**C#oders - Meta-Bizz Report**

**Team Members**

* 1. Name: Priyanshu Rao

Email ID: me210003056@iiti.ac.in

* 1. Name Atharva Mohite

Email ID: me200003016@iiti.ac.in

1. **Introduction**

**T**he Refractive Index (RI) of a material is a measure of how much the material bends light. It is defined as the ratio of the speed of light in a vacuum to the speed of light in the material.

The refractive index of a material can depend on several factors, including:

1. The composition and structure of the material

2. The wavelength of light passing through the material

3. The temperature of the material

4. The pressure on the material

5. The presence of impurities or defects in the material

6. The angle of incidence of light on the material

7. The polarization of the light incident on the material

8. The anisotropy of the material.

Our task is **to predict the refractive index of test materials** based on their given properties. This is a multi-variable regression task.

Machine Learning Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. The dependent variable, which is RI in our case, is continuous.

**2. Exploratory Data Analysis**

There are 137 features or independent variables in each data entry. We are given 659 data samples. The “Name” column consists of unique string-type data entries of names of the material. The output variable RI is of float data type.

| **Feature name** | **Data type (as originally given)** | **Description** |
| --- | --- | --- |
| Crystal Structure | Categorical | Seven different crystal structures: [1-Triclinic, 2-Monoclinic, 3-Orthorhombic, 4-Tetragonal, 5-Hexagonal, 6-Trigonal]. An ordered structure can affect the way that light interacts with the material and can therefore influence the refractive index. |
| Mohs Hardness | Continuous (float) | The mineral's hardness is defined by the Mohs scale. Mohs hardness is not related to the optical properties of a material, such as refractive index |
| Diaphaneity | Categorical | Diaphaneity refers to the transparency or translucency of a mineral, and it describes how well light can pass through a material. 1-Opaque, 2-Translucent, 3-Transparent  A material with a high refractive index bends light more than a material with a low refractive index, making it appear more optically dense. |
| Specific Gravity | Continuous (float) | Specific gravity is a measure of the material's density |
| Optical | Categorical | Optical properties are the characteristics of a material that describe its interactions with light, including its ability to transmit, reflect, absorb, or scatter light.1-Anisotropic, 2-Isotropic, 3-Uniaxial, 4-Biaxial  Refractive index is one of the key optical properties of a material |
| Dispersion | Continuous (float) | The change in the refractive index of light in the mineral as a function of frequency. |
| Chemical Composition | Discrete (int) | 126 columns of different chemical atoms/molecules like Hydrogen, Oxygen, Silicon, Calcium, Hydrated Water, and Iron |
| count | Discrete (int) | A redundant variable |
| Molar Mass | Continuous (float) | Molar mass is a physical property of a substance, while the refractive index is an optical property of a substance, they are not directly related. However, the composition of a substance can affect both properties. |
| Molar Volume | Continuous (float) | Molar volume is a physical property of a substance, while they have refractive index is an optical property of a substance; they are not directly related. However, the composition and structure of a substance can affect both properties. |
| Calculated Density | Continuous (float) | A material with a high calculated density will tend to have a higher refractive index because it will have a higher atomic number, which means that the material is more optically dense. |

The theoretical relationship between properties in the above table and the RI of the material mostly agrees with the experimental observations. The slope coefficients (or regression coefficients) for factors that vary proportionally with RI are positive, whereas those that vary inversely with RI are negative.

**3. Data pre-processing**

Data pre-processing generally includes a combination of data cleaning, data transformation, and data normalization techniques. We found no duplicate or missing entries in the provided dataset. In order to increase accuracy and for the sake of simplicity, it is essential to identify the least significant factors first and eliminate them.

For the 126 chemical composition variables, it was found that except for these six elements- Hydrogen, Oxygen, Silicon, Calcium, Hydrated Water, and Iron, the number of non-zero values for a variable is less than 200. Thus we considered only these six element columns.

It was observed that categorical variables - 'Crystal Structure,' 'Optical,' 'Dispersion,' etc. have classes with a representation of less than 5%. Hence we combine such classes/levels after considering their relatively small frequency.

Data transformation includes techniques such as feature scaling, encoding categorical variables, and creating new features from existing ones. The categorical feature classes were already assigned unique values. For all the continuous variables, as they have extremely high or low values (outliers), we preferred standardization over normalization. Among the two standardization techniques - z-score and min-max scaling, it was seen that the latter technique gave a better R2 score (see metrics section for R2-score).

After initial pre-processing, we trained several regression models on the training dataset (a randomly selected subset of 659 input data samples). We found that the Gradient boosting regressor gave the best R2 score. The predictive importance of a variable is directly proportional to the decrease in the R2 score over the test dataset when that variable is omitted during training and testing. This is a good measure of a factor’s significance, as some factors are categorical, and some are continuous.

Using this approach, the most significant factor is 'Optical,' followed by ‘Dispersion.’

There was no change in testing R2 score when 'Mohr hardness' and ‘count’ were separately omitted while training. Hence these factors were dropped.

**4. Building regression model**

The pre-processed training data now consist of 16 columns, including 1 RI column, and has 659 data samples.

This data is now split into training and testing data using sklearn’s train\_test\_split function. The test-train ratio is fine-tuned as a hyperparameter. The optimum test R2-score is achieved with a test-to-train ratio of 0.2.

Thus we have 527 train samples and 132 test samples.

We tried several regression models, out of which the ensemble regression models, like the Bagging Regressor, performed relatively well on the testing dataset. We chose to fine-tune only the two top-performing models - The Bagging Regressor and Gradient Boosting Regressor.

Sklearn library functions were used to instantiate the regression models.

**5. Gradient Boosting regressor**

# GradientBoostingRegressor (fine-tuned)

from sklearn.ensemble import GradientBoostingRegressor

model = GradientBoostingRegressor(loss='huber', n\_estimators=130, learning\_rate=0.1, max\_depth=1, alpha=0.9)

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

This code imports the GradientBoostingRegressor class from the sklearn.ensemble module, and creates an instance of the class with the specified parameters.

(GBR) is an ensemble learning method for regression problems, which combines multiple decision trees to improve the overall prediction performance. It is an extension of the popular gradient-boosting algorithm, which is used for classification problems.

The loss parameter is set to 'huber', a robust loss function that is less sensitive to outliers in the data.

The n\_estimators parameter specifies the number of trees in the ensemble model, and the learning\_rate parameter controls the step size of the gradient descent algorithm used to fit the model.

The max\_depth parameter sets the maximum depth of the trees in the ensemble, and the alpha parameter controls the regularization term in the model.

After instantiating the model, the code fits the model to the training data (X\_train, y\_train) and then uses the fitted model to make predictions on the test data (X\_test) stored in y\_pred.

GBR is widely used for regression problems due to its high predictive accuracy and its ability to handle large datasets and high-dimensional input spaces. However, it also has some limitations, such as its sensitivity to outliers and its tendency to overfit the training data if the number of trees is too large. In this case, you may want to try the next best-performing model ie Bagging Regressor.

**6. Performance metrics**

Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function. It measures how close a regression line is to a set of data points.

The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors over all instances in the test set. It is the arithmetic average of absolute errors.

R2 score is the percentage of the response variable variation explained by a linear model.

**Although R2 Score is a very popular metric used for evaluating the performance of linear regression models, MSE is better than R-squared if we are concerned about the model’s performance on unseen data.**

Nevertheless, in our case, both metrics agree on deciding the best-performing model.

**7. Results**

Following are the performance metrics of the various regression models on the test dataset.

| **Model name** | **Mean Squared error** | **Mean Absolute error** | **R2 score** |
| --- | --- | --- | --- |
| RandomizedSearchCV | 0.1340 | 0.1761 | 0.7953 |
| GridSearchCV | 0.1331 | 0.1747 | 0.7967 |
| DecisionTreeRegressor | 0.1660 | 0.1685 | 0.7464 |
| KNeighborsRegressor | 0.1245 | 0.1777 | 0.8098 |
| RandomForestRegressor | 0.1301 | 0.1739 | 0.8012 |
| BaggingRegressor (fine-tuned) | 0.1111 | 0.1604 | 0.8302 |
| Linear Regression | 0.1361 | 0.2201 | 0.7921 |
| SVM | 0.1364 | 0.2080 | 0.7917 |
| SGDRegressor | 0.1649 | 0.2468 | 0.7481 |
| BayesianRidge | 0.1444 | 0.2267 | 0.7795 |
| **GradientBoostingRegressor (fine-tuned)** | **0.1036** | **0.1338** | **0.8417** |
| XGBRegressor | 0.1454 | 0.1973 | 0.7779 |

**8. Conclusion**

In this study, different regression models is trained and evaluated on pre-processed data. It was observed that the R2 score improved by about 33% after combining levels with a frequency of less than 5%.

The most significant factor is **‘Optical’ or the ‘Optical Properties’**

The best model, based on its performance on the testing subset, is the **Gradient Boosting Regressor (84.17% R2 score).**