Data Science for Material Innovation

Design Credits Project Report (CHN2010)
Department of Chemical Engineering,
Indian Institute of Technology, Jodhpur

Supervisor: Dr. Deepak Arora (Associate Professor) | Hrishi Raj Singh (B20CH018)

Introduction

Polymers are one of modern society's most critical and enormously used materials. Many kinds of research are taking place on developing new polymers with specific properties. To determine if the new polymer is profitable. Some of the problems that occur are as follows:

- It requires preparing various samples.
- It requires performing several experiments.
- It is time-consuming.
- Not economical.

A new approach known as materials informatics has emerged to reduce the time-to-market and development cost of new products, ideally by two or more.

Material informatics uses Data Science on Large Datasets based on the idea of training machine learning algorithms on large databases to identify previously unrecognized trends or patterns.

Data Science :

Data Science is the domain of study that deals with vast volumes of data using modern tools to find and learn about unseen patterns.

Data Science requires Statistical visualization, AI, and Basic Mathematical concepts. It includes Artificial Intelligence, Machine Learning, and Deep Learning.

Data science involves several stages, including data acquisition, data cleaning, data exploration, modeling, and interpretation.

Data acquisition stage: data is collected from various sources such as databases, sensors, or surveys.

Data cleaning stage: the data is pre-processed to remove any noise or inconsistencies.

Data exploration stage: the data is analyzed using various statistical and visualization techniques to gain insights and identify patterns.

Modeling stage: machine learning algorithms are used to build models that can predict or classify future data.

Interpretation stage: the results of the models are interpreted and communicated to stakeholders.

Dataset:

This project's most important deliverable, as the most significant hurdle for the widespread use of polymer informatics, especially in design, is the lack of databases, not a lack of machine learning algorithms.

The Data Acquisition was done from online sources. A excel spreadsheet of nearly 1000 datas were prepared with columns including Polymer name, Glass Transition temperature, Young’s Modulus, Tensile strength at break and Density of the polymeric material with their references.

Dataset Excel File:
polymer_dataset.xlsx

https://docs.google.com/spreadsheets/d/1Oiyu3AONPnXZmKOTHTTU3x-0vKfM24PR/edit?usp=share_link&ouid=109510204791998159370&rtpof=true&sd=true

Material Name	Density(gm/	Glass Transition Tempera	youngs mod	tensile stren	Reference li
Teijin Tenax® E TPCL PEEK-HTA40 ThermoPlastic Consol	1.76	143	42.0 - 60.0	42.0 - 60.0	https://www.matweb.com/search/DataSheet.aspx?MatGUID=
Teijin Tenax® A rTPWFCX PEEK-HTA40 Composite	1.76	143	40	40	https://www.matweb.com/search/DataSheet.aspx?MatGUID=
Teijin Tenax® E TPUD PEEK-IM565 Composite	1.78	147	15.7	15.7	https://www.matweb.com/search/DataSheet.aspx?MatGUID=
Overview of materials for Acrylonitrile Butadiene Styrene (A	0.882 - 3.50	105 - 109	0.778 - 21.2	0.778 - 21.2	https://www.matweb.com/search/DataSheet.aspx?MatGUID=
Overview of materials for Acrylonitrile Butadiene Styrene (A	0.890 - 1.50	108 - 127	1.70 - 2.70	1.70 - 2.70	https://www.matweb.com/search/DataSheet.aspx?MatGUID=
Overview of materials for Acrylonitrile Butadiene Styrene (A	1.00 - 3.50	105 - 108	1.40 - 2.55	1.40 - 2.55	https://www.matweb.com/search/DataSheet.aspx?MatGUID=
Overview of materials for Acrylonitrile Butadiene Styrene (A	1.01 - 1.20	108 - 109	1.00 - 2.65	1.00 - 2.65	https://www.matweb.com/search/DataSheet.aspx?MatGUID=
Overview of materials for Polycarbonate/ABS Alloy, Unreinf	0.960 - 1.50	104 - 161	1.67 - 11.1	1.67 - 11.1	https://www.matweb.com/search/DataSheet.aspx?MatGUID=
Overview of materials for Acrylic, General Purpose, Moldec	0.700 - 1.30	102 - 122	0.950 - 3.79	0.950 - 3.79	https://www.matweb.com/search/DataSheet.aspx?MatGUID=

Figure1: Snippet of Excel Database of polymer

Data Visualization

To better understand the property and density values, different plots were made.

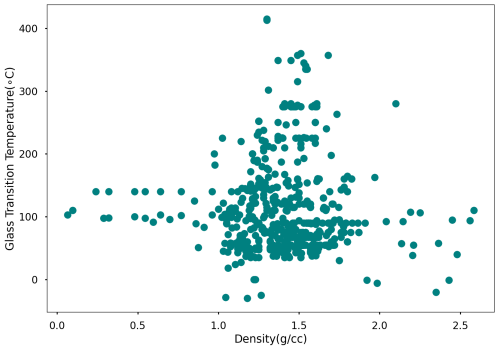

Figure 2: Scatter plot of Glass Transition Temperature (℃) v/s Density (gm/cc)

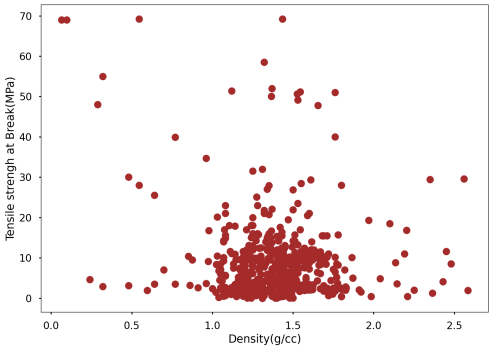

Figure 3: Scatter plot of Tensile Strength at Break(MPa) v/s Density (gm/cc)

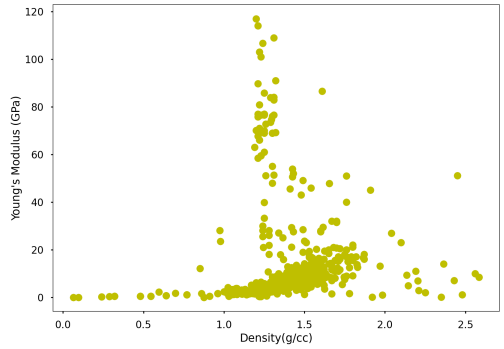

Figure 4: Scatter plot of Young's Modulus(GPa) v/s Density (gm/cc)

The Histogram plots were also Plotted for the different properties:

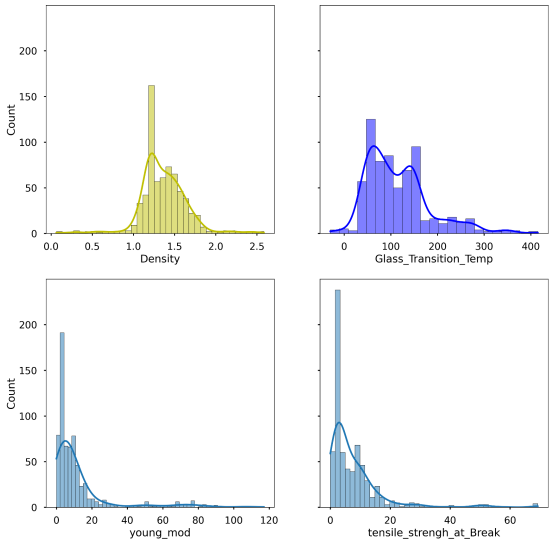

Figure 5: Histograms of properties of polymer.

Modeling

When you want to predict a narrow range of data, it is important to choose a model that is appropriate for the task at hand.

1. Linear regression: Linear regression is a simple and commonly used model for predicting numerical values within a narrow range. It assumes a linear relationship between the predictor variables and the response variable and can be used for both simple and multiple regression analysis.

2. Support vector machines (SVMs): SVMs are a powerful class of models that can be used for both regression and classification tasks. They work by finding the hyperplane that maximally separates the data into different classes or predicts the response variable within a narrow range.

3. Random forests: Random forests are an ensemble of decision trees that can be used for both regression and classification. They work by combining the predictions of multiple decision trees to reduce overfitting and improve the accuracy of the predictions.

4. XGBoost: XGBoost builds an ensemble of weak prediction models, typically decision trees, in a sequential manner. It combines the predictions from multiple models to create a strong predictive model. Each new model is trained to correct the errors made by the previous models, thereby improving the overall prediction accuracy.

The choice of the appropriate model depends on the specific task at hand, the nature of the data, and the desired level of accuracy. It is important to evaluate the performance of different models using appropriate metrics and choose the one that provides the best results for the task. Above are a few options that we applied for our Dataset.

Implementation

Dataset is trained using Machine Learning with Python programming language on the Google Colab Platform. The dataset was divided into two parts: training and testing in the ratio of 80: 20. The training set is used to train the machine AI model, and the testing set is used to test the implemented model, The different implemented model were-Linear regression, Support vector machines (SVMs), Random forests, XGBoost.

The True Density & predicted Density histograms were plotted, as Follows:

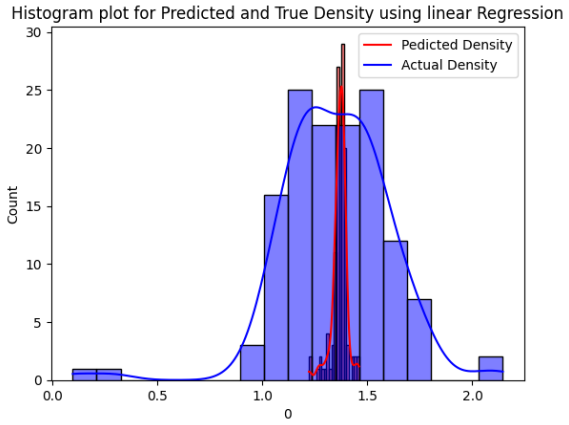

Figure 6: Histogram plot for Predicted and True Density using linear Regression

From linear Regression model, the Score is came out to be 7 % only, which is fairly low , hence it should not be considered for our Dataset.

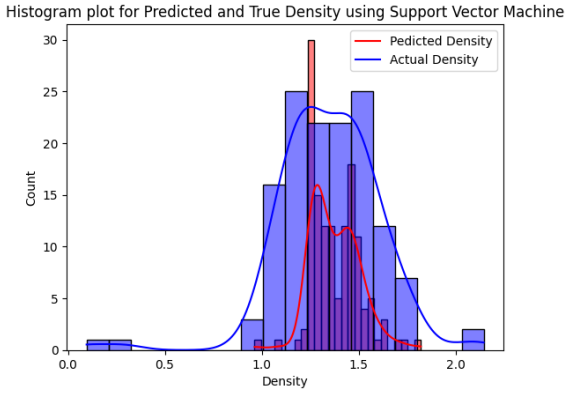

Figure 7:Histogram plot for Predicted and True Density using SVM.

From SVM model, the Score improved to 33 % only, But it is not good enough for prediction , hence it should not be considered for our Dataset.

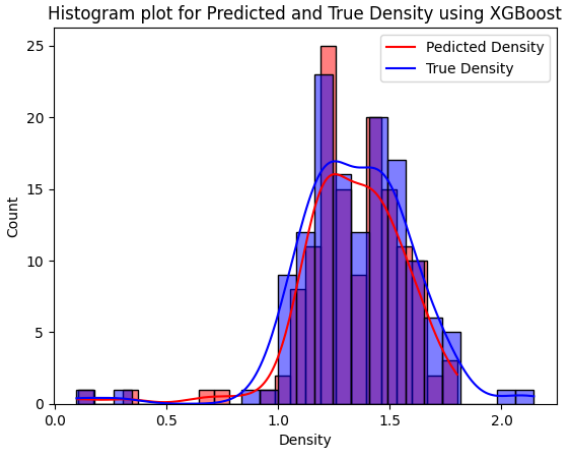

Figure 8: Histogram plot for Predicted and True Density using XGBoost

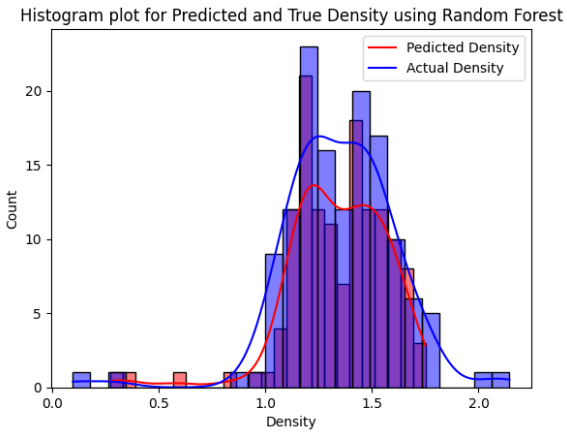

Figure 9: Histogram plot for Predicted and True Density using Random Forests

The result were Better in the Random Forests and XGBoost with 66% and 61% accuracy score respectively. Both model score are Comparable and upto the mark ,therefore can be considered for as Machine learning model for our Dataset. Linear regression and SVMs performed Poorly for our Datasets, which was quite obvious as they were based on linear relationship approach.

Density	Predicted_Density
1.23	1.2925
1.53	1.6009
0.961	0.99528
1.2	1.305
1.19	1.23585
1.25	1.12588
1.24	1.23325
1.33	1.07453
1.2	1.2

Figure 10: Snippet of CSV file generated from predicted Density comparison with Actual Density values using random forests

The Platform used for this Programming is Google Colab and Python is used as a programming language.

Conclusion:

From our Applied Machine learning model ,we have achieved an accuracy score of roughly 65%, on a database having nearly 1000 data points of polymer. it could be improved more by applying more suitable ML models. It would also be better for having more datasets of Polymers in our database, with exact and accurate values.

References:

- Polymer Informatic: Opportunities and Challenges, Debra J. Audus and Juan J. de Pablo ACS Macro Letters 2017 6 (10),1078-1082 DOI:10.1021/acsmacrolett.7b00228
- Machine learning in polymer informatics Wuxin Sha, Yan Li, Shun Tang, Jie Tian, Yuming Zhao, Yaqing Guo, Weixin Zhang, Xinfang Zhang, Songfeng Lu, Yuan-Cheng Cao, DOI: 10.1002/inf2.12167
- polymer Dataset excel sheet
- Data science for material innovation Colab file
- polymer_pred_using_random_forests.csv
- Search Engineering Material by Property Value
- (PDF) Machine-learning predictions of polymer properties with Polymer Genome (researchgate.net)
- [1] Cover Page Image
- pandas documentation — pandas 2.0.1 documentation (pydata.org)
- scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation
- XGBoost Documentation — xgboost 1.7.5 documentation
- sklearn.linear_model.LinearRegression — scikit-learn 1.2.2 documentation
- 1.4. Support Vector Machines — scikit-learn 1.2.2 documentation
- Matplotlib documentation — Matplotlib 3.7.1 documentation