Glass Classification based on Machine Learning Algorithms

Harshit Mathur, Aditya Surana

Abstract: ***Glass Industry is considered one of the most important industries in the world. The Glass is used everywhere, from water bottles to X-Ray and Gamma Rays protection. This is a non-crystalline, amorphous solid that is most often transparent. There are lots of uses of glass, and during investigation in a crime scene, the investigators need to know what is type of glass in a scene. To find out the type of glass, we will use the online dataset and machine learning to solve the above problem. We will be using ML algorithms such as Artificial Neural Network (ANN), K-nearest neighbors (KNN) algorithm, Support Vector Machine (SVM) algorithm, Random Forest algorithm, and Logistic Regression algorithm. By comparing all the algorithm Random Forest did the best in glass classification.***

Keywords: ***Glass Classification, Machine Learning, Support Vector Machine (SVM) algorithm, K-nearest neighbors (KNN) algorithm, Random Forest algorithm, Artificial Neural Network (ANN), Logistic Regression algorithm.***

I. INTRODUCTION

A glass classification problem study was conducted to assist within the criminal investigation. In the event of a crime, the remaining glass can be used as evidence if properly identified. The constant need for a lawsuit is the categorization of glass from the crime scene and the glass particles found associated with the crime. These glass particles tend to be very small. There is a need to view and compare these small pieces of glass that will be important during a forensic context [3]. Every sort of glass is made of different elements with different unit measurements and different Refractive Index. The property of the glass, particularly the refractive index, depends on the composition and treatments of the glass [2]. The elements used in making different types of glasses are Sodium(Na), Calcium(Ca), Magnesium(Mg), Barium(Ba), Silicon(Si), Aluminium (Al), Iron(Fe), Potassium(K). Furthermore, the Refractive Index(RI) also plays an important role in differentiating glass and its use. The glasses in this dataset are of 7 types according to their use. Building Windows Float glasses, Building Windows non-float glass, vehicle window float glass, containers, tableware, and headlamps. Float glass is a sheet of glass made by floating molten glass on a bed of liquid metal. This glass is generally premium quality with no finishing required and has structural plasticity during manufacturing. This method gives the sheet uniform thickness and very flat surfaces. To classify the glasses, various machine learning algorithms can be used to train and test the data.

II. WORK DONE

Not much work has been done on the classification of glass. Some researchers Mashael S. Aldayel [18] tried and tested algorithms like KNN. But to give more accuracy to a model, we need to implement more algorithms. There are plenty of algorithms for classification, so we need to select some better algorithms from Literature review, and implement better algorithms in python for classifying different types of glasses. Vivencio et al. [19] suggested a feature weighting nearest neighbor method based on a chi-square statistical test, to be used in association with a KNN classifier.

III. METHODOLOGY

Generally, categorical data is classified as a type of qualitative data [6]. For Glass Classification, we used the Glass Classification dataset from the UCI repository [1] and used Jupyter Notebook as our IDE. 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 pandas 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 and tested. We selected the best algorithm based on its accuracy. Figure 1 shows the steps followed to shortlist the algorithm.

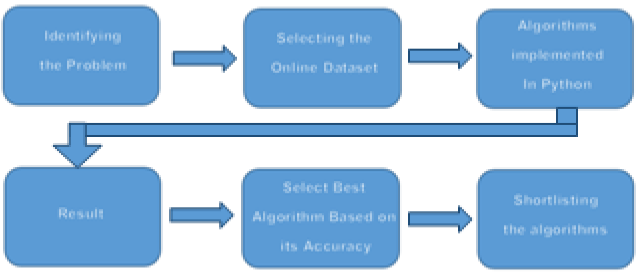
Identifying the Problem
Selecting the Online Dataset
Algorithms implemented In Python
Result
Select Best Algorithm Based on its Accuracy
Shortlisting the algorithms

Figure 1 : Methodology to be followed to solve this problem

A. Dimensionality Reduction

Reduction of set of random variables into set of principal variables is done by the process called Dimensionality Reduction.

Revised Manuscript Received on August 29, 2020.

Harshit Mathur, Department of Computer Science, Jaipur Engineering College and Research Centre, Jaipur, India. E-mail: harshitmathur10@gmail.com

Aditya Surana, Department of Computer Science, Jaipur Engineering College and Research Centre, Jaipur, India. E-mail: to.adityasurana@gmail.com

Retrieval Number: 100.1/ijitee.H6819069820
DOI: 10.35940/ijitee.H6819.0991120

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

It helps us in getting 2-D data so we can enhance our visualization of ML based models by making prediction regions vs prediction boundary curve for individual model. It removes the less significant features and focuses more on the significant features.

Dimension Reduction also helps in removing unnecessary or disorderly data occurrences in our training subset of data.

B. Features Selection

Feature selection is a method of deriving the subset of primitive features in various ways depending on the information they provide ,accuracy and forecasting errors.

C. Features Projection

For transforming the high dimensional data to low dimensional data, we use linear and non-linear reduction techniques. The transformation is done on the basis of the relationship among the features in a dataset. In this research, the dataset of 10 attributes, which are all related to the cell parameters is used. We applied PCA technique to derive components from the dataset.

D. Principal Component Analysis (PCA)

Principal Component Analysis is a analytical method to derive the most relevant features using the covariance matrix of the dataset. It transforms a large number of dimensions to 2-3 dimensions. It is used to tackle the problem of many features at once. PCA is generally used to reduce number of variables from common covariance datasets or various types of data such as discrete, compositional, ordinal, binary and symbolic data [13,14,15].

It is way of summarizing data without loosing any important information from original data. To maximize the variance, it converts the high dimensions to less attributes. It extracts x autonomous variables from y autonomous variables such that x<=y. The eigenvectors are calculated with the help of the covariance matrix generated for the dataset. The principal elements are those eigenvectors that have the biggest eigenvalues and these can be used to rebuild a large piece of original data's variance. After applying PCA, the dataset is ready for data mining and further machine learning techniques.

E. Model Selection

There are three categories of machine learning algorithms namely Supervised Learning, Unsupervised Learning and Reinforcement Learning. The notion of similarity amid data objects is used for solving many pattern recognition obstacles like categorization, classification, clustering, and prediction. 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 classification are types of supervised learning. Regression is used to predict the values in different scenarios whereas Classification 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 which is neither labeled nor classified, without any specific direction . This algorithm generates various types of clusters based on the data and then predicts the cluster to which the data belongs.

Reinforcement learning doesn't require data as a requirement. The model can be prepared even without any data. This algorithm learns from experience. It runs the algorithm on the given data process and generates the result. After that, it takes feedback on whether the result was correct or not and then performs a certain action based on the feedback.

Out of these, we used Classification Supervised Learning for our model. K-nearest neighbors (KNN) algorithm, Support Vector Machine (SVM) algorithm, Random Forest algorithm, Logistic Regression and Artificial Neural Network (ANN) algorithm were used for glass classification such that we can choose the best algorithm out of these (i.e. algorithm that has the best accuracy).

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 algorithm.SVM learning is one of many supervised ML methods. Profound patterns in complex datasets can be easily recognized 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 perform classification by calculating the hyper-plane that separates 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 of plane. Hyperplane's dimension is dependent on attribute's count.

In SVM, the main objective is to maximize the margin between data points and hyperplane.

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.[9-12]. 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.

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 getting majority of the votes, the selection is made. The accuracy of the model is directly proportional to the number of trees. Moreover, it also solves the problem of overfitting. We found the training time to be very less in contrast with other algorithms. The accuracy of prediction is also high even with such large dataset. Hence, we tried this algorithm for our model.

4. Artificial Neural Network (ANN):

It is a deep learning algorithm that is particularly inspired by the working of a biological brain. It has a multilayer structure 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 neurons of the previous layer to a response signal that gets transmitted to the neurons of the subsequent layer. This consists of forward transmission as well as reverse transmission [17].

5. Logistic Regression:

Similar to Linear Regression, Logistic Regression is a supervised ML algorithm but it is used for classification rather than regression. When the outcome variable is dichotomous, this powerful analytical technique is used [5]. It shows the linear relations between the independent variables and classifies them into binary form.The equation used by the basic logistic model is Ln (p/1-p)=a0+a1*x+a2*x (1) This is called the logistic function [8].

IV. RESULTS AND OBSERVATIONS

We used a dataset of 7 different types of glasses which contains 9 features and 1 output. We used different machine learning techniques to predict the type of glass based on its element measurements and refractive indices. We applied 5 different ML algorithms Support Vector Machine (SVM) algorithm, K-nearest neighbors (KNN) algorithm, Random Forest algorithm, Artificial Neural Network (ANN), and Logistic Regression algorithm. Out of all Random Forest was the best suited for the dataset for prediction, with an accuracy of 79.62%. SVM and KNN were the 2nd and 3rd best-suited algorithms with an accuracy of 77.77% and 74.07% respectively.

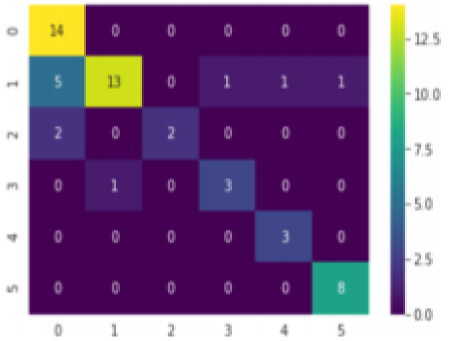

Figure 2 : Logistic Regression Confusion Matrix

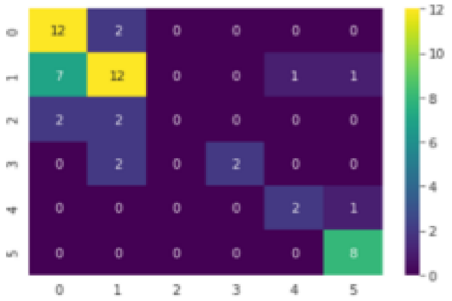

Figure 3 : Support Vector Machine (SVM) Confusion Matrix

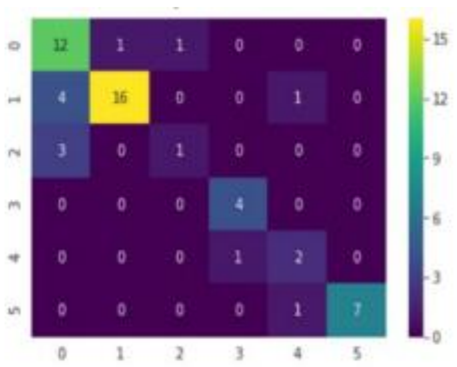

Figure 4: K-Nearest Neighbors (KNN) Confusion Matrix

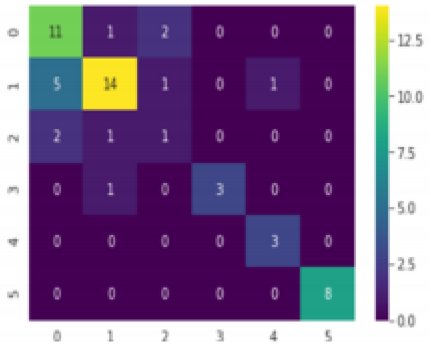

Figure 6: Random Forest Confusion Matrix

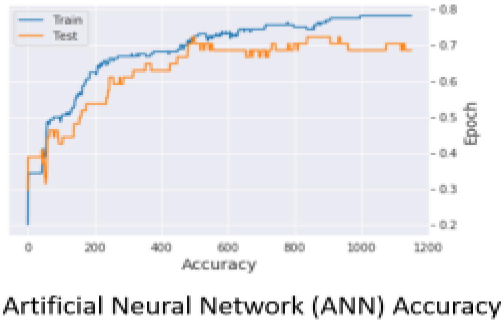

Figure 7: Artificial Neural Network (ANN) Accuracy

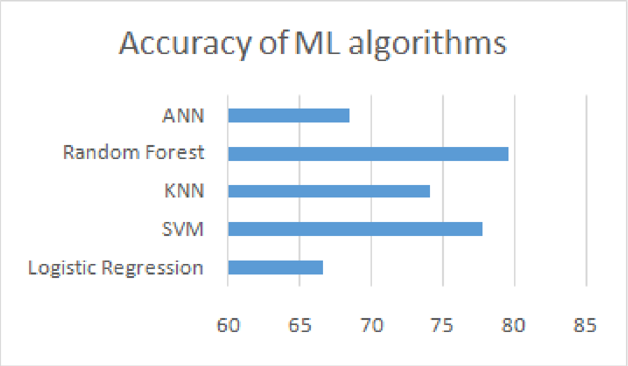

Figure 8: Accuracy of ML Algorithms

V. 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 accuracy of the model of algorithm.

The model based on Random Forest algorithm performed the best with 79.62% accuracy, followed by Support Vector Machine (SVM) with 77.77% accuracy.

ACKNOWLEDGMENT

We would like to thank Neha Bharti, a researcher, and professor at Jaipur Engineering College and Research Centre, for guiding us throughout the research.

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 word capitalized,” J. Name Stand. Abbrev., in press.
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. Devroye, L., Gyorfi, L., Krzyzak, A. & Lugosi, G. (1994) "On the strong universal consistency of nearest neighbor regression function estimates", Ann. Statist, 22: 1371– 1385.
11. Devroye, L. & Wagner, T.J. (1977) "The strong uniform consistency of nearest neighbor density estimates", Ann. Statist., 5: 536–540.
12. Devroye, L. & Wagner, T.J. (1982) "Nearest neighbor methods in discrimination, In Classification, Pattern Recognition and Reduction of Dimensionality", Handbook of Statistics, 2: 193–197. North-Holland, Amsterdam.
13. Jolliffe IT. 2002. Principal component analysis, 2nd edn New York, NY: Springer-Verlag.
14. Flury B. 1988. Common principal components and related models. New York, NY: Wiley.
15. Hallin M, Paindaveine D, Verdebout T. 2014. Efficient R-estimation of principal and common principal components. J. Am. Stat. Assoc. 109, 1071–1083. (10.1080/01621459.2014.880057)
16. Schapire, R. and Freund, Y. Boosting: Foundations and Algorithms. MIT Press, 2012.
17. Liu Zhongying, Zhao Qinghua, (Bidding for construction projects)2007,224-225
18. Mashael S. Aldayel. 2012. K-Nearest Neighbor classification for glass identification problem. ICCSII.2012.645452
19. D. P. Vivencio, E. Hruschka, M. Nicoletti, E. dos Santos, and S. Galvao, “Feature-weighted k-Nearest Neighbor Classifier,” Foundations of Computational Intelligence, 2007. FOCI 2007. IEEE Symposium, pp. 481–486, 2007.

AUTHORS PROFILE

Harshit Mathur currently pursuing bachelor’s in technology in computer science from Jaipur Engineering College and Research Centre, Jaipur. His research areas include Machine Learning, Data Science and Deep Learning.

Aditya Surana currently pursuing bachelor’s in technology in computer science from Jaipur Engineering College and Research Centre, Jaipur. His research areas include Machine Learning, Data Science and Deep Learning.