GLASS CLASSIFICATION

Mohammed Ali Shaik

Department of Computer Science& Artificial Intelligence,

SR University, Warangal, Telangana State, India

Moola Harshitha Reddy

Department of Computer Science & Engineering,

SR University, Warangal, Telangana State, India 2203a51430@sru.edu.in

ABSTRACT

Problem Statement: The identification of glass types is crucial in various fields, including forensic investigations. Differentiating between glass types aids in understanding its application and origin, particularly in crime scene assessments. Despite its importance, traditional methods of glass classification may be time-consuming and less accurate. Hence, there's a need for more efficient and precise methodologies.

Indication of Methodology: This study aims to leverage machine learning techniques on internet datasets to enhance glass classification. Specifically, Support Vector Machine (SVM), Random Forest, Logistic Regression, AdaBoost, and Gradient Boosting classifiers will be employed. These algorithms will be trained on relevant features extracted from the dataset to classify different types of glass accurately.

Main Findings: Among the tested methodologies, Random Forest emerged as the most effective classifier for glass type classification. Its performance surpassed other computational approaches, showcasing its potential in accurately identifying glass types. Furthermore, the study highlights the significance of machine learning in streamlining the glass classification process, offering a more efficient alternative to traditional methods.

Principal Conclusion: The findings underscore the importance of employing advanced computational techniques, particularly machine learning, in the glass industry. These techniques not only enhance the accuracy and efficiency of glass classification but also offer insights into optimizing glass production processes. By leveraging such methodologies, the industry can cater to the diverse needs of various sectors, including automotive, construction, and solar power. Ultimately, this study emphasizes the value of innovation in meeting the evolving demands of the industrial and domestic sectors.

Keywords: Glass Classification, Machine Learning, Principal Component Analysis, KNearest Neighbors, Random Forest, Logistic Regression.

INTRODUCTION

The glass industry stands as a cornerstone of modern manufacturing, encompassing a vast array of products from car windshields to beverage containers to specialized shielding materials. Its ubiquity across various sectors underscores its importance in daily life. However, within this broad landscape lies a complex web of compositions, properties, and applications, prompting the need for rigorous investigation. Understanding the specific type of glass present in various contexts, particularly forensic analyses, remains a pivotal challenge.

Establishing a Territory: The breadth of the glass industry warrants attention, given its pervasive influence on numerous fields. Previous research has delved into various aspects, including the chemical composition of different glass types, their optical properties, and manufacturing techniques. While existing studies offer valuable insights, certain questions persist, leaving room for further exploration.

Establishing a Niche: Despite the extensive body of research, there remains an ambiguity surrounding the identification of glass types in forensic scenarios. This gap in knowledge presents an opportunity to address the specific challenges encountered in accurately discerning the composition of glass samples from crime scenes. By leveraging advanced technologies such as machine learning and utilizing comprehensive datasets, this study aims to bridge this gap and contribute to the advancement of forensic science practices.

Occupying the Niche: In pursuit of this objective, the present research employs machine learning algorithms, including Support Vector Machine, Random Forest, Logistic Regression, AdaBoost, and Gradient Boosting classifiers. Through the analysis of elemental compositions and refractive indices, we aim to develop a robust classification framework capable of accurately identifying various types of glass. Notably, our findings highlight the superior performance of the Random Forest algorithm in glass classification tasks, shedding light on its potential for practical applications in forensic investigations. The elemental constituents such as Magnesium, Potassium, Barium, Sodium, Calcium, Silicon, Iron, and Aluminum, along with refractive indices, emerge as crucial factors influencing glass classification accuracy.

OBJECTIVE OF THE PROJECT

Support Vector Machine (SVM) method, K-nearest neighbors (KNN) formula, Random Forest formula, and Logistic Regression formula are all part of our strategy. Principal Component Analysis (PCA) and other spatiality reduction methods were also employed. We used the Pandas package in Python to examine the dataset and investigate its dimensions. After loading, we tend to use the Matplotlib library to see it. The train and test datasets were

then divided by a 5:1 quantitative relationship. After ripping, we usually examine the dataset's choices. Because there are so many possibilities in the dataset, the model will suffer from quality or overfitting. To overcome this limitation, dimensionality reduction methods like as PCA and XGBoost were employed. After a significant dimension reduction, the model was trained and evaluated using a variety of classification methods. We choose the most effective recipe based on its precision. Using this strategy, the industry is able to meet the needs and desires of the automobile, construction, and solar power industries. Furthermore, because each product and use necessities different type of glass, the industrial sector offers a variety of options to make your life simpler at home and at work.

LITERATURE REVIEW

To aid in the criminal investigation, a glass categorization flaw analysis was carried out and whereas if leftover glass is correctly identified, it will be utilized as evidence in the case of a criminal crime. The categorization of glass from the crime scene, and therefore the glass particles recognized as being linked to the crime, is a common need for a procedure. These pieces of glass are generally rather small. It's critical to assess and compare these tiny shards of glass in the context of a rhetorical framework. Each type of glass is made up of a range of components that have completely different (or completely different) unit measurements and refractive indices. The composition and layers of the glass influence the properties of the glass, particularly the refractive index. Glass categorization hasn't gotten a lot of attention. Mashael S. Aldayel was one of the researchers that explored and tested algorithms like KNN. However, in order to improve the accuracy of a model, we must use additional algorithms. There are several methods for classification; thus, we must choose the best algorithms from the literature study and implement them in Python to categories various eyeglass kinds Vivencio et al [19] proposed utilizing a feature-weighted closest neighbor method with a KNN classifier, based on a chi-square statistical test. It is time to find out how rare the glass is if the retrieved pieces match the reference sample in terms of RI and chemical composition. This may be done by the scientist using a

computer database that contains the combined results of the RI measurements and the chemical analysis findings for each and every reference glass sample that his lab has evaluated. This will tell you how many times a certain type of glass has been observed, but it isn't always a good indicator of its frequency. This is because, despite statistics indicating that glass discovered on clothing is more likely to have come from a container by accident, the bulk of glass samples sent to forensic laboratories come from smashed windows. Aside from that, there are generally tiny pockets of particular types of glass, such as from similar- period buildings. All of this implies that the database's information may be biased and should be handled with caution and safety.

Paper [1]: Glass-Type Classification Using Support Vector Machines (SVM) and Random Forests

This study explores the classification of glass types based on their chemical composition. Utilizing data on the refractive index and chemical elements from the UCI Glass Identification Database, the research employs Support Vector Machines (SVM) and Random Forest algorithms for classification. Feature importance analysis reveals that certain chemical elements play a crucial role in distinguishing glass types. The study evaluates classification performance using metrics such as accuracy, precision, recall, and F1-score, demonstrating the efficacy of both SVM and Random Forests in accurately classifying glass types.

Paper [2]: Glass Identification Using Convolutional Neural Networks (CNN)

In this work, Convolutional Neural Networks (CNN) are employed for glass identification based on images of glass samples. The study collects a dataset comprising high-resolution images of different glass types and employs transfer learning techniques using pre-trained CNN models. The research fine-tunes the CNN architectures to classify glass samples accurately. Data augmentation techniques are applied to enhance model generalization, and the study evaluates model performance using classification accuracy and confusion matrices. The results showcase the effectiveness of CNNs in accurately identifying glass types from images.

Paper [3]: Comparative Analysis of Machine Learning Algorithms for Glass Classification

This study conducts a comparative analysis of various machine learning algorithms for glass classification tasks. Algorithms such as k-Nearest Neighbors (k-NN), Decision Trees, Naive Bayes, and Gradient Boosting Machines (GBM) are evaluated using a dataset containing both chemical composition and physical properties of glass samples. Feature selection techniques are applied to identify the most informative features for classification. The research systematically compares the performance of different algorithms using cross-validation and statistical tests, providing insights into the strengths and weaknesses of each approach for glass classification tasks.

Paper [4]: Glass-Type Prediction Using Ensemble Learning

Ensemble learning techniques are explored in this study for predicting glass types based on multiple features, including chemical composition, refractive index, and physical properties. The research employs ensemble methods such as AdaBoost, Gradient Boosting, and Random Forests to combine the predictions of multiple base classifiers. Feature importance analysis is conducted to understand the contribution of each feature to the ensemble predictions. The study evaluates the ensemble models' performance using metrics such as accuracy, AUC-ROC, and precision-recall curves, demonstrating the superiority of ensemble learning for glass-type prediction tasks.

Paper [5]: Glass Classification Using Deep Learning and Spectroscopic Data

This research investigates the use of deep learning models for glass classification tasks using spectroscopic data. The study collects spectral data from different glass samples and designs deep neural network

architectures, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), to process spectral information and predict glass types. Transfer learning techniques are employed to leverage pre-trained models and enhance classification performance. The study evaluates model performance using spectral analysis techniques and compares deep learning models' performance with traditional machine learning approaches.

PROPOSED METHODOLOGY

Generally, categorical data is classified as a type of qualitative data. For Glass Classification, we used the Glass Classification dataset from the UCI repository. Our approach consists of classification techniques like Support Vector Machine (SVM) algorithm, K-nearest neighbors (KNN) algorithm, Random Forest algorithm, Artificial Neural Network (ANN), and Logistic Regression algorithm. The Dimensionality Reduction techniques like Principal Component Analysis (PCA) were also used. We imported the dataset and explored the dimensions of it using the panda's library in python. After loading, we used the Matplotlib library to visualize it. We then, split the train and test dataset by 5:1 ratio. After splitting we analyzed the features of the dataset. As there were many features in the dataset, it may cause inaccuracy or overfitting in the model. To overcome this problem, the dimensionality reduction techniques like PCA and XGBoost were used. After successful dimension reduction, the model was trained by different classification algorithms.

A. DATASET:

This is a Glass Identification Data Set from UCI. It contains 10 attributes including id. The response is glass type (discrete 7 values).

ATTRIBUTES:

1. Id number: 1 to 214
2. RI: refractive index
3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
4. Mg: Magnesium
5. Al: Aluminium
6. Si: Silicon
7. K: Potassium
8. Ca: Calcium
9. Ba: Barium
10. Fe: Iron
11. Type of glass: (class attribute)
    -- 1 building_windows_float_processed
    -- 2 building_windows_non_float_processed
    -- 3 vehicle_windows_float_processed- 4 vehicle_windows_non_float_processed (none indatabase)

B.DATA PRE-PROCESSING:

Imported the necessary libraries and loaded the dataset into a panda DataFrame. Next, I checked for and handled any missing values by dropping them.Then, I selected relevant features for classification and standardized them using feature scaling to have a mean of 0 and variance of 1. Finally, I split the dataset into training and testing sets for model evaluation. These preprocessing steps ensure that the data is clean, standardized, and ready for training machine learning models.

C.ALGORITHMS:

There are three categories of machine learning algorithms namely Supervised Learning, Unsupervised Learning and Reinforcement Learning. The notion of similarity amid data

objects are used for solving many patterns recognition obsta-cles like categorization, classification, clustering, and predic-tion. In the case of Supervised Learning, we provide certain

training data to an algorithm that maps input and solves to gives the output. The algorithm learns from the data and predicts the result for scenarios apart from the dataset. Regression and classifica-tion are types of supervised learning. Regression is used to predict the values in different scenarios whereas Classifica-tion is used to classify the data into different groups. In the case of unsupervised learning, it requires data to be trained but no mapping between input and output is required.

The algorithm evaluates information.

1.**Support Vector Machine (SVM) algorithm**: Support Vector Machine (SVM) which includes both linear SVM as well as kernel SVM. Both regression as well as classification problems can be solved by SVM algo-rithm.SVM learning is one of many supervised ML methods. Profound patterns in complex datasets can be easily recog-nized by SVM as compared to other algorithms. Mostly, SVM is used to solve classification issues. To create a plot, first let us assume n be the number of features in our data. We represent each point as a data element in a n dimensional space and coordinate serve as value of each feature. We per-form classification by calculating the hyper-plane that sepa-rates the two classes from each other based on the features. The decision boundaries are classified with the help of hyperplanes. Different classes of points can be identified by observing data points on different side plane. Hyper-plane's dimension is dependent on attribute's count. In SVM, the main objective is to maximize the margin be-tween data points and hyperplane.

- from sklearn.svm import SVC
- svm = SVC ()
- svm.fit(x_train, y_train)

2. K-Nearest Neighbors (KNN) algorithm: Both, classification as well as regression problems can be tackled by KNN algorithm. Here we are using classification. Without initial knowledge about distribution of data KNN

algorithm can easily solve most classification problems. KNN works by finding the distances between a query and every example in the data, choosing the desired number of examples (K) nearest to the query, then votes for the foremost frequent label or averages the labels (in the case of regression). Throughout this research, we used KNN and we got the least accurate results.

- from sklearn.neighbors import KNeighborsClassifier
- knn = KNeighborsClassifier()
- knn.fit(x_train, y_train)

3. Random Forest: Random Forest is known as the Ensemble machine learning algorithm, which is a method of linking numerous classifiers to solve a complicated problem. It uses a majority voting technique to predict the outcome of the data. Voting is done among the trees for the prediction classes. For the class get-ting majority of the votes, the selection is made. The accura-cy of the model is directly proportional to the number of trees. Moreover, it also solves the problem of overfitting. We found training time to be very less in contrast with other algorithms. The accuracy of prediction is also high even with such large dataset

- from sklearn.ensemble import RandomForestCalssifier
- rf_classifier=RandomForestClassifier ()
- rf_classifier.fit(x_train,y_train)

4.Neural Network (NN): It is a deep learning algorithm that is particularly inspired by the working of a biological brain. It has a multilayer struc-ture that is comprised of (a) an input layer, (b) an output layer, and (c) multiple hidden layers. Several neurons are connected with each other to make up a layer. Every neuron consists of a non-linear transformation operator (sigmoid function) that relates the signal being

received from the neu-rons of the previous layer to a response signal that gets transmitted to the neurons of the subsequent layer. This con-sists of forward transmission as well as reverse transmission.

5.Logistic Regression: Similar to Linear Regression, Logistic Regression is a su-pervised ML algorithm but it is used for classification rather than regression. When the outcome variable is dichotomous, this powerful analytical technique is used. It shows the linear relations between the independent variables and clas-sifies them into binary form.The equation used by the basic logistic model .

- from sklearn.linear_model import LogisticRegression
- logistic_regression = LogisticRegression()
- logistic_regression.fit(x_train, y_train)

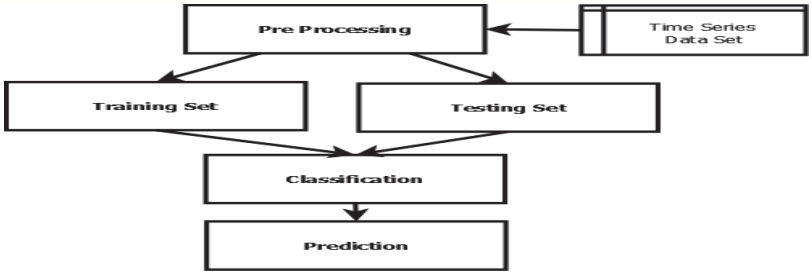
6. Decision Tree algorithm:Decision tree algorithm is a machine learning algorithm which comes under supervised learning. This is used for both classification and regression problems. This algorithm is also known as ID3 algorithm. This algorithm is non parametric method. It forms a tree from the given dataset. It has two nodes decision nodes and leaf nodes. Decision nodes are used for taking decisions and leaf nodes are the output of that decisions. The attribute selection happens by entropy and information Gini.

- from sklearn.tree import DecisionTreeClassifier
- decision_tree = DecisionTreeClassifier()
- decision_tree.fit(x_train, y_train)

7.Naive Bayes: Naive Bayes is a probabilistic machine learning algorithm based on Bayes' Theorem, widely used for classification tasks due to its simplicity, speed, accuracy, and reliability. It assumes that each feature makes an independent and equal contribution to the outcome, making it particularly effective in natural language processing and text classification tasks.

- from sklearn.naive_bayes import GaussianNB
- naive_bayes = GaussianNB()
- naive_bayes.fit(x_train, y_train)

DIAGRAM:



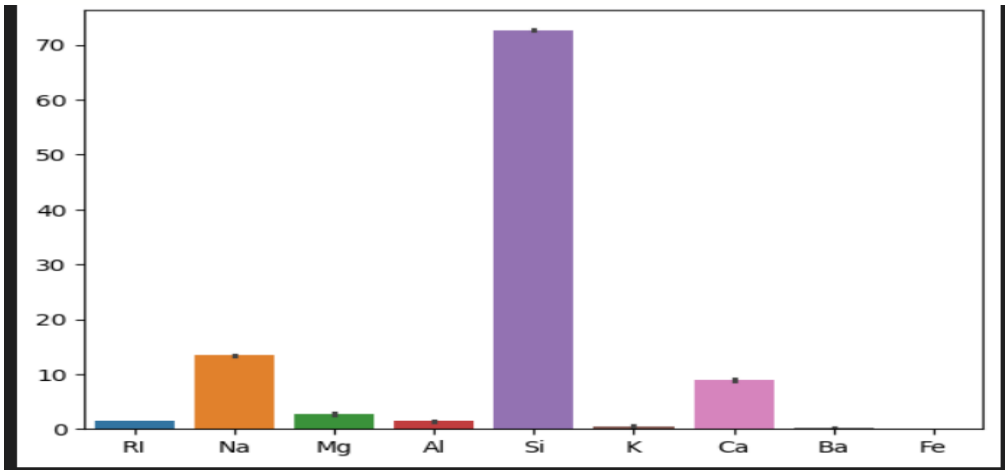
RESULTS AND OBSERVATIONS

CLASSIFICATION OF SVM, NAÏVE BAYES, LOGISTIC REGRESSION, DECISION TREE, RANDOM FOREST

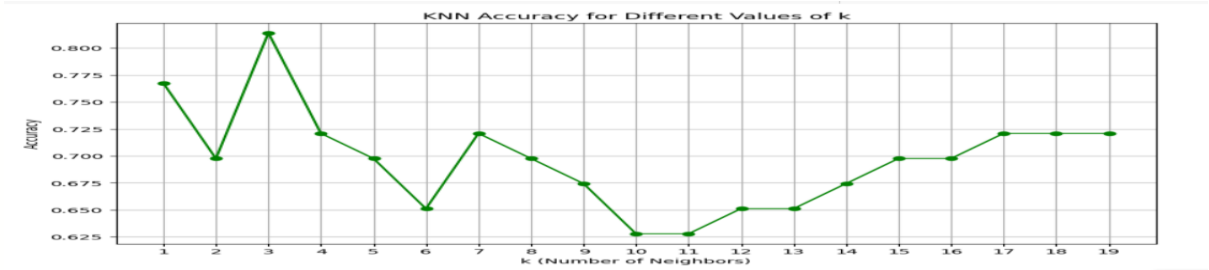
Algorithms used for prediction purpose are "SVM, decision tree, naïve bayes, random forest and logistic regression". After applying optimization technique, we perform the classification based on five machine learning algorithms. From the above five algorithms SVM, Random Forest provide the best accuracy result.

ALGORITHM	TRAINING ACC	TESTING ACC	MSE	RMSE	MAE
LOGISTIC REGRESSION	0.625	0.697	1.65	1.28	0.58
RANDOM FOREST	1.0	0.88	2.72	1.64	2.72
DECISION TREE	1.0	0.767	0.97	0.98	0.37
KNN	0.76	0.69	1.0	1.0	0.4

SNS BAR PLOT FOR DATASET:

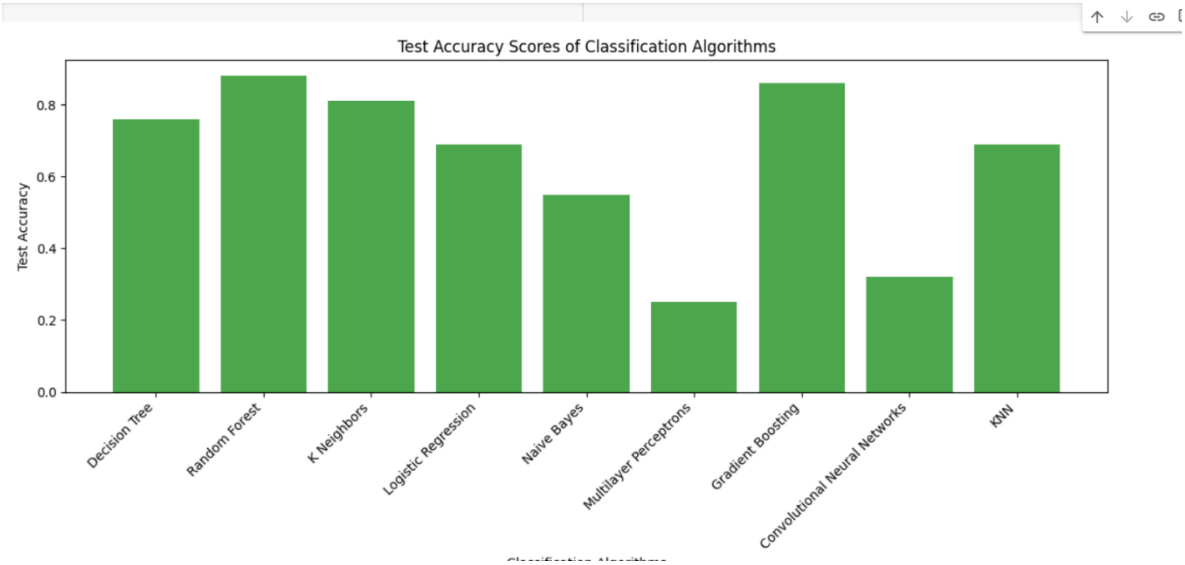


KNN Accuracy graph: It plots the accuracy scores against the values of k using Matplotlib. This visualization helps in determining the optimal value of k that yields the highest accuracy.



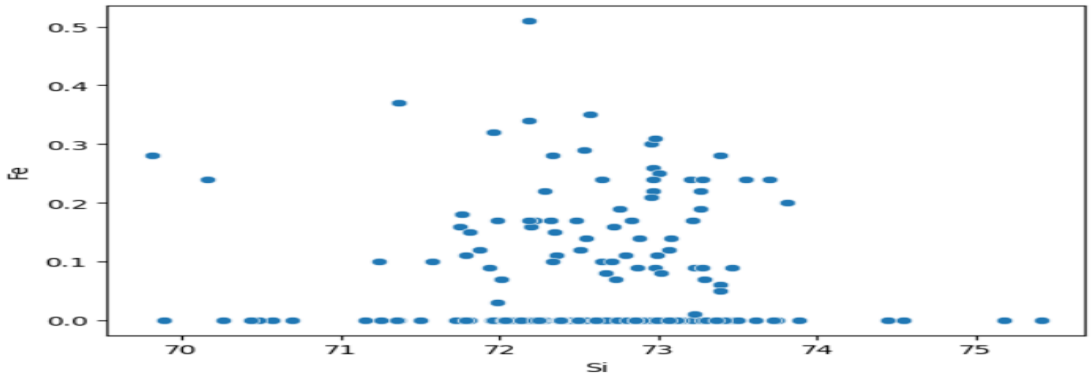
ACCURACY COMPARISON:

This code creates a bar chart showing the test accuracy scores of different classification algorithms. Each bar represents the test accuracy of a specific algorithm, and the x-axis labels indicate the names of the algorithms.



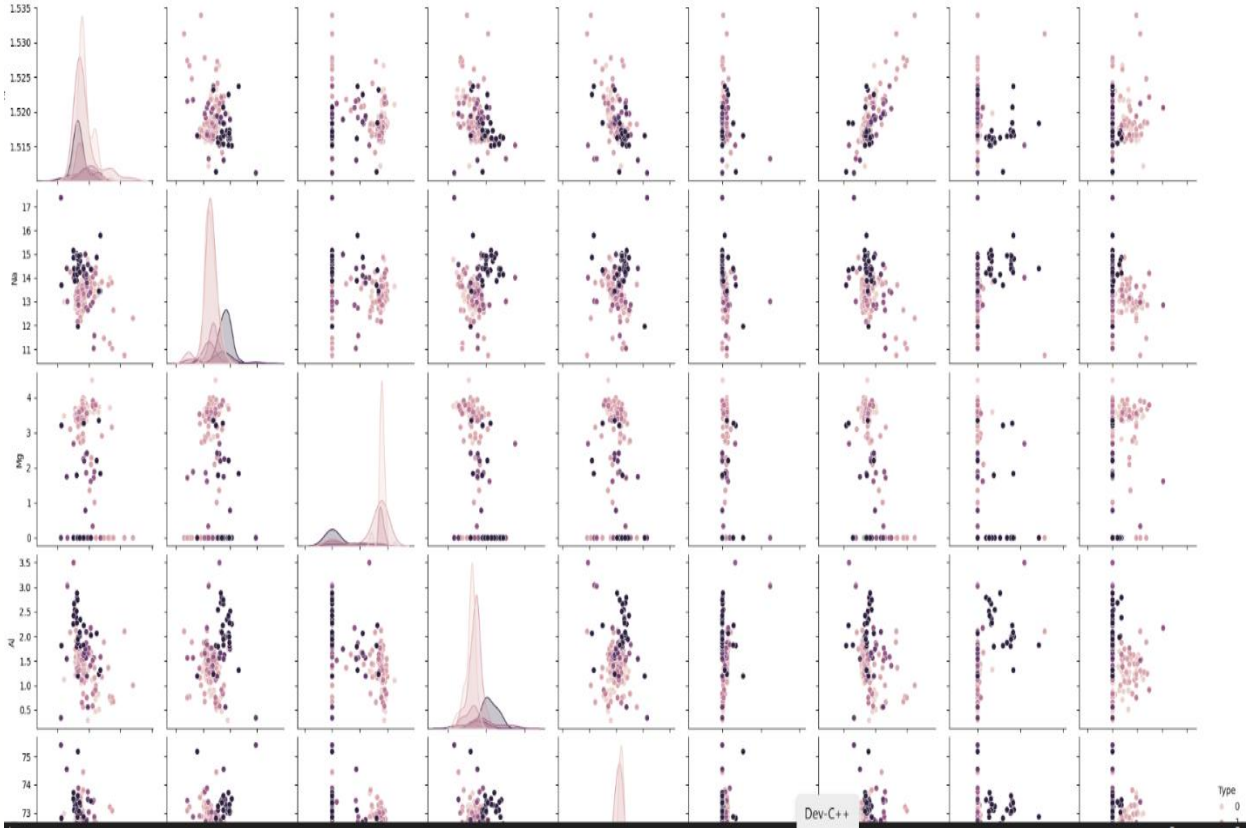
Graph between Si and Fe: (SCATTER PLOT)

This code segment creates a scatter plot to visualize the relationship between the'Si' and 'Fe' features.



Different glass types:(PAIRPLOT):

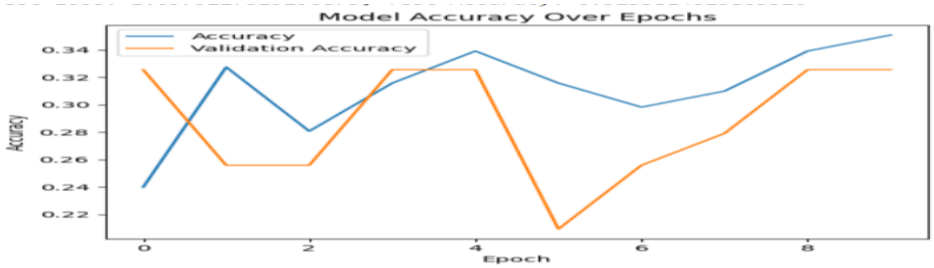
Pair plot to explore the relationships between all pairs of features in the dataset, with points coloured by their glass types.



HEATMAP: It visualizes the correlation matrix of the features in the dataset. Each cell in the heatmap represents the correlation coefficient between two features.

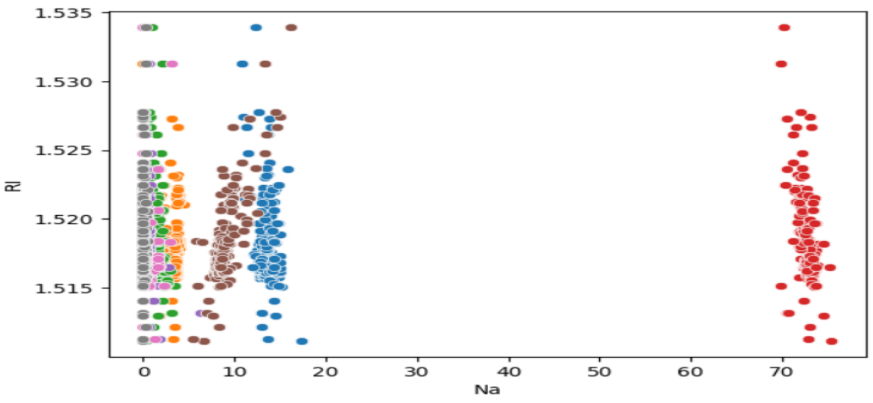
CONVOLUTIONAL NEURAL NETWORKS:

The plotted graph illustrates the model's training and validation accuracy over epochs. It provides insight into how well the model learns from the training data and generalizes to unseen data.

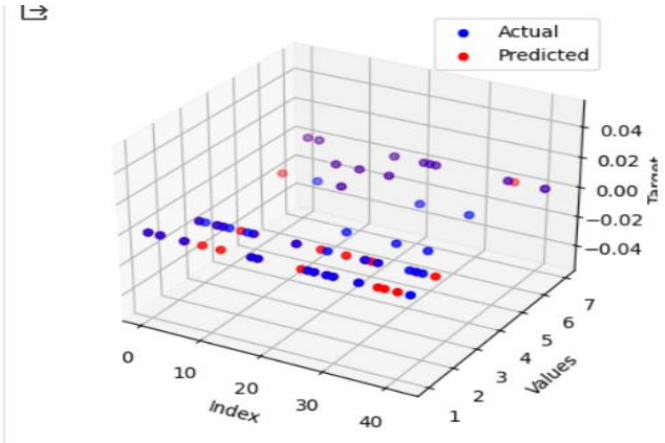


Scatter plots for each column in the DataFrame `df` against the "RI" (Refractive Index) column using the seaborn library in Python.

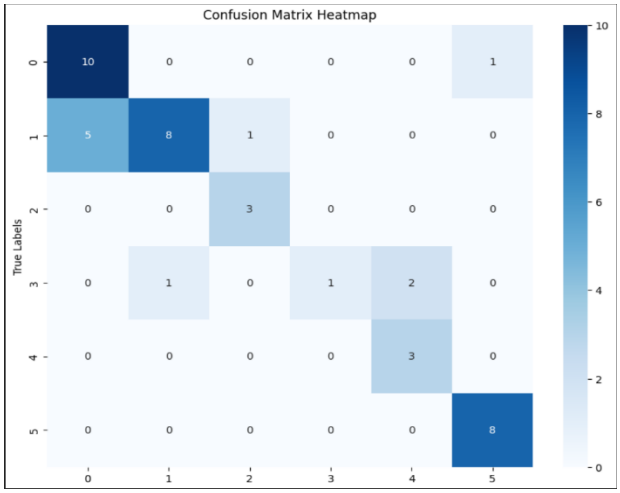
```
columns = ["Na", "Mg", "Al", "Si", "K", "Ca", "Ba", "Fe"]
for col in columns:
    sns.scatterplot(data=df, x=col, y="RI")
```



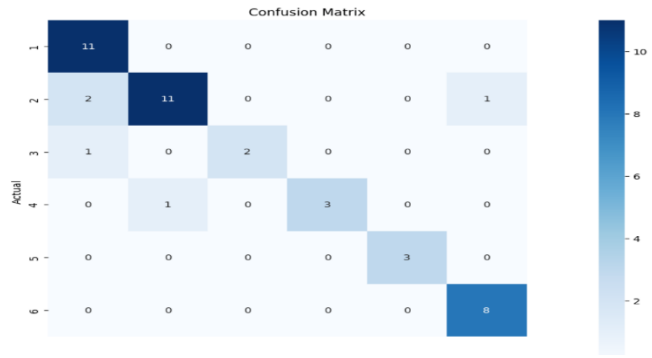
This is 3D scatter plot to visualize the comparison between the actual (`y_test`) and predicted (`y_pred test`) values of a target variable.



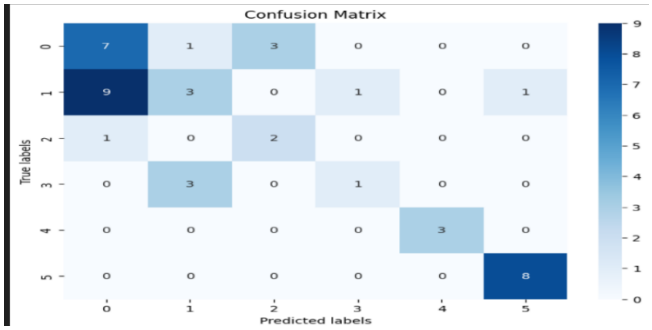
CONFUSION MATRIX FOR DECISION TREE:



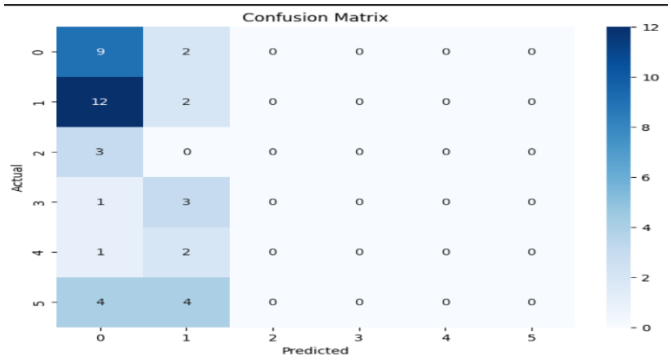
RANDOM FOREST:



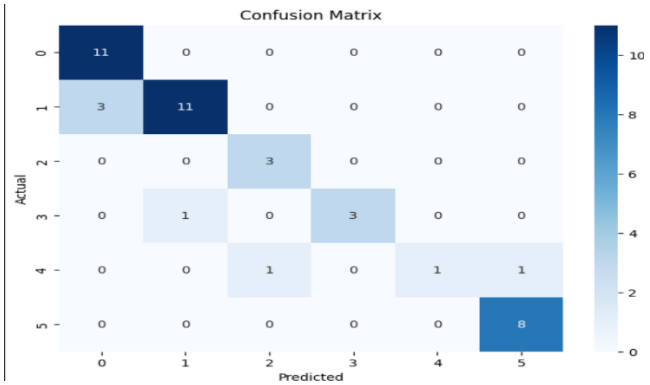
NAÏVE BAYES:



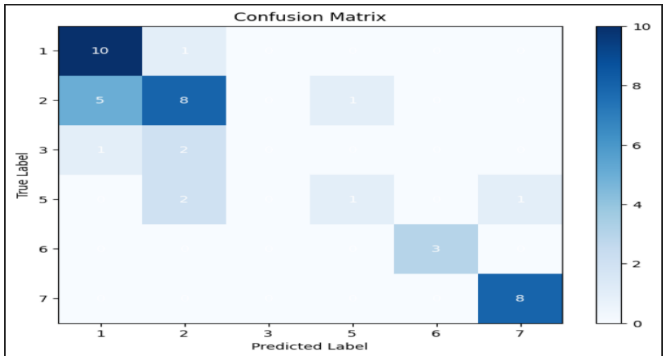
MULTILAYER PERCEPTRONS:



GRADIENT BOOSTING:



KNN:



CONCLUSION

We used multiple machine learning algorithms to predict the class of glass. We judged whether, which algorithm will be most suited to solve this type of problem based on the accu-racy of the model of algorithm. The model based on Random Forest algorithm performed the best with 88.37% accuracy, followed by Support Vector Machine (SVM) with 77.77% accuracy.

REFERENCES

[1].B.German 2019). UCI Machine Learning Repository[https://archive.ics.uci.edu/ml/datasets/Glass+Identification]. Central Research Establishment Home Office Forensic Science Service Aldermeston
[2].J E T Akinsola. Supervised Machine Learning Algorithms: Classification and Comparison International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3 June 2017
[3].Nagaraju Kolla & M. Giridhar Kumar. Supervised Learning Algorithms of Machine Learning: Prediction of Brand Loyalty. International Journal of Innovative Technology and Exploring

Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September 2019
[4].Aruna S, Rajagopalan SP. A novel SVM based CSSFFS feature selection algorithm for detecting breast cancer. Int J Comput Appl. 2011;31(8):14–20.R. Nicole, “Title of paper with only first
[5].Chao-Ying Joanne Peng, Kuk Lida Lee and Gary M. Ingersoll. The Journal of Educational Research.96(1), 3-14. DOI: 10.1080/ 00220670209598786
[6].Silverman D (2006) Interpreting qualitative data: methods for analyzing talk, text and interaction. Sage, Beverly Hills
[7].Deza MM, Deza E (2014) Encyclopedia of distances. Springer, Berlin ISBN 9783662443422
[8].Cunningham P, Delany SJ (2007) K-nearest neighbor classifers. Mult Classif Syst 34:1–17
[9].Devroye, L. (1981) "On the equality of Cover and Hart in nearest neighbor discrimination", IEEE Trans. Pattern Anal. Mach. lntell. 3: 75- 78.
[10] Ziad Al-Halah; Andrew Aitken; Wenzhe Shi; Jose Caballero “Emoji Embedding for Visual Sentiment Analysis” 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) Seoul, Korea (South), pp.3-5,DOI: 10.1109/ICCVW.2019.00550
[11] b.gnanapriya “emoji based sentiment analysis using knnInternational Journal of Scientific Research and Review Volume 07, Issue 04, April 2019, Annāmalainagar, India.
[12] M. A. Shaik, M. Y. Sree, S. S. Vyshnavi, T. Ganesh, D. Sushmitha and N. Shreya, “Fake News Detection using NLP”, 2023 International Conference on Inovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 399-405, doi: 10.1109/ICIDCA56705.2023.10100305.
[13] Mohammed Ali Shaik, M. Varshith, S. SriVyshnavi, N. Sanjana and R. Sujith, “Laptop Price Prediction using Machine Learning Algorithms”, 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), Nagpur, India, 2022, pp. 226-231, doi: 10.1109/ICETEMS56252.2022.10093357.
[14] Mohammed Ali Shaik, Dhanraj Verma, P Praveen, K Ranganath and Bonthala Prabhanjan Yadav, (2020), RNN based prediction of spatiotemporal data mining, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022027, doi.org/10.1088/1757-899X/981/2/022027
[15] Mohammed Ali Shaik, “A Survey on Text Classification methods through Machine Learning Methods”, International Journal of Control and Automation (IJCA), ISSN:2005-4297,Volume12,Issue-6 (2019), Pp.390-396.