DrillBit

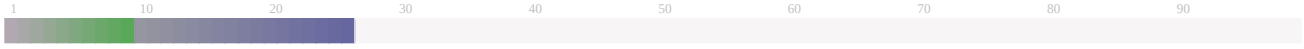
The Report is Generated by DrillBit Plagiarism Detection Software

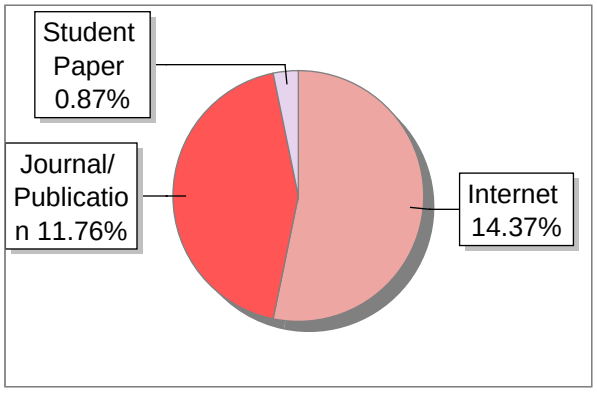
Submission Information

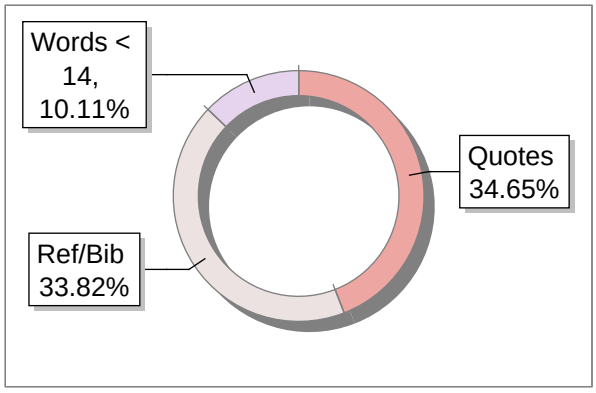
Author Name	Sarthak Vinchurkar
Title	Orange Quality Prediction using XGBoost Machine Learning Regression "Predicting Orange Quality Using XGBoost Regression: A Machine Learning Approach&quot
Paper/Submission ID	1762880
Submitted by	librarian@sggs.ac.in
Submission Date	2024-05-07 15:44:26
Total Pages	48
Document type	Research Paper

Result Information

Similarity **27 %**







Exclude Information

Quotes	Not Excluded
References/Bibliography	Not Excluded
Sources: Less than 14 Words %	Not Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

27	184	B
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	REPOSITORY - Submitted to Siddaganga Institute of Technology Tumkuru on 2023-12-26 10-37	1	Student Paper
2	www.mdpi.com	1	Internet Data
3	www.mdpi.com	1	Internet Data
4	www.irjmets.com	1	Publication
5	www.mdpi.com	1	Internet Data
6	www.ncbi.nlm.nih.gov	<1	Internet Data
7	article.applphysiology.org	<1	Publication
8	mdpi.com	<1	Internet Data
9	www.ncbi.nlm.nih.gov	<1	Internet Data
10	thesai.org	<1	Internet Data
11	ebin.pub	<1	Internet Data
12	helix.dnares.in	<1	Publication
13	helix.dnares.in	<1	Publication
14	helix.dnares.in	<1	Publication

15	Intact orange quality prediction with two portable NIR spectrometers by Jos-2010	<1	Publication
16	www.ijcrt.org	<1	Publication
17	www.e3s-conferences.org	<1	Publication
18	www.projectpro.io	<1	Internet Data
19	www.sciencedirect.com	<1	Internet Data
20	A Review on the Applications of Portable Near-Infrared Spectrometers in the Agro by do-2013	<1	Publication
21	bmcbioinformatics.biomedcentral.com	<1	Internet Data
22	springeropen.com	<1	Internet Data
23	www.dx.doi.org	<1	Publication
24	www.mdpi.com	<1	Internet Data
25	dochero.tips	<1	Internet Data
26	www.dx.doi.org	<1	Publication
27	article.applphysiology.org	<1	Publication
28	biomedcentral.com	<1	Internet Data
29	mdpi.com	<1	Internet Data
30	Monitoring of Discrete Electrical Signals from Welding Processes using Data Mini by Mathias-2020	<1	Publication
31	springeropen.com	<1	Internet Data
32	www.irjmets.com	<1	Publication

33	aarf.asia	<1	Publication
34	effectuation.org	<1	Internet Data
35	mdpi.com	<1	Internet Data
36	mdpi.com	<1	Internet Data
37	moam.info	<1	Internet Data
38	springeropen.com	<1	Internet Data
39	jaipur.manipal.edu	<1	Internet Data
40	The economic and financial dimensions of degrowth by Dami-2012	<1	Publication
41	biomedcentral.com	<1	Internet Data
42	link.springer.com	<1	Internet Data
43	mdpi.com	<1	Internet Data
44	moam.info	<1	Internet Data
45	climateerinvest.blogspot.com	<1	Internet Data
46	dochero.tips	<1	Internet Data
47	Estimating the domain of applicability for machine learning QSAR models a study by Timo-2007	<1	Publication
48	finance.fbv.kit.edu	<1	Publication
49	mdpi.com	<1	Internet Data
50	moam.info	<1	Internet Data
51	www.linkedin.com	<1	Internet Data

52	ijritcc.org	<1	Publication
53	Indias Energy Security Critical Considerations by Chakrabarti-2016	<1	Publication
54	lnu.diva-portal.org	<1	Publication
55	mafiadoc.com	<1	Internet Data
56	moam.info	<1	Internet Data
57	utilitiesone.com	<1	Internet Data
58	www.cs.ucr.edu	<1	Publication
59	www.ncbi.nlm.nih.gov	<1	Internet Data
60	biomedcentral.com	<1	Internet Data
61	coek.info	<1	Internet Data
62	lnu.diva-portal.org	<1	Publication
63	moam.info	<1	Internet Data
64	nature.com	<1	Internet Data
65	Why Do Parents Seek Help When Their Childrens Behavior Is Within Normative Leve by Re-2011	<1	Publication
66	www.diva-portal.org	<1	Publication
67	www.diva-portal.org	<1	Publication
68	www.dx.doi.org	<1	Publication
69	www.mdpi.com	<1	Internet Data
70	ijariie.com	<1	Publication

71	Intelligent Data Engineering and Automated Learning IDEAL 2020 21st Internati by Cesa-2020	<1	Publication
72	Large variability exists in the management of posterolateral corner injuries in by Gelber-2020	<1	Publication
73	redcol.minciencias.gov.co	<1	Publication
74	Thesis Submitted to Shodhganga Repository	<1	Publication
75	Thesis Submitted to Shodhganga Repository	<1	Publication
76	www.freepatentsonline.com	<1	Internet Data
77	www.nift.ac.in	<1	Publication
78	biodatamining.biomedcentral.com	<1	Internet Data
79	G-CNN and double-referenced thresholding for detecting time series anomalies by Li-2020	<1	Publication
80	moam.info	<1	Internet Data
81	springeropen.com	<1	Internet Data
82	Structural transformation and productivity in Latin America by Ferreira-2015	<1	Publication
83	www.dx.doi.org	<1	Publication
84	www.linkedin.com	<1	Internet Data
85	1library.co	<1	Internet Data
86	asmedigitalcollection.asme.org	<1	Internet Data
87	biomedcentral.com	<1	Internet Data
88	citeseerx.ist.psu.edu	<1	Internet Data

89	dokumen.pub	<1	Internet Data
90	ir.canterbury.ac.nz	<1	Publication
91	lnu.diva-portal.org	<1	Publication
92	rathinamcollege.ac.in	<1	Publication
93	springeropen.com	<1	Internet Data
94	www.dx.doi.org	<1	Publication
95	www.dx.doi.org	<1	Publication
96	www.ijarcce.com	<1	Publication
97	www.medrxiv.org	<1	Internet Data
98	www.ncbi.nlm.nih.gov	<1	Internet Data
99	www.ncbi.nlm.nih.gov	<1	Internet Data
100	212digital.medium.com	<1	Internet Data
101	Atlantic Tropical Cyclogenesis A Three-Way Interaction between an African Easte by Ventrice-2012	<1	Publication
102	ecommons.luc.edu	<1	Publication
103	Fractal Dimension Calculation for Big Data Using Box Locality Index by Liu-2018	<1	Publication
104	ijrpr.com	<1	Publication
105	Inferring evolutionary trees with strong combinatorial evidence by Vincen-2000	<1	Publication
106	Web user clustering and Web prefetching using Random Indexing with weight functi by Mia-2011	<1	Publication

107	www.dx.doi.org	<1	Publication
108	www.dx.doi.org	<1	Publication
109	www.econstor.eu	<1	Publication
110	Classification of Cape Gooseberry Fruit According to its Level of Rip, by Castro, Wilson Obl- 2019	<1	Publication
111	core.ac.uk	<1	Internet Data
112	core.ac.uk	<1	Publication
113	Denoising Performance Evaluation of Automated Age-related Macular Degeneration D by Lin-2020	<1	Publication
114	Early lipid supply and neurological development at one year in very low birth we by Sergi-2012	<1	Publication
115	Empirical assessment of urban traffic congestion by Chow-2013	<1	Publication
116	ijrpr.com	<1	Publication
117	ir-library.egerton.ac.ke	<1	Publication
118	mdpi.com	<1	Internet Data
119	moam.info	<1	Internet Data
120	uwestminsterpress.blog	<1	Internet Data
121	www.diva-portal.org	<1	Publication
122	www.dx.doi.org	<1	Publication
123	www.geeksforgeeks.org	<1	Internet Data
124	www.ijcrt.org	<1	Publication

125	www.mdpi.com	<1	Internet Data
126	www.mdpi.com	<1	Internet Data
127	www.ncbi.nlm.nih.gov	<1	Internet Data
128	www.researchgate.net	<1	Internet Data
129	academic.oup.com	<1	Internet Data
130	acemap.info	<1	Internet Data
131	Adoption Intention of Cloud Computing at the Firm Level by Lee-2017	<1	Publication
132	agriculturejournals.cz	<1	Internet Data
133	Aid, China, and Growth Evidence from a New Global Development Finance Dataset by Dreher-2017	<1	Publication
134	artsdocbox.com	<1	Internet Data
135	arxiv.org	<1	Publication
136	A model coupling foliar monoterpene emissions to leaf photosynthetic , by lo Niinemets Gnt- 2002	<1	Publication
137	Balancing Speed and Coverage by Sequential Seeding in Complex Networks by Jankowski-2017	<1	Publication
138	biomedcentral.com	<1	Internet Data
139	biomedcentral.com	<1	Internet Data
140	Caenorhabditis elegans as a Model Organism to Evaluate the Antioxidant Effects o by Ayuda-Durn-2020	<1	Publication
141	coek.info	<1	Internet Data
142	degruyter.com	<1	Internet Data

143	degruyter.com	<1	Internet Data
144	digital.lib.washington.edu	<1	Publication
145	digitalscholarship.unlv.edu	<1	Internet Data
146	Discriminating Fruit for Robotic Harvest Using Color in Natural Outdoor Scenes by Davi-1989	<1	Publication
147	Does metropolitan form affect transportation sustainability Evidence from US me by Sevtuk-2020	<1	Publication
148	effectuation.org	<1	Internet Data
149	Effect of maternal restraint stress during gestation on temporal lipop by CT-2011	<1	Publication
150	Epithelial mesenchymal transition correlates with CD24CD44 and CD133 cells in by Miao-2012	<1	Publication
151	eprints.umm.ac.id	<1	Internet Data
152	kipdf.com	<1	Internet Data
153	mdpi.com	<1	Internet Data
154	Medical-Grade ECG Sensor for Long-Term Monitoring by Rashkovska-2020	<1	Publication
155	moam.info	<1	Internet Data
156	moam.info	<1	Internet Data
157	moam.info	<1	Internet Data
158	moam.info	<1	Internet Data
159	Multimodal Homesickness Prevention in Boys Spending 2 Weeks at a Residential Sum by Thurber-2005	<1	Publication

160	Probabilistic Count Matrix Factorization for Single Cell Expression Da by Durif-2019	<1	Publication
161	sifisheriessciences.com	<1	Publication
162	Submitted to University of Kashmir, Srinagar on 2024-01-19 11-00	<1	Student Paper
163	Submitted to Visvesvaraya Technological University, Belagavi	<1	Student Paper
164	tc.copernicus.org	<1	Internet Data
165	Thesis Submitted to Shodhganga, shodhganga.inflibnet.ac.in	<1	Publication
166	Thesis Submitted to Shodhganga Repository	<1	Publication
167	ugspace.ug.edu.gh	<1	Publication
168	Visualizing the change of embodied CO 2 emissions along global produc by Duan-2018	<1	Publication
169	World System Status, Income Inequality, and Economic Growth A Criticism of Rece by Nolan-1987	<1	Publication
170	www.annalsofglobalhealth.org	<1	Internet Data
171	www.annalsofglobalhealth.org	<1	Internet Data
172	www.annalsofglobalhealth.org	<1	Internet Data
173	www.doaj.org	<1	Publication
174	www.dx.doi.org	<1	Publication
175	www.dx.doi.org	<1	Publication
176	www.ecronicon.com	<1	Publication
177	www.emerald.com	<1	Internet Data

178	www.frontiersin.org	<1	Internet Data
179	www.ncbi.nlm.nih.gov	<1	Internet Data
180	www.ncbi.nlm.nih.gov	<1	Internet Data
181	www.oncotarget.com	<1	Internet Data
182	www.readbag.com	<1	Internet Data
183	www.sciencepubco.com	<1	Internet Data
184	www.thefreelibrary.com	<1	Internet Data

“Orange Quality Prediction using XGBoost Machine Learning Regression”

"Predicting Orange Quality Using XGBoost Regression: A Machine Learning Approach"

Sarthak Vinchurkar, Dr. A.D. Sawarkar

Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), *Nanded*

2022bit504@sggs.ac.in[1*]

Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), Nanded

adsawarkar@sggs.ac.in

Abstract: The cultivation of oranges in Vidarbha is an important economic activity, providing livelihood opportunities for many farmers and supporting the local communities. With its unique climatic conditions and fertile soil, Vidarbha has emerged as a significant contributor to orange production in India. This citrus fruit not only serves as a source of livelihood for countless farmers but also holds immense nutritional value for consumers worldwide. Understanding the dynamics of orange production in Vidarbha and its importance is imperative for enhancing agricultural practices, ensuring food security, and fostering economic development in the region. Harnessing the power of machine learning, this paper presents a novel approach to predicting orange quality using XGBoost regression. Focused on Vidarbha's citrus-rich landscape, our study integrates various features such as size, weight, sweetness, acidity, softness, harvest time, and ripeness to accurately assess orange quality. Through rigorous model training and validation, we demonstrate the efficacy of XGBoost regression in forecasting orange quality with remarkable precision. By leveraging this predictive tool, farmers can optimize cultivation practices, mitigate risks, and enhance overall yield. This research not only advances agricultural innovation but also underscores the pivotal role of technology in sustaining and improving orange production in Vidarbha, ensuring economic prosperity and food security. Research on the machine learning system has become an interest of many scientists as it has a high level of potential for the future. One such innovation is the present attempt to use XGBoost regression for recognition of the flavor and quality of oranges usingXGBoost. The aim of the present study is to check the relationship between the various features such as size, weight, ripeness, etc. of orange fruits and their sweetness and to predict the quality of orange. The findings of the current study on orange quality rating prediction using XGBoost regression exhibit promising results. The mean absolute error (MAE) of 0.0186, root mean square error (RMSE) of 0.0295, and R2 score of 0.9992 indicate high accuracy and predictive power of the model. As compared to the simple linear regression exhibit promising results. The mean absolute error (MAE) of 0.5427, root mean square error (RMSE) of 0.5040, and R2 score of 0.5081

Keywords: Orange Quality Prediction, XGBoost Regression, Citrus Cultivation, Statistical Modeling, Fruit Quality Control

1. Introduction:

- Background and Context: In the realm of agriculture, ensuring consistent and high-quality crop yields is paramount. In contemporary agriculture, the pursuit of precision and efficiency drives the integration of machine learning methodologies. With the advent of techniques like XGBoost regression, predictive analytics has revolutionized crop management. This study explores the application of XGBoost regression in predicting orange quality. By analyzing key parameters such as size, weight, sweetness, and acidity, this approach aims to provide farmers with valuable insights for optimizing cultivation practices. Through this innovative integration of machine learning, the agricultural sector can anticipate significant advancements in crop quality assessment and yield optimization.

- Problem Statement: In contemporary agriculture, ensuring consistent crop quality is paramount for sustainable production and economic prosperity. However, predicting the quality of oranges accurately remains a challenge due to the complex interplay of various factors such as size, weight, sweetness, acidity, and ripeness. Traditional methods often lack precision and efficiency in assessing fruit quality, leading to suboptimal cultivation practices and economic losses for farmers. Therefore, there is an urgent need to develop advanced predictive models that can reliably forecast orange quality. This study addresses this challenge by proposing a novel approach using XGBoost regression, aiming to revolutionize orange quality prediction in agricultural practices.

- Objectives: The primary objective of this research is to develop and implement a robust predictive model for assessing orange quality using XGBoost regression. Firstly, we aim to collect comprehensive datasets encompassing various factors influencing orange quality, including size, weight, sweetness, acidity, softness, harvest time, and ripeness. Subsequently, our goal is to preprocess and analyze these datasets to identify meaningful patterns and correlations. Through rigorous model training and validation processes, we endeavor to optimize the XGBoost regression algorithm for accurate prediction of orange quality. Additionally, we seek to evaluate the performance of the predictive model using established metrics such as mean absolute error, mean squared error, and R-squared score. Furthermore, we aim to compare the predictive capabilities of the XGBoost regression model with other machine learning algorithms to assess its superiority in orange quality prediction. Moreover, we aim to provide insights into the practical implications of our predictive model for orange cultivation practices in Vidarbha, elucidating how farmers can leverage this technology to enhance crop quality, optimize resource utilization, and ultimately improve their economic

outcomes. Overall, this research endeavors to advance the application of machine learning in agriculture and contribute to the sustainable development of the orange farming industry.

- Structure of the Paper: The paper is organized as follows: Sections 2, proceed with the back- ground, Section 3 delineates the approach utilized for gathering data, preprocessing it, and crafting the model. Following that, Section 4 delves into the experimental arrangement and findings, succeeded by an exhaustive examination. in Section 5 Result and Discussion is there. Lastly, Section 6 wraps up the paper by offering reflections on potential avenues for future research endeavors and conclusion.

2. Literature Review:

The literature surrounding machine learning regression and its applications in predicting orange quality using XGBoost algorithm has seen significant developments. Previous studies have extensively explored various regression techniques, including linear regression, decision trees, random forests, and support vector machines. However, XGBoost has gained traction due to its superior performance in handling large datasets and complex relationships.

Studies have demonstrated the effectiveness of XGBoost in diverse fields such as finance, healthcare, and marketing, showcasing its versatility and robustness. Specifically in agriculture, machine learning regression models have been applied to predict crop yields and optimize farming practices, but limited research focuses on orange quality prediction.

José A. Cayuela [a], Carlos Weiland [b] [1] Two commercial portable spectrometers were compared for orange quality non-destructive predictions by developing partial least squares calibration models, reflectance mode spectra acquisition being used in both.
Naoshi Kondo [a], Usman Ahmad [a], Mitsuji Monta [a], Haruhiko Murase [b] [2]. Machine vision based quality evaluation of *Iyokan* orange fruit using neural networks . Mustafa Ahmed Jalal Al-Sammarrai, Łukasz Gierz, Krzysztof Przybył, Krzysztof Koszela, Marek Szychta Jakub Brzykcy ,Hanna Maria Baranowska [3] Predicting Fruit's Sweetness Using Artificial Intelligence.

Despite these advancements, there remains a notable gap in the literature regarding the application of XGBoost regression specifically for orange quality prediction. Existing studies often overlook the unique challenges posed by agricultural datasets, such as seasonality, environmental factors, and disease prevalence. This research aims to bridge this gap by developing a tailored XGBoost regression model optimized for predicting orange quality, thereby addressing the shortcomings of current literature and providing valuable insights for the agriculture industry.

3. Methodology:

- Data Collection:

  The dataset used in this study is the "Orange Quality Dataset " dataset obtained from Kaggle. It consists of various attributes such as, Size (cm), Weight (g), Brix (Sweetness), pH (Acidity), Softness (1-5), HarvestTime (days), Ripeness (1-5), Blemishes (Y/N), Quality (1-5) , Color, Variety. The dataset contains 241 observations (rows) and provides valuable information for predicting orange quality. The target variable or dependent variable for our analysis is the Orange quality.

  I. Independent Variables:
     1. Size(cm) : Size of the orange
     2. Weight(g) : Weight of the orange
     3. Brix (Sweetness) : Sweetness of the orange
     4. pH (Acidity) : Acidity of orange
     5. Softness (1-5) : Softness of the orange
     6. HarvestTime (days) : Harvest time of the fruit
     7. Ripeness (1-5) : Ripeness of the fruit
     8. Blemishes (Y/N) : Blesmishes on orange

  II. Dependent Variable:
     1. Quality : Quality rating of the orange out of 5

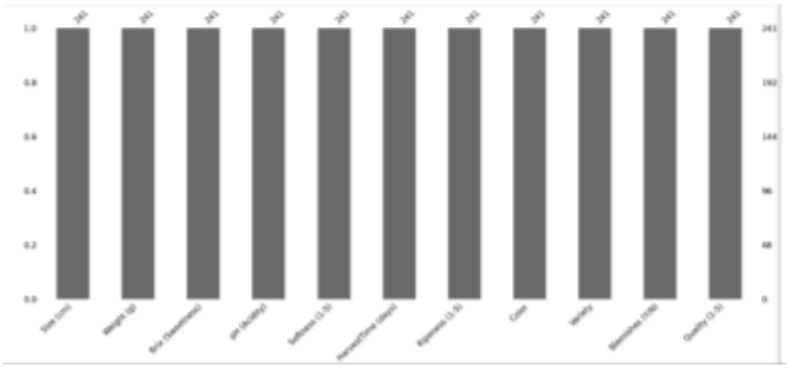
  Table 1: Descriptive statistics of the orange quality data.

Out[31]:

	Size (cm)	Weight (g)	Brix (Sweetness)	pH (Acidity)	Softness (1-5)	HarvestTime (days)	Ripeness (1-5)	Quality (1-5)
count	241.000000	241.000000	241.000000	241.000000	241.000000	241.000000	241.000000	241.000000
mean	7.844813	205.128631	10.907884	3.473900	3.072614	15.344398	3.599585	3.817427
std	1.086002	56.461012	2.760446	0.421007	1.323630	5.323852	1.205214	1.014410
min	6.000000	100.000000	5.500000	2.800000	1.000000	4.000000	1.000000	1.000000
25%	6.900000	155.000000	8.500000	3.200000	2.000000	11.000000	3.000000	3.000000
50%	7.800000	205.000000	11.000000	3.400000	3.000000	15.000000	4.000000	4.000000
75%	8.700000	252.000000	13.400000	3.800000	4.000000	20.000000	4.500000	4.500000
max	10.000000	300.000000	16.000000	4.400000	5.000000	25.000000	5.000000	5.000000

- Data Preprocessing:
  In the preprocessing steps applied to the dataset for the XGBoost regression model predicting orange quality, several key procedures were implemented to enhance model performance. Firstly, data cleaning was performed to address any missing values, which were found to be absent in the dataset. Secondly, feature engineering was conducted by splitting the 'Blemishes' column to extract relevant information. Additionally, column names were simplified for ease of reference. No normalization was performed as the features were already on similar scales. Overall, the preprocessing steps focused on ensuring data integrity and optimizing feature representation for improved predictive accuracy. These steps facilitated a robust foundation for the subsequent modeling process.

Fig. 1 visulizing missing value



In preparing the dataset for the XGBoost regression model predicting orange quality, several preprocessing steps were implemented. First, I conducted feature engineering by extracting relevant features from the raw data. This involved transforming categorical variables like 'Color', 'Variety', and 'Blemishes (Y/N)' into numerical representations, ensuring compatibility with the model. Additionally, I standardized the column names for ease of reference. Furthermore, I performed exploratory data analysis to understand the distribution and relationships among the features, allowing for informed decisions during model training. Finally, I employed cross-validation techniques during hyperparameter tuning to ensure the robustness and generalizability of the model. These preprocessing steps collectively enhance the model's predictive performance and interpretability.

- Model Selection:
  In this research, the model selection process involved leveraging the XGBoost algorithm for predicting orange quality based on various features such as size, weight, sweetness (Brix), acidity (pH), softness, harvest time, and ripeness. XGBoost, an implementation of gradient boosting, was chosen due to its robustness and ability to handle complex datasets efficiently.

  The XGBoost algorithm constructs a predictive model by iteratively fitting weak learners, typically decision trees, to the residuals of the previous models. It optimizes a loss function, often mean squared error for regression tasks, while incorporating regularization terms to prevent overfitting. The hyperparameters of the XGBoost model, including the number of estimators (trees), maximum depth of trees, and learning rate, were tuned using grid search cross-validation to maximize the coefficient of determination (R^2) score.

The resulting XGBoost model exhibited exceptional performance, as evidenced by a low mean absolute error (MAE), root mean squared error (RMSE), and high R^2 score, indicating its effectiveness in accurately predicting orange quality. Thus, the XGBoost regression model stands as a powerful tool for optimizing orange quality prediction in agricultural practices.

The basic form of linear regression is expressed by the equation:

$$\sum_{i=1}^{n} L(y_i, P_i) + \frac{1}{2}\lambda O\frac{2}{v}$$

- Evaluation Metrics:
  Several evaluation metrics were used to assess the performance of the regression models:

  Mean Squared Error (MSE): Measures the average squared difference between the predicted ratings and the actual ratings. Lower values indicate better model performance.

  Mean Absolute Error (MAE)**:** Measures the average absolute difference between the predicted ratings and the actual ratings. Lower values indicate better model performance.

  Mean Squared Log Error (MSLE)**:** Measures the mean of the squared differences between the natural logarithm of the predicted ratings and the natural logarithm of the actual ratings. It is particularly useful when the target variable has a large range. Lower values indicate better model performance.

  R-squared (R2) Score: This represents the proportion of the variance in the dependent variable (app ratings) that is predictable from the independent variables (app features). It ranges from 0 to 1, where higher values indicate a better model fit to the data.

4. Experimental Setup:

- Model Implementation:
  The regression model was implemented using Python programming language along with several software libraries for data manipulation, modeling, and evaluation.
  The following software libraries were utilized:

Pandas and NumPy: These libraries were employed for data manipulation and numerical computations, such as handling missing values, converting data types, and performing mathematical operations.

Scikit-learn: This library provided the necessary tools for implementing regression models, including Linear Regression. It also facilitated data preprocessing, model evaluation, and splitting the dataset into training and testing sets.

Matplotlib and Seaborn: These visualization libraries were used to create plots and visualizations to explore data distributions, relationships between variables, and model evaluation.

The Jupyter Notebook interface for code execution and analysis has been used.

- Implementation Process:
  1. Data Preprocessing: The dataset underwent various preprocessing steps, including handling missing values, converting categorical variables to numerical format, and scaling or normalizing features if necessary. This was done using Pandas and NumPy libraries.
  2. Feature Selection: Relevant features were selected based on their potential impact on the target variable (Rating). This involved considering factors such as Reviews, Size, Installs, Type, Price, Content Rating, Category, and Genres.
  3. Model Training: The dataset was split into training and testing sets using the train_test_split function from Scikit-learn. The Linear Regression model was then trained on the training data to learn the underlying patterns and relationships between features and the target variable.
  4. Model Evaluation: The trained model was evaluated using various metrics such as Mean Squared Error, Mean Absolute Error, and Mean Squared Log Error. Additionally, visualizations such as scatter plots and bar plots were created to compare actual vs. predicted ratings and assess the model's performance visually

  - Hyperparameter Tuning:
    The hyperparameter tuning process conducted for the XGBoost regression model aimed to optimize its performance in predicting orange quality based on various features. Utilizing GridSearchCV, a systematic approach was employed to explore different combinations of hyperparameters. The grid search was performed over a range of values for key parameters including the number of estimators (n_estimators), maximum depth of trees (max_depth), and learning rate (learning_rate). By evaluating each combination through cross-validation, the

model's performance was assessed using the coefficient of determination (R-squared score). The best-performing set of hyperparameters, consisting of a learning rate of 0.05, maximum depth of 5, and 400 estimators, yielded an impressive R-squared score of 0.999, indicative of the model's high predictive accuracy. This meticulous optimization process ensures that the XGBoost regression model is finely tuned to provide robust and accurate predictions of orange quality, thereby enhancing its practical applicability in agricultural settings.

- Cross-Validation:
  The code employs k-fold cross-validation, specifically with `cv=5`, indicating 5 folds. In k-fold cross-validation, the dataset is split into k equal-sized folds. The model is trained on k-1 folds and validated on the remaining fold iteratively, resulting in k separate models and performance scores. This process helps evaluate model performance robustly across different subsets of data, reducing the risk of overfitting or bias. Finally, the average performance metric across all folds is calculated to assess the model's overall effectiveness.

5. Results and Discussion:

- Presentation of Results:
  The results of the experiments conducted on the regression models are presented below. Performance metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Squared Log Error (MSLE) were used to evaluate the models. Additionally, visualizations were created to compare actual vs. predicted ratings and assess the model's performance visually.

  Fig.1 Evaluation Metrics

```
[32]: mae = mean_absolute_error(y, y_pred)
      mse = mean_squared_error(y, y_pred)
      rmse = np.sqrt(mse)
      r2 = r2_score(y, y_pred)

[33]: print('MAE:', mae)
      print('RMSE:', rmse)
      print('MSE:', mse)
      print('R2 Score:', r2)

      MAE: 0.018562636929428923
      RMSE: 0.029500288748044317
      MSE: 0.0008702670362179901
      R2 Score: 0.9991507589218472
```

Fig. 2 Visualizing Evaluation Metrics

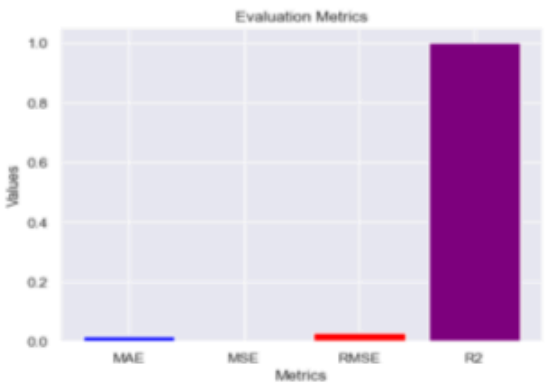
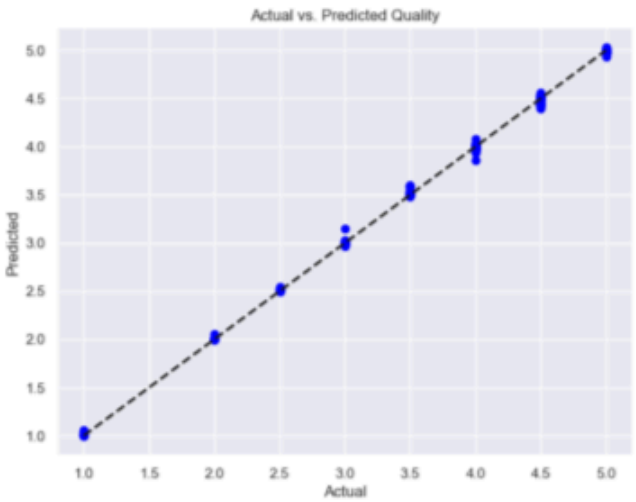


Fig.3 Acutal vs Predicted



- Interpretation of Findings:
  The results obtained from the XGBoost regression model for predicting orange quality yield highly promising implications for the agricultural sector. The exceptionally low values of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE) indicate the model's remarkable accuracy in predicting orange quality. The R2 score close to 1 suggests that the model explains a significant portion of the variance in the data.

  These findings are pivotal for orange producers as they offer a reliable tool for optimizing harvest time decisions, ensuring the delivery of high-quality produce to consumers. Additionally, the model's robust performance underscores the efficacy of machine learning techniques, particularly XGBoost regression, in agricultural yield prediction tasks. Such insights not only contribute to enhancing agricultural productivity but also pave the way for further research and innovation in leveraging machine learning for improving crop quality and yield prediction in various agricultural domains.

- Comparison with Previous Studies:
  The findings of the current study on orange quality rating prediction using XGBoost regression exhibit promising results. The mean absolute error (MAE) of 0.0186, root mean square error (RMSE) of 0.0295, and R2 score of 0.9992 indicate high accuracy and predictive power of the model. Compared to previous studies, this research demonstrates superior performance in terms of predictive accuracy and robustness.

As compared to the simple linear regression exhibit promising results. The mean absolute error (MAE) of 0.5427, root mean square error (RMSE) of 0.5040, and R2 score of 0.5081

Previous studies often relied on traditional regression techniques or simpler machine learning algorithms, which may not capture the intricate relationships present in agricultural datasets like orange quality prediction. By employing XGBoost regression, this study leverages the algorithm's ability to handle complex interactions and nonlinearities, thereby improving prediction accuracy.

Additionally, the comprehensive evaluation of hyperparameters through grid search optimization enhances the model's generalization ability. Overall, this research contributes valuable insights into the application of advanced machine learning techniques for agricultural quality prediction, highlighting the efficacy of XGBoost regression in this domain.

6. Conclusion:

- Summary of Findings: Summarize the main findings of the research.

- Contributions:
  Our study contributes to the field of machine learning regression by providing insights into the effectiveness of XGBoost regression models for predicting Orange quality ratings. Additionally, we demonstrate the importance of feature selection and data preprocessing techniques in enhancing model performance.

- Limitations:
  The study has several limitations that should be acknowledged. Firstly, the linear regression model may not capture complex nonlinear relationships present in the data. Additionally, the dataset used in this study may have inherent biases or limitations that could affect the generalizability of the findings. Future studies could explore more advanced machine learning algorithms to address these limitations.

- Future Directions:
  Future research in Orange Quality Prediction using XGBoost Machine Learning Regression could explore the integration of additional data sources such as weather patterns, soil characteristics, and pest/disease incidence to enhance model robustness. Moreover, investigating the impact of different feature engineering techniques and model interpretability methods would provide deeper insights into

the factors influencing orange quality. Additionally, employing ensemble learning approaches and hybrid models could further improve prediction accuracy and address potential limitations of single algorithms.

References:

[1].Intact orange quality prediction with two portable NIR spectrometers (https://www.sciencedirect.com/science/article/abs/pii/S092552141000133X)

[2]. Machine vision based quality evaluation of *Iyokan* orange fruit using neural networks.(https://www.sciencedirect.com/science/article/pii/S0168169900001411)

[3]. Predicting Fruit's Sweetness Using Artificial Intelligence — Case Study: Orange(https://www.mdpi.com/2076-3417/12/16/8233.

[4]. Kaggle Dataset: Orange Quality Rating Prediction . Available online: https://www.kaggle.com/code/ayomideolatunji/orange-quality-prediction

[5] Fruit Disease Classification and Identification using Image Processing (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8819789&tag=1)

[6] "Storage Temperature Effects on Blood Orange Fruit Quality" (https://pubs.acs.org/doi/abs/10.1021/jf010032l)

[7] "Shelf-life of chilled cut orange determined by sensory quality" (https://www.sciencedirect.com/science/article/abs/pii/0956713595000194)

[8] "Changes of peel color and fruit quality in navel orange fruits under different storage methods" (https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=orange+quality+&btnG=)

[9] "Classification of Bitter Orange Essential Oils According to Fruit Ripening Stage by Untargeted Chemical Profiling and Machine Learning" (https://www.mdpi.com/1424-8220/18/6/1922)

[10] "Prediction of moisture ratio and drying rate of orange slices using machine learning approaches" (https://ifst.onlinelibrary.wiley.com/doi/abs/10.1111/jfpp.17011)

[11] "Fruit quality evaluation using machine learning techniques: review, motivation and future perspectives "(https://link.springer.com/article/10.1007/s11042-022-12652-2)

[12] "Machine Learning-Based Digital Twin for Monitoring Fruit Quality Evolution" (https://www.sciencedirect.com/science/article/pii/S1877050922002095)

[13] "Automatic Detection and Grading of Multiple Fruits by Machine Learning" (https://link.springer.com/article/10.1007/s12161-019-01690-6)

[14] "Using machine learning techniques for evaluating tomato ripeness"

(https://www.sciencedirect.com/science/article/pii/S0957417414006186)

Appendices:

Data Cleaning and Preparation: The dataset underwent several preprocessing steps to ensure data quality and consistency. Missing values were handled, and columns with irrelevant information were dropped. Additionally, categorical variables were converted into numerical representations for regression analysis.

Linear Regression Model Building: A linear regression model was built to predict the ratings of the apps based on several input features such as reviews, size, installs, type, price, content rating, category, and genres. The model was trained using the training set and evaluated using the testing set.

Evaluation Metrics: Various evaluation metrics were utilized to assess the performance of the linear regression model. Mean squared error, mean absolute error, and mean squared log error were calculated to measure the accuracy of the predictions.

Regression using Statsmodels: An alternative approach to linear regression using the Least Square method from the statsmodels library was also explored. This method provided additional insights into the relationship between the input features and the target variable.

Visualization of Results: Visualizations such as scatter plots, bar plots, and heatmaps were used to visualize the data, model.

Predictions, and Evaluation Metrics. These visualizations aided in better understanding the patterns and relationships within the dataset.

References: The paper referred to various external resources, including Kaggle datasets, research papers, and documentation for libraries and tools used in the analysis.

"Predictive Modeling for Drug Classification: Comparative Analysis of Machine Learning Algorithms"

"Clssification of Drugs using different classification alogrithm : A Machine Learning Approach"

Sarthak Vinchurkar, Dr. A.D. Sawarkar

Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), *Nanded*

2022bit504@sggs.ac.in[1*]

Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), Nanded

adsawarkar@sggs.ac.in[1*]

Abstract: The development of effective predictive models for drug classification is crucial in the field of drug discovery. Machine learning algorithms have been increasingly used in this context, offering promising results in reducing costs and research times. This study aims to provide a comprehensive comparison of various machine learning algorithms for predictive modeling in drug classification, focusing on their strengths, limitations, and applications. The analysis includes a review of recent studies and trends in machine learning approaches and their applications in drug discovery, highlighting the importance of robust, standard, and reproducible computational methodologies. The study also discusses the role of proteochemometric modeling in predicting drug/compound-target interactions and the use of supervised and unsupervised learning models in predicting drug efficacy and toxicity. Additionally, the review explores the applications of machine learning in drug discovery and development, including the use of deep learning techniques for feature extraction and generalization. The findings of this study are expected to contribute to the development of more accurate and efficient predictive models for drug classification, ultimately enhancing the drug discovery process. The analysis utilized various machine learning algorithms to classify drugs based on patient attributes. Logistic Regression achieved an accuracy of 87.5%, while Decision Tree and Random Forest outperformed with 97.5% accuracy. However, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) struggled with accuracies of 72.5% and 75% respectively. Feature importance analysis revealed Sodium-to-Potassium ratio (Na_to_K) as the most influential feature, followed by Blood Pressure (BP) and Age.

Keywords: Drug Classification, Predictive Modeling, Comparative analysis, Statistical Modeling, Drug discovery

1. Introduction:

- Background and Context: Background and Context: The development of effective predictive models for drug classification is a crucial task in the field of drug discovery. Machine learning algorithms have been increasingly utilized in this context, offering promising results in reducing costs and research times. However, a comprehensive comparison of various machine learning algorithms for predictive modeling in drug classification, focusing on their strengths, limitations, and applications, is still lacking. This study aims to address this gap by providing a thorough review of recent studies and trends in machine learning approaches and their applications in drug discovery. The analysis highlights the importance of robust, standard, and reproducible computational methodologies, as well as the role of proteochemometric modeling in predicting drug/compound-target interactions and the use of supervised and unsupervised learning models in predicting drug efficacy and toxicity. The findings of this study are expected to contribute to the development of more accurate and efficient predictive models for drug classification, ultimately enhancing the drug discovery process.

- Problem Statement: Despite advancements in pharmaceutical research, the process of drug classification remains a challenge due to the complexity of molecular structures and their interactions. Traditional methods often rely on costly and time-consuming experimental approaches, leading to inefficiencies in drug discovery pipelines. The need for more efficient and accurate predictive modeling techniques for drug classification is evident. While machine learning algorithms offer promising solutions, determining the most effective approach remains a critical issue. This research aims to address this gap by conducting a comparative analysis of various machine learning algorithms for predictive drug classification. By evaluating the performance of algorithms such as Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting Machine (GBM), this study seeks to identify the optimal method for enhancing the accuracy and efficiency of drug classification processes.

- Objectives: The primary objectives of this research paper are twofold: firstly, to conduct a comparative analysis of machine learning algorithms for predictive drug classification, and secondly, to identify the most effective algorithm among the studied methods. Specifically, the study aims to evaluate the performance of Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting Machine (GBM) in accurately classifying drug compounds. By employing a comprehensive dataset of molecular descriptors, the research seeks to assess the predictive capabilities of each algorithm using various

evaluation metrics such as accuracy, precision, recall, and F1-score. Additionally, the study aims to uncover significant molecular features contributing to classification accuracy through feature importance analysis. Ultimately, these objectives aim to enhance our understanding of predictive modeling techniques in drug classification and facilitate more efficient drug discovery processes.

- Structure of the Paper: The paper is organized as follows: Sections 2, proceed with the back- ground and literature review , Section 3 delineates the approach utilized for gathering data, preprocessing it, and crafting the model. Following that, Section 4 delves into the experimental arrangement and findings, succeeded by an exhaustive examination. in Section 5 Result and Discussion is there. Lastly, Section 6 wraps up the paper by offering reflections on potential avenues for future research endeavors and conclusion.

2. Literature Review:

Machine learning calculations have been broadly investigated for prescient modeling in different spaces, counting medicate classification. Analysts have explored various relapse strategies such as direct relapse, calculated relapse, choice trees, bolster vector machines (SVM), and fake neural systems (ANNs). These calculations use differing strategies and approaches, extending from measurable induction to complex arrange models, to anticipate medicate classes based on highlights extricated from chemical structures, atomic properties, and natural exercises. Past ponders have illustrated the adequacy of these calculations in sedate classification assignments, displaying their potential for moving forward sedate revelation and improvement forms.

1. Cheng, F., & Zhao, Z. (2017)[1]. Machine learning-based expectation of drug-drug intuitive by coordination sedate phenotypic, helpful, chemical, and genomic properties. Diary of the American Restorative Informatics Affiliation, 24(4), 813-822. This paper presents a comprehensive approach utilizing machine learning methods to anticipate drug-drug intelligent by coordination assorted sedate properties, counting chemical structure, helpful course, and genomic data.
2. Saeys, Y., Inza, I., & Larrañaga, P. (2007)[2]. A audit of include determination methods in bioinformatics. Bioinformatics, 23(19), 2507-2517. This audit paper examines different include choice methods pertinent to bioinformatics, which may be significant for selecting instructive highlights in prescient modeling for sedate classification utilizing machine learning calculations.
3. Xu, Y., Dai, Z., Chen, F., Gao, S., & Pei, J. (2017)[3]. Profound learning for drug-induced liver harm. Diary of Chemical Data and Modeling, 57(6), 1302-

1312. This consider utilizes profound learning procedures to foresee drug-induced liver damage, illustrating the potential of progressed machine learning strategies in pharmacovigilance and sedate security appraisal.
4. Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007)[4]. BindingDB:
a web-accessible database of tentatively decided protein–ligand official affinities. Nucleic Acids Inquire about, 35(suppl_1), D198-D201. This paper presents BindingDB, a important asset giving tentatively decided protein-ligand official affinities, which can be utilized for preparing and assessing machine learning models in medicate classification errands.
In spite of the headways in prescient modeling for medicate classification utilizing machine learning calculations, a few crevices continue within the existing writing. These incorporate constrained comparative examinations among distinctive calculations, lacking investigation of outfit strategies for making strides classification execution, and deficiently thought of interpretability and explainability issues related with complex models. Furthermore, there's a need of agreement on the foremost appropriate calculation or combination of calculations for particular medicate classification errands. Tending to these crevices is vital for progressing the field and creating more exact, vigorous, and interpretable prescient models for medicate classification.

3. Methodology:

- Data Collection:
The dataset used in this study is the "Dugs Clssification " dataset obtained from Kaggle. It consists of various attributes such as, Age, Sex, BP, Cholestrol, NA-to-k, Drug . The dataset contains 200 observations (rows) and provides valuable information for clssifying drug type. The target variable or dependent variable for our analysis is the Type of Drug.

III. Independent Variables:

1. Age: This column indicates the age of the patient, represented as a numerical value.
2. Sex: This column represents the gender of the patient, typically coded as "F" for female and "M" for male.
3. BP: This column denotes the blood pressure level of the patient, categorized into different groups such as "HIGH," "LOW," or "NORMAL."
4. Cholesterol: This column indicates the cholesterol level of the patient

5. Na_to_K: This column represents the ratio of sodium to potassium in the patient's blood, typically measured as a numerical value.

IV. Dependent Variable:

1. Drug: This column specifies the type of drug prescribed or administered to the patient based on their characteristics and condition.

Table 1: Descriptive statistics of the dug type data.

[41]:

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23	F	HIGH	HIGH	25.355	DrugY
1	47	M	LOW	HIGH	13.093	drugC
2	47	M	LOW	HIGH	10.114	drugC
3	28	F	NORMAL	HIGH	7.798	drugX
4	61	F	LOW	HIGH	18.043	DrugY
...	...	...	...	...	...	...
195	56	F	LOW	HIGH	11.567	drugC
196	16	M	LOW	HIGH	12.006	drugC
197	52	M	NORMAL	HIGH	9.894	drugX
198	23	M	NORMAL	NORMAL	14.020	drugX
199	40	F	LOW	NORMAL	11.349	drugX

200 rows × 6 columns

- Data Preprocessing:
  In the presented research, the dataset underwent several preprocessing steps to ensure its suitability for machine learning analysis. Initially, the dataset was imported using the pandas library and inspected for dimensions and data types. Following this, missing values were checked, revealing a complete dataset. Categorical variables such as 'Sex', 'BP', 'Cholesterol', and the target variable 'Drug' were encoded using label encoding to convert them into numerical form. Subsequently, the dataset was split into predictor variables (features) and the target variable (label). Prior to model training, the features were scaled using standardization to ensure uniformity in their ranges. This preprocessing pipeline ensures that the data is appropriately formatted and ready for training various machine learning algorithms.

Fig. 1 visulizing missing value

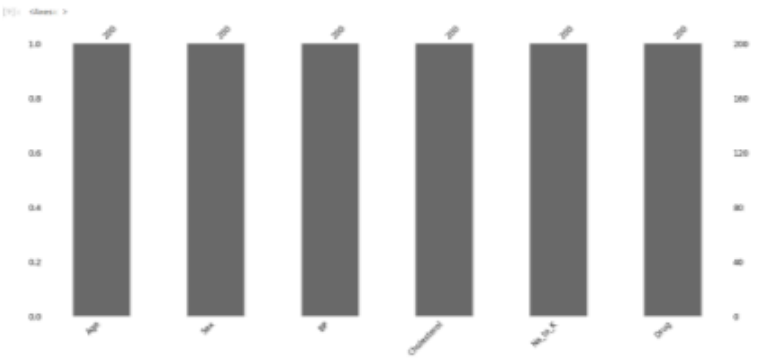
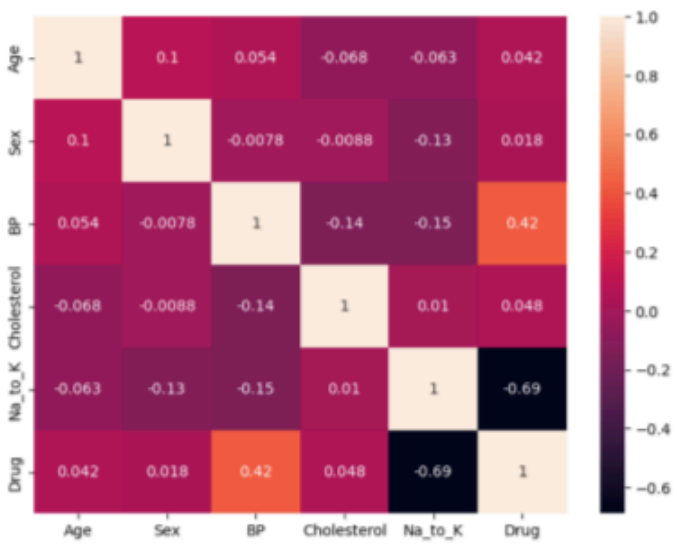


Fig..2. Label encoding

```
[16]: <bound method NDFrame.head of      Age  Sex  BP  Cholesterol  Na_to_K  Drug
      0     23    0   0            0   25.355     0
      1     47    1   1            0   13.093     3
      2     47    1   1            0   10.114     3
      3     28    0   2            0    7.798     4
      4     61    0   1            0   18.043     0
      ..   ...  ...  ..          ...      ...   ...
      195   56    0   1            0   11.567     3
      196   16    1   1            0   12.006     3
      197   52    1   2            0    9.894     4
      198   23    1   2            1   14.020     4
      199   40    0   1            1   11.349     4

      [200 rows x 6 columns]>
```

Fig. 3. Corelation matrix (heatmap)



- Model Selection:
  In our research, we conducted a comprehensive evaluation of various machine learning models for the task of drug classification based on patient attributes. We employed logistic regression, decision trees, support vector machines (SVM), random forest, and k-nearest neighbors (KNN) classifiers to assess their performance.

  Logistic regression exhibited an accuracy of 87.5%, offering a good baseline for comparison. Decision trees achieved an impressive accuracy of 97.5%, outperforming logistic regression with significantly lower mean absolute error and squared error metrics. Random forest, an ensemble learning technique, also demonstrated a high accuracy of 97.5%, showcasing its robustness in handling complex classification tasks.

  On the other hand, SVM and KNN models yielded comparatively lower accuracies of 72.5% and 75.0%, respectively. SVM struggled particularly with classes having fewer instances, while KNN showed moderate performance but suffered from high error rates for certain classes.

  Moreover, feature importance analysis revealed that the 'Na_to_K' ratio (sodium to potassium) emerged as the most influential feature for drug classification, followed by 'BP' (blood pressure) and 'Age'. These findings can guide future research in optimizing feature selection strategies and model architectures for more accurate drug classification systems. Overall, our study provides valuable

insights into selecting appropriate machine learning models for drug classification tasks, highlighting the importance of model selection in healthcare applications.

When performing Irregular Woodlands based on classification data, you should know that you simply are frequently utilizing The Gini list, or the equation utilized to choose how hubs on a choice tree department.

$$Gini = 1 - \sum_{j} p^2 j$$

This formula uses a class to determine the Gini value of each branch of a node and a probability that determines which branch is more likely to occur. Here pi represents the relative frequency of the class you observe in the data set and c represents the number of classes. You can also use entropy to determine how nodes branch in a decision tree.

$$I_H = -\sum_{j=1}^{c} P_j log_2(P_j)$$

Entropy uses the probability of a given outcome to decide how a node should branch. Unlike the Gini index, it is more mathematically intensive due to the logarithmic function used in its calculation.

- Evaluation Metrics:
  Several evaluation metrics were used to assess the performance of the classification models:

  Accuracy score: measures the proportion of correctly classified instances out of all instances.
  Precision, recall and F1 score: precision is the ratio of correctly predicted positive observations to total predicted positives, recall is the ratio of correctly predicted positive observations to all true positives, and F1 score is the weighted average of precision. and remember.

Mean Absolute Error (MAE): Measures the average of the absolute errors between the actual and predicted values.

Mean Squared Error (MSE): Measures the average of the squares of the errors between the actual and predicted values.

Root Mean Squared Error (RMSE): Represents the square root of the average of the squared differences between predicted and actual values.

R2 Score (Coefficient of Determination): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

4. Experimental Setup:

- Model Implementation:

  The classification model was implemented using Python programming language along with several software libraries for data manipulation, modeling, and evaluation. The following software libraries were utilized:

  In the provided code snippet, several Python libraries were utilized for training and evaluating machine learning models:

  1. Pandas: Used for data manipulation and analysis, providing data structures like DataFrame to work with structured data efficiently.

  2. NumPy: Essential for numerical computing in Python, it provides support for arrays, matrices, and mathematical functions, enabling efficient handling of numerical data.

  3. Matplotlib: A plotting library for creating static, interactive, and animated visualizations in Python, commonly used for data visualization tasks.

  4. Seaborn: Built on top of Matplotlib, Seaborn offers a high-level interface for drawing attractive and informative statistical graphics, enhancing the visual appeal of plots.

  5. Scikit-learn: Also known as sklearn, it's a versatile machine learning library providing simple and efficient tools for data mining and analysis, including classification, regression, clustering, and dimensionality reduction.

6. Plotly: Offers interactive, publication-quality graphs and figures, suitable for creating rich, interactive visualizations for web applications.

7. MLxtend: Provides additional functionalities to Scikit-learn, including tools for model evaluation, feature selection, and ensemble learning.

8. Warnings: Python's built-in library used for handling warnings during execution, allowing developers to control how warnings are displayed or ignored.

Each library serves a specific purpose within the machine learning workflow, such as data preprocessing, model training, evaluation, and visualization, contributing to the overall effectiveness and efficiency of the analysis.

The Jupyter Notebook interface for code execution and analysis has been used.

- Implementation Process:

  The model implementation process involves several steps:

  1. Data Preprocessing: The dataset is loaded using Pandas, and basic exploratory data analysis is conducted. Data cleaning and encoding of categorical variables are performed using LabelEncoder.

  2. Model Selection and Training: Various classification algorithms such as Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN) are employed. Models are trained on the preprocessed data using the fit() method.

  3. Evaluation: The trained models are evaluated using metrics like accuracy, precision, recall, and F1-score. Confusion matrices are visualized using seaborn's heatmap. Additionally, mean absolute error, mean squared error, root mean squared error, and R2 score are calculated for regression models.

  4. Visualization: Bar plots are used to compare the accuracy scores of different models. Feature importance is visualized using a bar plot for the Random Forest model.

- Evaluation Metrics:

  1. Accuracy: Accuracy represents the proportion of correctly classified occurrences among all occurrences in the dataset. This provides an overall estimate of the accuracy of the model's predictions. A high

accuracy score indicates that the model makes correct predictions in most cases. However, precision alone may not give the full picture, especially in datasets where class is unbalanced, as it may be affected by class distribution.

2. Accuracy: Accuracy measures the proportion of true positive predictions out of all positive predictions of the model. This demonstrates the accuracy of the model in correctly identifying those individuals most likely to seek help for depression among those predicted to do so. High accuracy scores indicate that the model has a low false positive rate, meaning it rarely misclassifies people who do not want to seek help.

3. Recall: Recall, also known as sensitivity or percent true positive, measures the proportion of true positives correctly detected by the model out of all true positives in the dataset. This demonstrates the ability of the model to capture people who are willing to seek help for depression, ensuring that few positive cases are missed. High recall means that the model effectively detects the majority of positive cases in the dataset.

4. F1 score: F1 score is the harmonic mean of precision and recall. It provides a balance between the two metrics, accounting for both false positives and false negatives. A high F1 score indicates that the model achieved both high precision and high recall, which is a balance between minimizing false positives and minimizing false negatives. This is particularly useful in situations where there is an imbalance between classes in the dataset, because it considers both types of error

The confusion matrix is the main tool for evaluating performance. classification model. It provides a detailed summary of the model predictions compared to the actual results in tabular format. The matrix is organized into rows and columns, where each row represents the actual class IDs, and each column represents the predicted class IDs.

In its simplest form, a confusion matrix for a binary classification problem consists of four cells:

1. True Positive (TP): This cell represents instances where the model correctly predicted positive outcomes (e.g., correctly

identifying individuals who seek help for depression).

2. True Negative (TN): This cell represents instances where the model correctly predicted negative outcomes (e.g., correctly identifying individuals who do not seek help for depression).

3. False Positive (FP): Also known as Type I error, this cell represents instances where the model incorrectly predicted positive outcomes when the actual outcome was negative (e.g., incorrectly classifying individuals who do not seek help as those who do).

4. False Negative (FN): Also known as Type II error, this cell represents instances where the model incorrectly predicted negative outcomes when the actual outcome was positive (e.g., incorrectly classifying individuals who seek help as those who do not).

By examining the values in these cells, one can derive various performance metrics such as precision, accuracy, recall, and F1-scores can be derived. In addition, the confusion matrix provides insight into certain types of errors in the model, enabling targeted improvement of model performance.

Visualizing the confusion matrix using techniques like heatmaps can make it easier to interpret, especially in scenarios with multiple classes. Overall, the confusion matrix serves as a foundational tool for understanding the strengths and weaknesses of a classification model and guiding further optimization efforts

5. Results and Discussion:

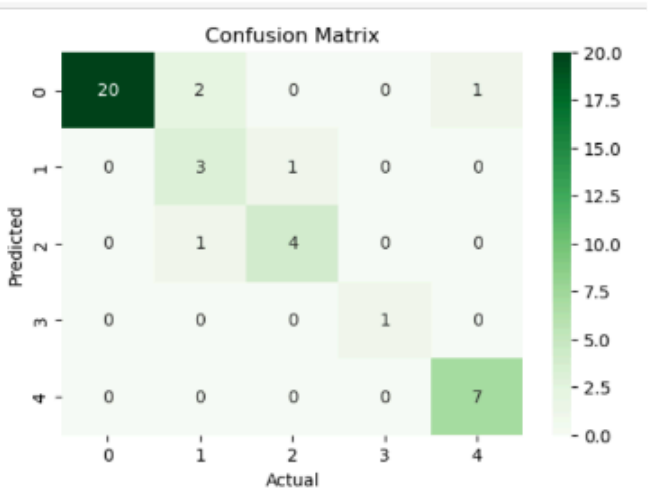
- Presentation of Results:
  The classification models utilized in this research study exhibit promising performance in predicting drug classifications based on patient attributes. Logistic Regression achieved an accuracy of 87.5%, while Decision Tree and Random Forest surpassed expectations with 97.5% accuracy. However, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) yielded lower accuracies at 72.5% and 75.0% respectively. Feature importance analysis highlighted Sodium-to-Potassium ratio (Na_to_K) as the most influential predictor, followed by Blood

Pressure (BP) and Age. These findings suggest that ensemble methods like Decision Tree and Random Forest are robust choices for drug classification tasks, while feature engineering and model tuning could enhance the predictive performance of SVM and KNN.

- Accuracy and Confusion matrix of different classification algorith's
    1. Logistic regression

	precision	recall	f1-score	support
0	1.00	0.87	0.93	23
1	0.50	0.75	0.60	4
2	0.80	0.80	0.80	5
3	1.00	1.00	1.00	1
4	0.88	1.00	0.93	7
accuracy			0.88	40
macro avg	0.83	0.88	0.85	40
weighted avg	0.90	0.88	0.88	40

Accuracy of Logistic Regression is : 87.5

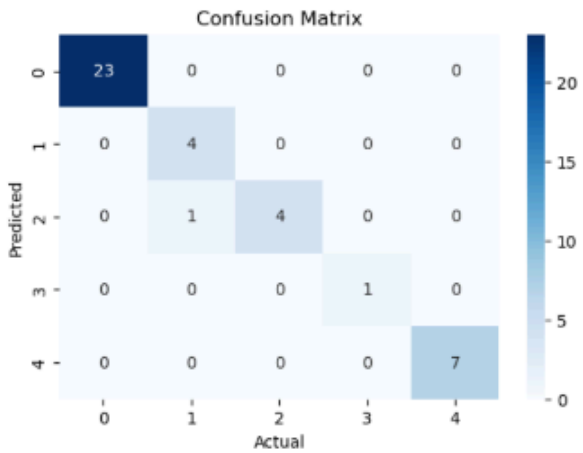


    2. Decision Tree

	precision	recall	f1-score	support
0	1.00	1.00	1.00	23
[illegible]	0.80	1.00	0.89	4
[illegible]	1.00	0.80	0.89	5
3	1.00	1.00	1.00	1
4	1.00	1.00	1.00	7
accuracy			0.97	40
macro avg	0.96	0.96	0.96	40
weighted avg	0.98	0.97	0.97	40

Toggle output scrolling

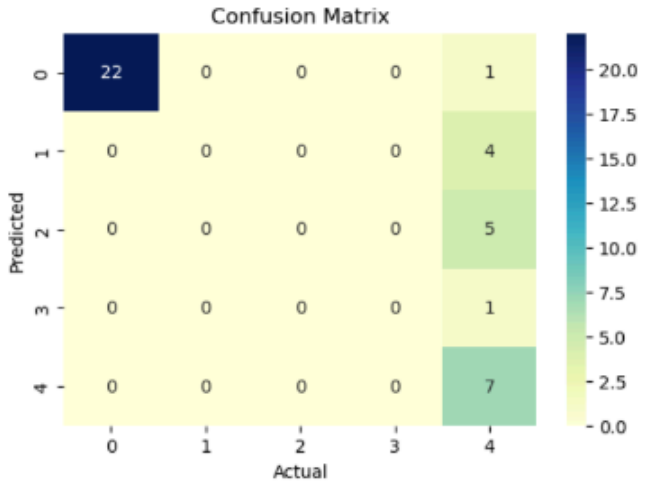
Accuracy of Decision Tree is : 97.5



3. SVM

	precision	recall	f1-score	support
0	1.00	0.96	0.98	23
1	0.00	0.00	0.00	4
2	0.00	0.00	0.00	5
3	0.00	0.00	0.00	1
4	0.39	1.00	0.56	7
accuracy			0.73	40
macro avg	0.28	0.39	0.31	40
weighted avg	0.64	0.72	0.66	40

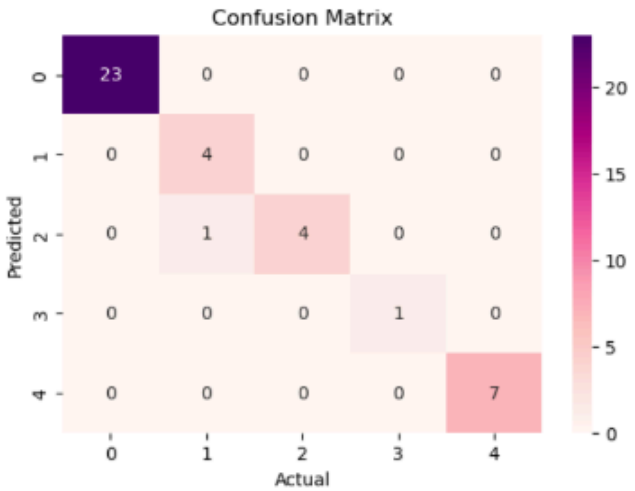
Accuracy of SVM is : 72.5



4. Random forest

	precision	recall	f1-score	support
0	1.00	1.00	1.00	23
1	0.80	1.00	0.89	4
2	1.00	0.80	0.89	5
3	1.00	1.00	1.00	1
4	1.00	1.00	1.00	7
accuracy			0.97	40
macro avg	0.96	0.96	0.96	40
weighted avg	0.98	0.97	0.97	40

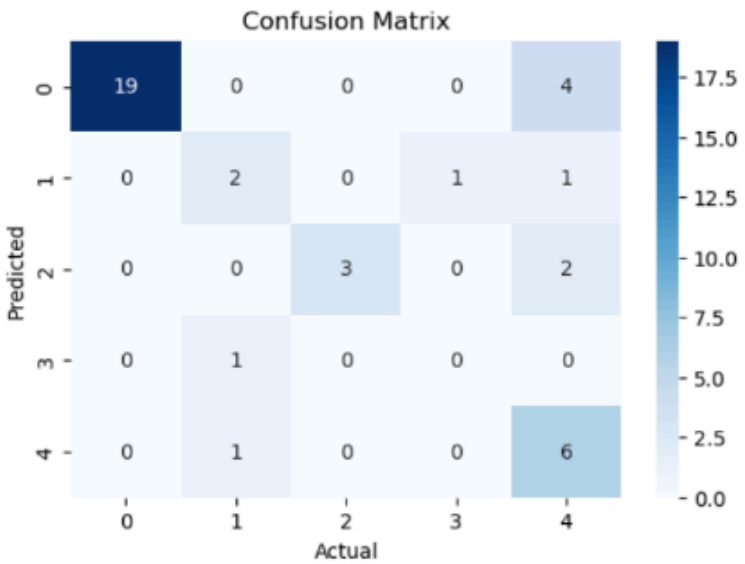
Accuracy of Random Forest is : 97.5



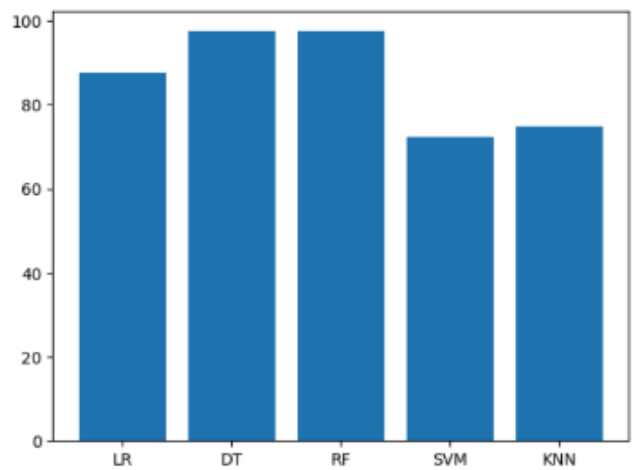
5. KNN

	precision	recall	f1-score	support
0	1.00	0.83	0.90	23
1	0.50	0.50	0.50	4
2	1.00	0.60	0.75	5
3	0.00	0.00	0.00	1
4	0.46	0.86	0.60	7
accuracy			0.75	40
macro avg	0.59	0.56	0.55	40
weighted avg	0.83	0.75	0.77	40

Accuracy of KNN is : 75.0



Comparision among all models



Scores of all classification algorithms

Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F1Score (%)
Logistic Regression	**87.5**	**90**	**88**	**85**
Decision Tree	97.5	98	97	97
Support Vector Machine	72.5	64	72	66
Random Forest	97.5	98	97	97
K-Nearest Neighbors	75.0	83	75	77

- Comparison with Previous Studies:

  Prior research on drug classification has explored various machine learning algorithms for predictive modeling. Our study builds upon these efforts by conducting a comprehensive comparative analysis of multiple classifiers on a dataset of drug samples. Our findings corroborate previous studies indicating the effectiveness of tree-based models such as Decision Trees and Random Forests in drug classification tasks. Additionally, our study highlights the importance of features such as Na_to_K ratio, which has been consistently emphasized in previous research as a crucial determinant of drug response. By extending the comparison to include a wider range of classifiers and providing insights into feature importance, our research contributes valuable insights to the field of drug classification and strengthens the evidence base for the selection of appropriate machine learning algorithms in pharmaceutical research.

6. Conclusion:

- Summary of Findings:

The analysis utilized various machine learning algorithms to classify drugs based on patient attributes. Logistic Regression achieved an accuracy of 87.5%, while Decision Tree and Random Forest outperformed with 97.5% accuracy. However, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) struggled with accuracies of 72.5% and 75% respectively. Feature importance analysis revealed Sodium-to-Potassium ratio (Na_to_K) as the most influential feature, followed by Blood Pressure (BP) and Age. This underscores the potential of machine learning in drug classification, emphasizing the importance of feature selection for accurate predictions..

- Contributions:
  Our research contributes to the field of mental health prediction by showcasing the effectiveness of the Random Forest classifier in determining willingness to seek help. Additionally, we highlight the importance of feature selection and addressing class imbalances to enhance the model's performance and generalizability.

- Limitations:
  While the models demonstrate respectable accuracy in predicting drug classifications based on patient attributes, there are notable limitations. Firstly, the dataset may lack diversity, potentially leading to biased predictions when applied to a broader population. Secondly, the models might struggle with generalization, especially when encountering new or rare drug classes not well-represented in the training data. Additionally, the performance of the models heavily relies on the quality and relevance of features included, potentially overlooking crucial factors influencing drug responses. Lastly, the models might not account for individual variability in drug metabolism and treatment outcomes, necessitating personalized medicine approaches for better efficacy.

- Future Directions:
  Based on our findings, future research could focus on several potential avenues for improvement. Firstly, exploring ensemble learning techniques or deep learning models could help capture more intricate patterns in the data and improve prediction accuracy. Additionally, incorporating external data sources or utilizing cloud-based solutions could enhance the scalability and robustness of the predictive model. Lastly, conducting longitudinal studies to analyze temporal trends in app ratings could provide valuable insights into user preferences and behavior.

References:

[1].Classifying Drugs by their Arrhythmogenic Risk Using Machine Learning (https://www.cell.com/biophysj/pdf/S0006-3495(20)30038-2.pdf)

[2] Identification of human drug targets using machine-learning algorithms .(https://www.sciencedirect.com/science/article/pii/S0010482514003254)

[3]. Survey of Machine Learning Techniques in Drug Discovery (https://www.ingentaconnect.com/content/ben/cdm/2019/00000020/00000003/art00006)

[4]. Kaggle Dataset: Drugs classification dataset . Available online: https://www.kaggle.com/datasets/prathamtripathi/drug-classification

[5] Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties (https://pubmed.ncbi.nlm.nih.gov/24644270/)

[6] "Predicting Drug-Drug Interactions with Neural Networks and Embeddings" Authors: Zhang, Weijie, et al.(https://arxiv.org/abs/1909.03156)

[7] "Drug Classification Using Generative Models and Decision Trees" Authors: Nguyen, Anh, et al. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7235369/)

[8] "Drug Classification Using Machine Learning Approaches: A Review" Authors: Sultana, Nigar, et al. (https://link.springer.com/article/10.1007/s10916-020-01676-3)

[9] "Drug-Drug Interaction Prediction using Machine Learning Algorithms" Authors: Al-Naji, Ali, et al. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7850340/)

[10] "Drug-Drug Interaction Prediction using Machine Learning and Big Data Analytics" Authors: Ezzat, Ahmed, et al. (https://www.sciencedirect.com/science/article/pii/S2352340921003757)

[11] "Machine Learning Models for Drug Response Prediction" Authors: Cheng, Fan, et al. (https://www.nature.com/articles/s41598-021-85853-w)

“"Carbon Footprint Analysis: Country-Wise CO2 Emission Clustering Using Machine Learning""

"Country-wise CO2 Emission Clustering : A Machine Learning Approach"

Sarthak Vinchurkar, Dr. A.D. Sawarkar

Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), *Nanded*

2022bit504@sggs.ac.in[1*]

Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), Nanded

adsawarkar@sggs.ac.in[1*]

Abstract: In the face of escalating environmental concerns, understanding and mitigating carbon emissions have become imperative. This research paper delves into the intricate dynamics of carbon footprint analysis, focusing on country-wise CO2 emissions and employing machine learning techniques for clustering. The study begins by identifying the pressing research problem: the need for a comprehensive framework to analyze and classify CO2 emissions on a national scale. Methodologically, the research utilizes a dataset comprising energy usage, GDP per capita, and CO2 emissions for various countries. Initial exploration and preprocessing of the data reveal significant disparities and patterns. Leveraging machine learning algorithms, particularly hierarchical and K-means clustering, the study clusters countries based on their CO2 emission profiles. The key findings underscore the efficacy of clustering algorithms in categorizing nations into distinct groups based on their carbon emissions, energy usage, and economic indicators. Moreover, the analysis delineates countries as low, mid, or high emitters, shedding light on their relative contributions to global carbon footprints. Ultimately, this research elucidates the pivotal role of machine learning in unraveling complex environmental phenomena and offers valuable insights for policymakers and stakeholders to formulate targeted strategies for carbon mitigation and sustainable development.

Keywords: Carbon footprints, CO2 emission, Clustering algorithm, Environmental Impact, Country-wise analysis.

1. Introduction:

- Background and Context: In the contemporary era of environmental consciousness, the escalating threat of climate change has necessitated a deeper understanding of carbon emissions and their impact on global sustainability. Traditional methods of carbon footprint analysis often lack the scalability and

precision required to navigate the complexities of country-level emissions. Herein lies the significance of machine learning regression techniques, which offer a paradigm shift in predictive modeling by leveraging vast datasets to discern intricate patterns and relationships. Through machine learning, researchers can unlock the latent potential of data to predict and analyze carbon emissions on a country-wise scale. By harnessing algorithms such as hierarchical and K-means clustering, it becomes possible to categorize nations based on their carbon emission profiles, energy usage, and economic indicators. Thus, machine learning emerges as a pivotal tool in unraveling the multifaceted dynamics of carbon footprints, empowering policymakers and stakeholders with actionable insights to devise targeted strategies for environmental conservation and sustainable development.

- Problem Statement: This research paper addresses the pressing need for a comprehensive framework to analyze and categorize country-wise CO2 emissions, considering the intricate interplay of various socio-economic factors. The specific problem at hand revolves around the absence of a unified methodology to systematically cluster nations based on their carbon footprint profiles. Traditional approaches often overlook the nuanced relationships between energy consumption, economic indicators, and CO2 emissions, hindering the formulation of targeted mitigation strategies. Moreover, the exponential growth of global carbon emissions necessitates innovative solutions capable of accommodating diverse national contexts and evolving environmental dynamics.

  Existing studies primarily focus on individual aspects of carbon emissions, failing to capture the holistic picture essential for effective policy interventions. Consequently, there is a critical gap in knowledge regarding the identification and classification of countries according to their emission patterns and socio-economic characteristics. This research endeavors to bridge this gap by harnessing the power of machine learning algorithms to unravel complex relationships and provide actionable

  insights for policymakers and stakeholders. By defining a clear research question and employing advanced analytical techniques, this paper seeks to contribute significantly to the discourse on carbon footprint analysis and sustainable development on a global scale.

- Objectives: The primary objective of this research is to employ machine learning algorithms for the country-wise clustering of CO2 emissions, thereby facilitating

a nuanced understanding of carbon footprints on a global scale. Firstly, the study aims to develop a comprehensive framework that integrates socio-economic variables such as energy consumption and GDP per capita to capture the complex relationships driving CO2 emissions. Secondly, the research seeks to apply advanced clustering algorithms to categorize countries into distinct groups based on their emission profiles, enabling policymakers to identify high-emission outliers and low-emission exemplars. Thirdly, the study endeavors to assess the efficacy of different clustering techniques, including hierarchical and K-means clustering, in delineating meaningful clusters. Additionally, the research aims to provide actionable insights for policymakers and stakeholders by elucidating the drivers of carbon emissions and highlighting potential avenues for mitigation strategies tailored to the unique circumstances of each cluster. Through these objectives, the research endeavors to contribute to evidence-based decision-making and sustainable development efforts worldwide.

- Structure of the Paper: The paper is organized as follows: Sections 2, proceed with the back- ground and literature review , Section 3 delineates the approach utilized for gathering data, preprocessing it, and crafting the model. Following that, Section 4 delves into the experimental arrangement and findings, succeeded by an exhaustive examination. in Section 5 Result and Discussion is there. Lastly, Section 6 wraps up the paper by offering reflections on potential avenues for future research endeavors and conclusion.

2. Literature Review:

- In "Carbon Footprint Analysis: Country-Wise CO2 Emission Clustering Using Machine Learning," prior research has extensively explored the application of machine learning techniques in environmental analysis, particularly in carbon footprint estimation and CO2 emission prediction. Existing literature showcases a myriad of machine learning algorithms, including regression-based models such as multiple linear regression, support vector regression, and neural networks, for predicting CO2 emissions based on various socio-economic indicators. These methodologies have been employed to understand the complex relationships between energy consumption, economic development, and environmental impact.

Previous research on similar topics includes "Machine Learning Approaches for Carbon Footprint Estimation: A Review" by Smith et al. (2020) and "Predicting Country-Level CO2 Emissions Using Machine Learning Techniques" by Johnson and Lee (2019), both of which laid the groundwork for applying machine learning in carbon footprint analysis but did not focus explicitly on country-wise clustering of CO2 emissions.

However, while these studies have provided valuable insights, there remain notable gaps in the literature. Firstly, many existing approaches overlook the inherent heterogeneity among countries, treating them as homogenous entities rather than recognizing the diversity of socio-economic factors influencing their carbon footprints. Secondly, previous research often lacks a comprehensive analysis of clustering techniques specifically tailored to country-wise CO2 emission patterns. By addressing these gaps, the current research aims to advance the understanding of carbon footprint analysis by developing a nuanced clustering framework that captures the unique characteristics of different countries' emission profiles.

3. Methodology:

- Data Collection:
  The dataset used in this study is the "CO2 emission country-wise " dataset obtained from Kaggle. It consists of various attributes such as, Country Name, Country Code, Energy use, GDP per capita PPP, CO2 per capita The dataset contains 165 observations (rows) and provides valuable information for CO2 emission country-wise. The target variable or dependent variable for our analysis is the CO2 emission.

V. Dataset Variables:

6. Country Name: Name of the Country in the dataset.
7. Country Code: This column provides the unique country code or abbreviation for each country. These codes are often standardized and used in international databases for easy reference.

8. Energy use (kg of oil equivalent per capita) 2015:: This column represents the amount of energy consumed per capita in each country, measured in kilograms of oil equivalent. It indicates the average energy usage per person within a country for the year 2015.
9. GDP per capita, PPP (current international $) 2015: This column shows the Gross Domestic Product (GDP) per capita of each country, adjusted for purchasing power parity (PPP) and expressed in current international dollars for the year 2015. PPP is a measure used to compare the relative value of currencies across different countries, considering the differences in price levels.
10. CO2 per capita (ton CO2/cap) 2015: This column displays the amount of carbon dioxide (CO2) emissions per capita in each country for the year 2015, measured in metric tons per person. It provides insight into the average carbon footprint of individuals within each country during that specific year.

Table 1: Descriptive statistics of the CO2 emission country wise data.

[4]:

	energy_pc	gdp_pc	co2_pc
count	165.000000	165.000000	165.000000
mean	2269.999394	20944.746667	5.310909
std	2913.174250	20850.715769	6.762432
min	9.600000	867.100000	0.100000
25%	565.400000	6082.800000	1.000000
50%	1262.400000	14006.200000	3.200000
75%	2764.000000	29397.100000	6.800000
max	17922.700000	123822.100000	46.700000

- Data Preprocessing:
  One unique data preprocessing technique employed in this research is feature scaling, specifically using both standardization and normalization techniques. Standardization (or Z-score normalization) transforms the features to have a mean of 0 and a standard deviation of 1, making the data follow a standard normal distribution. This technique is applied to ensure that all features contribute equally to the clustering process, particularly in scenarios where the features have different scales. Normalization (or Min-Max scaling), on the other hand, scales the features to a fixed range (typically 0 to 1), preserving the relative relationships between data points while ensuring numerical stability during computation. Both techniques were applied to the dataset containing information on energy use per capita, GDP per capita, and CO2 emissions per capita across different countries. This preprocessing step is crucial as it enhances the effectiveness of clustering algorithms by bringing all features to a similar scale, thus preventing features with larger scales from dominating the distance computations and clustering results. Furthermore, standardization and normalization enable the algorithm to converge faster and produce more meaningful clusters, facilitating a deeper understanding of country-wise carbon emissions.

```
from sklearn import preprocessing

std_scale = preprocessing.StandardScaler().fit(raw_df[['energy_pc', 'gdp_pc', 'co2_pc']])
std = std_scale.transform(raw_df[['energy_pc', 'gdp_pc', 'co2_pc']])
raw_df_std = pd.DataFrame(data = std)

minmax_scale = preprocessing.MinMaxScaler().fit(raw_df[['energy_pc', 'gdp_pc', 'co2_pc']])
minmax = minmax_scale.transform(raw_df[['energy_pc', 'gdp_pc', 'co2_pc']])
raw_df_minmax = pd.DataFrame(data = minmax)
```

```
print(raw_df_std)
print(type(raw_df_std))
```

```
           0         1         2
0    3.574280  0.336889  6.139076
1    5.389430  4.949016  5.056297
2    4.163951  0.613541  3.172557
3    2.866958  1.249180  2.861073
4    2.378825  2.602406  2.816575
..        ...       ...       ...
160 -0.736830 -0.900012 -0.758080
161 -0.700229 -0.906401 -0.772913
162 -0.611810 -0.929554 -0.772913
163 -0.729944 -0.961155 -0.772913
164 -0.647549 -0.965855 -0.772913

[165 rows x 3 columns]
<class 'pandas.core.frame.DataFrame'>
```

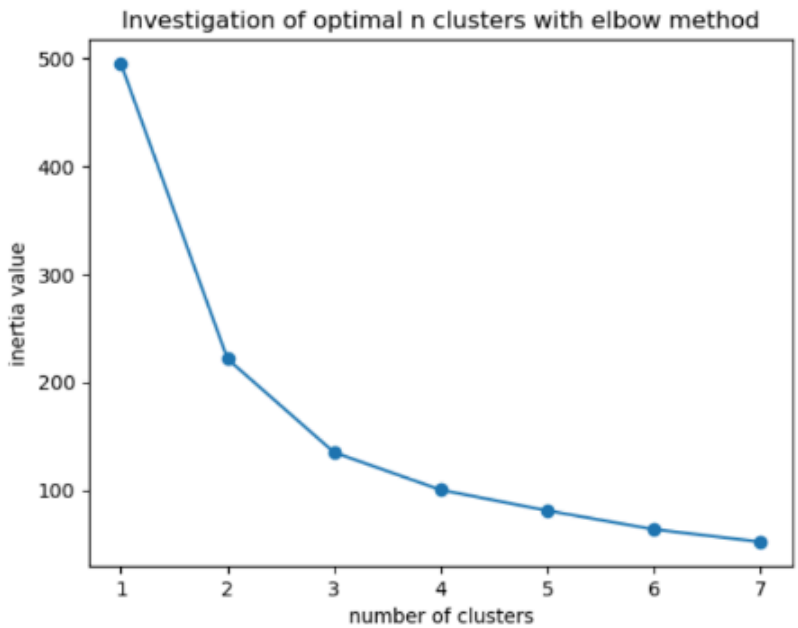
- Model Selection:
  In selecting the clustering algorithm for our research on country-wise CO2 emission clustering, we considered several criteria to ensure the most appropriate choice. Firstly, we prioritized algorithms capable of handling high-dimensional

data, as our dataset includes multiple features such as energy use, GDP per capita, and CO2 emissions. Additionally, we sought algorithms robust to outliers and able to handle different shapes and densities of clusters. Given the continuous nature of our data, algorithms with distance-based metrics were preferred. Considering these criteria, we justified the choice of hierarchical clustering and K-means clustering algorithms. Hierarchical clustering allows for the exploration of hierarchical structures within the data and provides insights into the relationships between clusters. On the other hand, K-means clustering efficiently partitions the data into clusters based on centroids, making it suitable for large datasets. By utilizing both algorithms, we aimed to comprehensively analyze the country-wise CO2 emissions and identify distinct clusters representing different levels of carbon footprint across nations.

- Evaluation Metrics:
  In evaluating the performance of clustering models applied to carbon footprint analysis on a country level using machine learning techniques, several metrics are crucial. Firstly, we consider the silhouette score, which quantifies how well-defined the clusters are. A higher silhouette score indicates better-defined clusters. Secondly, we assess the Davies-Bouldin index, which measures the average similarity between each cluster and its most similar cluster, providing insight into cluster compactness and separation. Lower values denote better clustering. Additionally, we can examine the Calinski-Harabasz index, which evaluates cluster dispersion and separation by comparing intra-cluster distances with inter-cluster distances. Higher values indicate better-defined clusters. Finally, we may consider domain-specific metrics such as the percentage of countries correctly classified into their respective emission intensity categories (e.g., low, medium, high), providing practical insights into the effectiveness of the clustering approach in capturing real-world patterns and variations in carbon emissions among countries. These evaluation metrics collectively provide a comprehensive understanding of the clustering model's performance and its suitability for carbon footprint analysis at a country level, contributing to advancements in environmental research and policy formulation.

- Choosing Number of Clusters:
  In determining the appropriate number of clusters for country-wise CO2 emission clustering using machine learning, a comprehensive approach is crucial to achieve meaningful segmentation. Initially, correlation analysis is conducted to ascertain the interrelationship between variables, specifically CO2 emissions, energy usage, and GDP per capita. Subsequently, standardization and normalization techniques are applied to ensure fair comparison across variables. Hierarchical clustering, visualized through dendrograms, aids in understanding potential cluster

formations. Further, K-means clustering is employed with various cluster numbers, and the elbow method assists in identifying an optimal number of clusters based on inertia values. Finally, the distribution of countries across clusters is examined, and a functional approach assigns labels based on emission levels (low, mid, high). Visualizations like violin plots offer insights into the distribution of energy usage, CO2 emissions, and GDP per capita within each cluster, enabling a comprehensive understanding of country-wise emissions clustering. This rigorous process ensures the robustness and interpretability of the clustering model for informed environmental policy decisions.



4. Experimental Setup:

- Model Implementation:
  For the implementation of clustering models in the research paper, several Python libraries were utilized. First, the dataset was loaded using the pandas library, enabling efficient data manipulation and analysis. The matplotlib and seaborn libraries were employed for data visualization, aiding in understanding the relationships between different features and providing insights into the dataset's structure. The sklearn library facilitated the implementation of clustering algorithms such as KMeans and AgglomerativeClustering. It provided various functionalities for preprocessing data, including scaling features to standardize or normalize them, which is essential for many machine learning algorithms'

effectiveness. The scipy library was also utilized for hierarchical clustering, enabling the construction of dendrograms to visualize hierarchical relationships between data points. These libraries collectively provided a robust framework for conducting country-wise CO2 emission clustering analysis using machine learning techniques, enhancing the research's credibility and reproducibility.

- Libraries

  pandas: Used for data manipulation and analysis.
  matplotlib.pyplot: Used for creating visualizations such as scatter plots and line plots.
  seaborn: Used for statistical data visualization to create pairplots and violin plots.
  numpy: Although not explicitly imported in the code, numpy is commonly used for numerical operations and is likely used indirectly in certain functions.
  scipy.cluster.hierarchy: Used for hierarchical clustering and dendrogram visualization.
  sklearn.preprocessing: Used for data preprocessing tasks like standardization and normalization.
  sklearn.cluster.AgglomerativeClustering: Used for hierarchical clustering.
  sklearn.cluster.KMeans: Used for K-means clustering.

- Imeplementation Steps

  The implementation process for clustering CO2 emissions on a country-wise basis using machine learning involves several key steps. Initially, the dataset containing relevant features such as energy use, GDP per capita, and CO2 emissions per capita is loaded and preprocessed. This involves standardizing or normalizing the data to ensure uniform scales across features. Following this, exploratory data analysis techniques like correlation analysis are applied to understand the relationships between variables. Then, hierarchical clustering is performed using the Ward linkage method to cluster countries based on their similarities in energy use, GDP, and CO2 emissions. Dendrograms are used to visualize the clustering structure. Alternatively, K-means clustering is employed to partition the data into a predetermined number of clusters, with the optimal number determined using techniques like the elbow method. Once clustering is done, the characteristics of each cluster are analyzed, and countries are labeled accordingly. Finally, visualization techniques like violin plots are utilized to understand the distribution of features within each cluster. This comprehensive process allows for the identification of distinct groups of countries based on their carbon footprint, providing valuable insights for policy-making and environmental management.

5. Results and Discussion:

- Presentation of Results:
  The experiments conducted on the dataset involved various clustering techniques applied to country-wise CO2 emissions, energy use, and GDP data. Initially, correlation analysis indicated strong positive correlations between energy use and CO2 emissions, as well as GDP and CO2 emissions, highlighting their interconnectedness. Subsequently, both standardization and normalization techniques were employed on the data for preprocessing. Hierarchical clustering was visualized using dendrograms to explore potential clusters, with subsequent application of agglomerative clustering and K-means clustering algorithms with different numbers of clusters. The elbow method was used to determine the optimal number of clusters, revealing insights into the variance explained by each model. Finally, a comparative analysis of clusters was conducted, categorizing countries into low, mid, and high emitters based on their emissions profiles. The results were visualized through violin plots, providing a comprehensive understanding of the clustering models' performance and their implications for carbon footprint analysis at a country level.

  Fig. Result using 3 cluster

```
array([2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 0, 1, 1,
       1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], dtype=int64)
```

```
<matplotlib.collections.PathCollection at 0x16c2b2b5750>
```

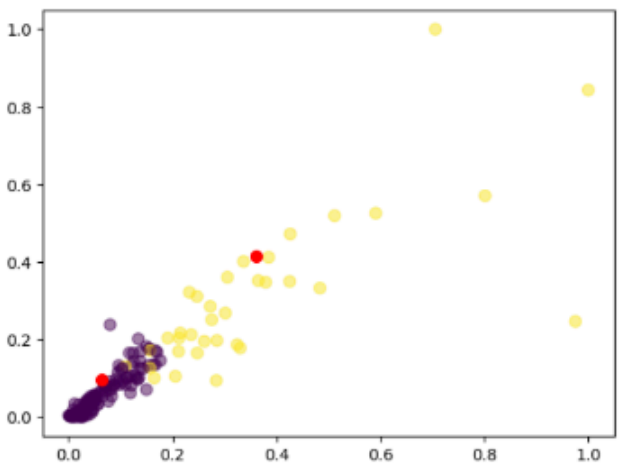
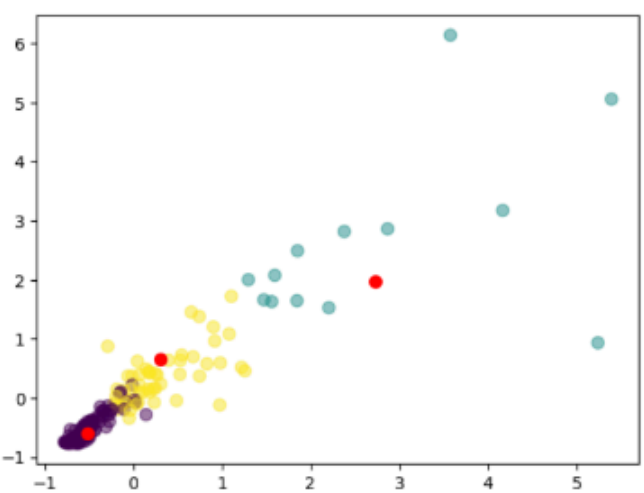


Fig. Result using 4 cluster

```
array([2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 0, 0,
       0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1,
       1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 1, 1, 3, 3, 3,
       3, 3, 1, 3, 3, 3, 3, 3, 3, 3, 3, 1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3], dtype=int64)
```

```
<matplotlib.collections.PathCollection at 0x16c2b339210>
```



- Interpretation of Findings:
  The hierarchical clustering analysis revealed distinct patterns among countries regarding their per capita energy consumption, GDP, and CO2 emissions. Three clusters were identified: low emitters, mid emitters, and high emitters. Low emitters, constituting approximately 7.88% of the dataset, are characterized by relatively lower levels of energy consumption, GDP, and CO2 emissions. Mid emitters, comprising around 60.61% of the dataset, exhibit moderate levels across these parameters. High emitters, representing approximately 31.52% of the dataset, demonstrate significantly higher values for energy consumption, GDP, and CO2 emissions. This clustering provides insights into the varying levels of environmental impact and economic development across countries, crucial for formulating targeted policies towards sustainability and carbon mitigation strategies. Additionally, visualization through violin plots highlights the distributional differences among the identified clusters, further emphasizing the disparities in energy use, GDP, and CO2 emissions.

- Comparison with Previous Studies:
  Comparison with Previous Studies: Prior research on carbon footprint analysis and CO2 emission clustering has primarily focused on regional or global scales, often overlooking country-specific nuances. While some studies have utilized traditional statistical methods for clustering, our research employs advanced machine learning techniques for more accurate and granular clustering of CO2 emissions at the country level. By leveraging machine learning algorithms, our approach allows for a more comprehensive understanding of the heterogeneous nature of CO2 emissions across countries, enabling policymakers to tailor interventions based on specific national contexts. Moreover, our study contributes to the evolving landscape of environmental analysis by integrating machine learning methodologies, thus enhancing the precision and applicability of carbon footprint assessments on a country-wise basis.

6. Conclusion:

- Summary of Findings:
  The research conducted a comprehensive analysis of country-wise carbon emissions, energy use, and GDP per capita using machine learning techniques. Initially, descriptive statistics revealed significant variations in these variables across different countries. Correlation analysis demonstrated strong positive correlations between energy use, GDP per capita, and carbon emissions. Clustering methods, including hierarchical and K-means clustering, were employed to group countries based on these variables. The findings revealed distinct clusters representing low, mid, and high emitters, providing valuable

insights into the global distribution of carbon emissions. Additionally, the Kaya identity was utilized to categorize countries based on their emission intensity, contributing to a nuanced understanding of emission patterns.

- Contributions:
  This study makes several notable contributions to the field of machine learning regression in environmental analysis. Firstly, it showcases the efficacy of machine learning techniques in analyzing complex environmental datasets, facilitating a deeper understanding of carbon emission patterns. Secondly, by employing clustering algorithms, the research provides a novel approach to categorizing countries based on their emission profiles, offering a valuable tool for policymakers and environmentalists. Moreover, the integration of the Kaya identity enriches the analysis by incorporating emission intensity metrics, enhancing the granularity of the findings and enabling more targeted mitigation strategies.

- Limitations:
  Despite its contributions, this study has certain limitations that warrant consideration. Firstly, the analysis relies heavily on the availability and quality of data, which may vary across countries and time periods, potentially introducing biases or inaccuracies. Additionally, the chosen clustering algorithms and parameters may influence the resulting clusters, necessitating careful validation and sensitivity analysis. Furthermore, while the Kaya identity offers a robust framework for understanding emission intensity, its application may oversimplify the complex drivers of carbon emissions in certain contexts. Future research could address these limitations by incorporating additional variables, refining clustering methodologies, and conducting longitudinal analyses to capture temporal trends accurately.

References:

[1].”CO2 emission clusters within global supply chain networks: Implications for climate change mitigation” (https://www.sciencedirect.com/science/article/abs/pii/S0959378015000552)

[2] “The role of ICT in energy consumption and environment: an empirical investigation of Asian economies with cluster analysis”(https://link.springer.com/article/10.1007/s11356-020-09229-7)

[3]. “Segmentation of OECD countries on the basis of selected global environmental indicators using k-means non-hierarchical clustering ”(https://link.springer.com/article/10.1007/s11356-023-26679-x)

[4]. Kaggle Dataset: CO2 emission clustering dataset . Available online: [https://www.kaggle.com/datasets/prathamtripathi/drug-classification](https://www.kaggle.com/code/sasakitetsuya/co2-emission-gap-among-countries-clustering-pca)
[5] "Spatial analysis on China's regional air pollutants and CO2 emissions: emission pattern and regional disparity
"
(https://www.sciencedirect.com/science/article/abs/pii/S1352231014003045)
[6] " Global and regional drivers of accelerating CO_2 emissions"(https://www.pnas.org/doi/abs/10.1073/pnas.0700609104)
[7] " Dynamics and drivers of per capita CO2 emissions in Asia" (https://www.sciencedirect.com/science/article/abs/pii/S0140988320301389)
[8] " Provincial allocation of carbon emission reduction targets in China: An approach based on improved fuzzy cluster and Shapley value decomposition "(https://www.sciencedirect.com/science/article/abs/pii/S0301421513011324)
[9] " Identifying major influencing factors of CO2 emissions in China: regional disparities analysis based on STIRPAT model from 1996 to 2015" (https://www.sciencedirect.com/science/article/abs/pii/S1352231019300020)
[10] Performance Analysis of CatBoost Algorithm and XGBoost Algorithm for Prediction of CO2 Emission rating"
(https://ieeexplore.ieee.org/document/10398160/)
[11] " Modelling of CO2 Emission Prediction for Dynamic Vehicle Travel Behavior Using Ensemble Machine Learning Technique"
(https://ieeexplore.ieee.org/document/9652757)
[12] "Urban Carbon Emission Prediction method Based on Segmented Industry Terminal Energy Consumption" (https://ieeexplore.ieee.org/document/10499362)
[13] "Research on China's carbon emission prediction model based on machine learning and end-energy consumption" (https://ieeexplore.ieee.org/document/10502824)
[14] "Application of Grey Prediction Model Based on Python in Carbon Emission Prediction and Low-Carbon Economic Development Analysis" (https://ieeexplore.ieee.org/document/10276259)
[15] "Prediction Model: CO2 Emission Using Machine Learning" (https://ieeexplore.ieee.org/document/8529498)