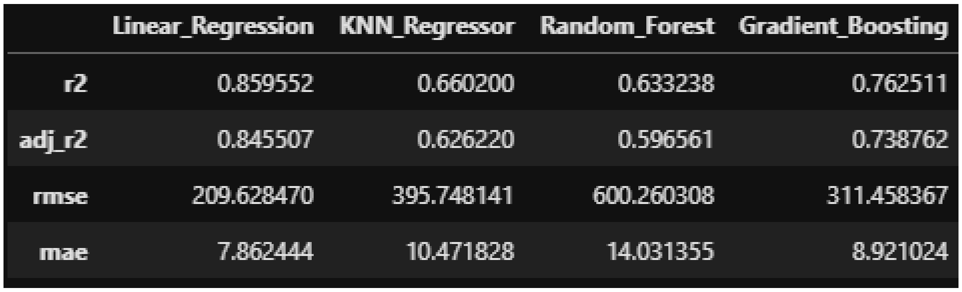

0.85
0.80
0.75
0.70
0.65
0.60
Linear Regression
KNN Regressor
Random Forest
Gradient Boosting

RMSE:


600
550
500
450
400
350
300
250
200
Linear Regression
KNN Regressor
Random Forest
Gradient Boosting

MAE:


14
13
12
11
10
9
8
Linear Regression
KNN Regressor
Random Forest
Gradient Boosting



	Linear_Regression	KNN_Regressor	Random_Forest	Gradient_Boosting
r2	0.859552	0.660200	0.633238	0.762511
adj_r2	0.845507	0.626220	0.596561	0.738762
rmse	209.628470	395.748141	600.260308	311.458367
mae	7.862444	10.471828	14.031355	8.921024

The best metric to evaluate the models is the MAE as its value tells us that the model is predicting a particular amount more or less on average than the actual value.

The other metric that is useful is the R-2 score. The R-2 score value tells us how far the data deviates from a linear model. A horizontal line may predict the data better than the models with negative R-2 score. A possible reason for a negative R-2 score is there are many different compounds' values in the small dataset (as can be seen in the data exploration stage) that causes the MSE of the model to be more than MSE of the baseline causing the R-2 to be negative.

$$R^2 = 1 - \frac{\mathrm{MSE(model)}}{\mathrm{MSE(baseline)}}$$

However, this data does not yield any negative R-2 score for the models tested.

Thus, from these metrics, the best model can be seen to be Linear Regression, with the lowest RMSE, lowest R-2 and adjusted R-2 score, and the lowest Mean Absolute Error. However, we must keep in mind that RMSE can be a misleading metric at times. The lesser the value of MAE and higher the adjusted R-2 score, the better the performance of the model.

Conclusion:

New drug synthesis is an extremely challenging but crucial field for the future of the survival of our species. Linear Regression followed by Boosting gave us promising results (for the appropriate metrics MAE and Adjusted R-2), given the small size of this dataset (due to computational limitations). Even though the Gradient Boosting gives quite good results, but the simpler model i.e., Linear Regression is a better choice to run for a larger dataset of thousands of molecules.

Future Work:

The next step would be to have more molecules for training. Beyond this project, there are many other important properties to predict that aid drug discovery based on their proximity with existing molecules. They include but are not limited to electron ionized mass spectra, solubility, the number of H-bond donors for a molecule, and the number of H-bond acceptors for a molecule. Other models that may be explored are SVR and Neural Networks for Regression.

Acknowledgement:

We would like to thank Dr. Wang and the Teaching Assistant, Linh Ho Manh for their unwavering encouragement and readily available support during the trying aspects of this project.

REFERENCES:

- https://pubs.acs.org/doi/abs/10.1021/ci00057a005
- https://en.wikipedia.org/wiki/Druglikeness
- https://pubmed.ncbi.nlm.nih.gov/22823020/

- https://pubs.acs.org/doi/10.1021/ci100050t
- https://en.wikipedia.org/wiki/Chemical_similarity#:~:text=The%20most%20popular%20similarity%20measure,0.85%20(for%20Daylight%20fingerprints).
- https://www.azom.com/article.aspx?ArticleID=19609
- https://www.acdlabs.com/download/app/physchem/making_sense.pdf
- https://en.wikipedia.org/wiki/Polar_surface_area
- https://www.youtube.com/watch?v=U86Qn9V33Y8
- https://thedatascientist.com/performance-measures-rmse-mae/
- https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/#:~:text=Root%20Mean%20Square%20Error%20(RMSE)%20is%20the%20standard%20deviation%20of,the%20line%20of%20best%20fit.
- https://www.researchgate.net/figure/nterpretation-of-typical-MAPE-values_tbl1_257812432
- https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/#:~:text=The%20RMSE%20is%20the%20square,an%20absolute%20measure%20of%20fit.&text=Lower%20values%20of%20RMSE%20indicate%20better%20fit.
- https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html
- https://cran.r-project.org/web/packages/MultivariateRandomForest/MultivariateRandomForest.pdf
- https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60
- www.pubchem.com
- www.webbook.nist.gov/chemistry
- https://towardsdatascience.com/regression-an-explanation-of-regression-metrics-and-what-can-go-wrong-a39a9793d914