International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8355
***Article Received:** 12 July 2023 **Revised:** 07 September 2023 **Accepted:** 20 September 2023*

Harnessing Data-Driven Insights: Predictive Modeling for Diamond Price Forecasting using Regression and Classification Techniques

Md Shaik Amzad Basha[1], Peerzadah Mohammad Oveis[2]
[1]GITAM School of Business
Gandhi Institute of Technology and Management (Deemed to be university)
Bangalore, India
amjuamjad66@gmail.com
[2]GITAM School of Business
Gandhi Institute of Technology and Management (Deemed to be university)
Bangalore, India
peeroveis@gmail.com

Abstract—In the multi-faceted world of gemology, understanding diamond valuations plays a pivotal role for traders, customers, and researchers alike. This study delves deep into predicting diamond prices in terms of exact monetary values and broader price categories. The purpose was to harness advanced machine learning techniques to achieve precise estimations and categorisations, thereby assisting stakeholders in informed decision-making. The research methodology adopted comprised a rigorous data preprocessing phase, ensuring the data's readiness for model training. A range of sophisticated machine learning models were employed, from traditional linear regression to more advanced ensemble methods like Random Forest and Gradient Boosting. The dataset was also transformed to facilitate classification into predefined price tiers, exploring the viability of models like Logistic Regression and Support Vector Machines in this context. The conceptual model encompasses a systematic flow, beginning with data acquisition, transitioning through preprocessing, regression, and classification analyses, and culminating in a comparative study of the performance metrics. This structured approach underscores the originality and value of our research, offering a holistic view of diamond price prediction from both regression and classification lenses. Findings from the analysis highlighted the superior performance of the Random Forest regressor in predicting exact prices with an R2 value of approximately 0.975. In contrast, for classification into price tiers, both Logistic Regression and Support Vector Machines emerged as frontrunners with an accuracy exceeding 95%. These results provide invaluable insights for stakeholders in the diamond industry, emphasising the potential of machine learning in refining valuation processes.

Keywords- Diamond Valuation, Machine Learning Prediction, Regression Analysis, Classification Techniques, Price Stratification, Ensemble Models.

I. INTRODUCTION

In the intricate realm of gemmology, the valuation of diamonds stands as a cornerstone, influencing decisions ranging from trade and investment to consumer choices. Diamonds, often characterized by their cut, clarity, color, and carat, have held both economic and symbolic value for centuries. However, with evolving markets and a plethora of factors influencing their price, a systematic and precise method of predicting diamond prices has become increasingly pertinent. Historically, diamond prices were determined through expert assessments and benchmarked diamond rates. But with the dawn of the digital age and the accumulation of vast datasets detailing diamond attributes and their corresponding prices, a golden opportunity emerges: a machine learning application to accurately predict diamond valuations.

The significance of this problem is manifold. For traders, accurate price predictions mean better investment decisions. For consumers, it translates to informed purchases and for researchers, it offers a rich avenue to explore the interplay of data science and gemmology. While several studies have touched upon the domain of diamond price prediction, the majority have either focused on traditional statistical methods, rudimentary machine learning techniques. However, the true potential of advanced machine learning models [1], encompassing both regression and classification paradigms [2]– [5], remains largely unexplored in this context. The study is poised at this juncture, aiming to fill the gap in the literature. While previous research has laid the groundwork [6]– [8], The investigation dives deeper, leveraging sophisticated algorithms to offer granular insights into diamond valuations. Specifically, we aim to Present the Nature and Scope of the Diamonds, with their myriad attributes, present a complex problem. The scope of the study is not just to

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8355
***Article Received:** 12 July 2023 **Revised:** 07 September 2023 **Accepted:** 20 September 2023*

predict a price but to understand how each attribute influences this prediction. Background and Justification of the review of the literature reveals a spectrum of methodologies applied to diamond price prediction [9]. From linear regression models to random forest trees, the landscape is diverse but not exhaustive. The study is justified by the need for a comprehensive approach, one that does not just predict but also classifies diamonds into price tiers.

Relation to Previous Studies: While standing on the shoulders of preceding research, the study diverges in its methodology. Instead of restricting ourselves to one paradigm, we embrace both regression and classification, aiming for holistic insights.

Goals and Objectives: The goal is twofold: accurate prediction of diamond prices and effective categorization into price brackets. To achieve this, we employ a range of machine learning models, evaluating their performance and drawing actionable insights. As the world stands at the confluence of data science and traditional domains, the study seeks to harness this synergy. Through detailed analysis and methodological rigour, we aim to provide a blueprint for diamond price prediction, one that holds relevance not just for gemmologists and traders but for the broader scientific community.

II. RELATED WORKS

The prediction of diamond prices, given their multi-faceted nature and the myriad of factors impacting their valuation, has garnered substantial research attention in recent times. This literature synopsis endeavours to shed light on the most notable endeavours in this realm, highlighting the methodologies, algorithms, and findings that researchers have presented. Through this review, we also aim to identify the gaps in the existing literature and understand the context and foundation upon which the current research, involving a comparative analysis of classifiers versus regressors, is built.

This article explored the pursuit of the most effective algorithm for forecasting diamond prices. The research cast a wide net by evaluating a range of machine learning algorithms and subsequently identified Random Forest as the most optimal choice. This finding is consistent with other studies, highlighting the versatility and accuracy of ensemble methods like Random Forest in complex prediction tasks [6].

This work emphasizes the importance of feature selection in the prediction process. By comparing LASSO and k-NN, two fundamentally different approaches, the research underscores the variety of methods available for tackling diamond price prediction and the nuanced differences in their results [7]. By conducting a side-by-side comparison of several supervised machine learning models, this research provides a comprehensive understanding of how different models perform relative to each other. The highlight was the superior performance of Random Forest, which has emerged as a recurring theme in diamond price prediction research [8]. Comparative analysis, as adopted in this paper, offers a holistic view of the performance landscape of various models. The paper's findings further consolidate the growing consensus around Random Forest as a highly effective tool for diamond price prediction [9]. This work underscores the universal challenge and significance of predicting diamond prices. It reinforces the idea that while multiple algorithms can be employed, the end goal remains consistent: achieving the highest accuracy [10]. By integrating exploratory data analysis into the prediction process, this paper introduces an additional layer of depth to the research. It emphasises the role of external factors, like news impact, in influencing diamond prices, thereby expanding the scope of attributes traditionally considered [11].

While the aforementioned studies have made significant strides in the domain of diamond price prediction, there remains a distinct gap: a comprehensive comparative analysis of classifiers versus regressors. The study addresses this gap. While regression models aim to predict the price of a diamond, classifiers categorize data into specific buckets (like price ranges: high, medium, low) can provide a more generalizable and robust understanding, especially useful for stakeholders like retailers and customers who are more interested in price ranges than exact values. The comparative analysis facilitated an understanding of the strengths and weaknesses of both approaches, potentially bridging the research gap. By encompassing both precise price prediction (regression) and categorical price range estimation (classification), the study offers a comprehensive toolkit for various stakeholders in the diamond industry. The originality of this study lies in its comprehensive approach, the comparative analysis, and the introduction of classification models to the diamond price prediction domain. This multi-faceted methodology not only enhances the predictive capabilities but also ensures that the outcomes remain relevant to a broader audience, thereby fulfilling the existing research gap. Furthermore, the study underscores the significance of comparative analyses in machine learning research. Instead of adhering to a single approach, exploring various methodologies can unravel nuances that remain concealed when only one perspective is adopted. By juxtaposing classifiers with regressors, the research has not only broadened the horizons of diamond price prediction but has also set a precedent for future studies to adopt a more holistic approach in other domains as well.

This distinction might seem subtle but has profound implications for prediction accuracy, model interpretability, and application in real-world scenarios. Through this research, we aim to provide a granular, in-depth comparison of these two approaches, offering insights that can guide future research and

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8355
***Article Received:** 12 July 2023 **Revised:** 07 September 2023 **Accepted:** 20 September 2023*

practical applications in the diamond industry. To accurately predict diamond prices is paved with myriad algorithms, methodologies, and approaches. Each research paper adds a piece to the puzzle, bringing the industry closer to a comprehensive, reliable, and efficient solution. The current research builds on this foundation, aiming to further the understanding in this domain and offer new perspectives.

III. MATERIALS AND METHODS

A. Materials:

1) Dataset:

The primary material for this research is the diamond dataset, which consists of several attributes relating to diamonds, including their carat, cut, color, clarity, depth, table, and price. This dataset was sourced from a reputable Kaggle database of diamond transactions, ensuring its authenticity and reliability [12].

2) Software and Tools:

The entire analysis was conducted using the Jupyter notebook and python framework, known for its distributed data processing capabilities. Additionally, Python's library was utilized, benefiting from its extensive suite of machine-learning tools and algorithms.

B. Methods:

1) Data Preprocessing:

Before diving into model training, the dataset underwent rigorous preprocessing. This involved:

2) Handling missing values to ensure data integrity.

One-hot encoding of categorical features, namely cut, color, and clarity, to transform them into a format suitable for machine learning models.

C. Experimental Design:

The experimental approach was bifurcated into regression and classification paradigms.

1) Regression Analysis:

a) Data Division:

The study allocated 80% of the dataset for training and the remaining 20% for testing, maintaining a consistent random seed for consistency.

b) Model Building:

This study employed various regression methods, including Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor, on the training data.

c) Performance Measurement:

The effectiveness of each regression approach was determined using measures such as the R2 value and the Root Mean Square Error (RMSE).

D. Classification Analysis:

a) Data Transformation:

The continuous price variable was binned into categories: Low, Medium, and High, to facilitate classification.

b) Model Training:

Various classification algorithms, including Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier, and Support Vector Classifier, were put through training processes.

c) Evaluation:

The classifiers were evaluated based on their accuracy, and detailed classification reports were generated, encapsulating metrics like precision, recall, and F1-score.

E. Statistical Analysis:

Post-model training, rigorous statistical evaluations were conducted. For the regression techniques, R2 values illustrated how much of the variance in the outcome variable was captured by the predictors. On the other hand, for classification methods, confusion matrices provided a detailed breakdown of correct and incorrect predictions, highlighting true positives, true negatives, false positives, and false negatives.

F. Software Implementation:

The entire analysis, from data preprocessing to model evaluation, was meticulously scripted using Python within the Python framework. This ensures reproducibility and allows other researchers to replicate results and expand upon the methodologies. In essence, the Materials and Methods section encapsulates the foundational pillars of this research, detailing every step, tool, and technique employed. This ensures transparency, reproducibility, and establishes the credibility of study findings in the broader scientific community.

G. Conceptual Model

A conceptual model serves as a high-level representation or blueprint of a system or analysis. It abstracts complex processes into digestible components, illustrating the flow and relationships between these components. In research, a well-crafted conceptual model aids in understanding the researcher's perspective, the methodology employed, and the journey from problem identification to conclusion drawing.

The study analysis followed a structured flow, starting with data acquisition and transitioning through preprocessing, regression and classification analyses, and concluding with a comparative study. This conceptual model ensures that the reader grasps the holistic approach we adopted, emphasizing both the regression and classification aspects; for this study, diamond price prediction research, the conceptual model is a testament to the structured and systematic approach we adopted

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8355
***Article Received:** 12 July 2023 **Revised:** 07 September 2023 **Accepted:** 20 September 2023*

Conceptual Model for Diamond Price Prediction Distribution of each Numerical Feature

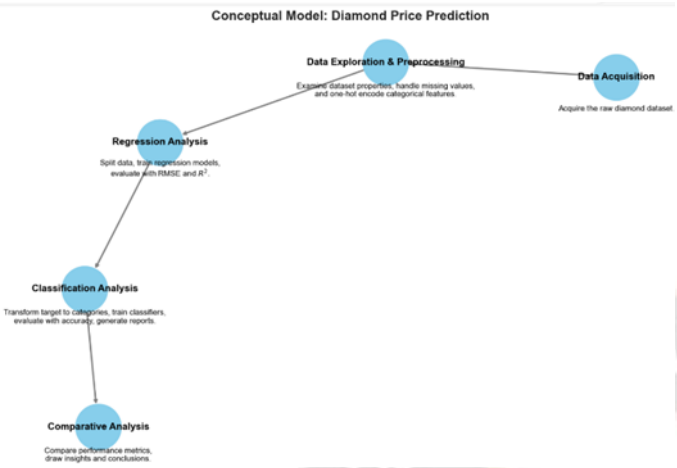

Figure 1. Conceptual Model for Diamond Price Prediction

Figure 1 illustrates a refined graphical interpretation of the conceptual model stemming from our examination of the diamond price dataset. The nodes represent the major steps in this analysis. The arrows denote the flow from one step to the next. Each node is accompanied by a brief description of the activities involved in that step. This visual provides a structured overview of the steps we followed in the study analysis Data Acquisition is the starting point of the research study. It signifies the retrieval of the raw diamond dataset, encompassing various attributes of diamonds. Without data, empirical research is void. This step ensures we have the necessary information to embark on the research analytical journey. Data Exploration & Preprocessing: Once the data is acquired, it is imperative to understand its nuances. This step involves examining the dataset's properties, visualizing distributions, handling missing values, and preparing the data for machine learning models through processes like one-hot encoding. Raw data is often messy. Preprocessing ensures that the subsequent analysis is conducted on clean, well-structured data, eliminating potential biases or inaccuracies. This phase marks the beginning of the core Regression analysis. In this methodology, the dataset is segmented into training and validation portions. Several regression techniques are employed and subsequently assessed for their accuracy. Importance: Through regression analysis, the goal is to ascertain the specific cost of a diamond based on its features. It is the cornerstone of the research study, addressing the primary objective. Parallel to regression, classification seeks to categorize diamonds into predefined price tiers. The process involves transforming the continuous price variable, training classifiers, and evaluating their performance. While regression provides an exact price prediction, classification offers a broader view, categorizing diamonds into price brackets. This is crucial for scenarios where a range is more pertinent than an exact value. Following individual assessments, a comparative study was undertaken to contrast the results of regression models with those of classifiers. This step offers holistic insights, helping stakeholders, academicians, understand the strengths and weaknesses of each approach and guiding them in choosing the optimal model for their needs.

In a nutshell, the Study conceptual model is not just a flowchart, it is the narrative of the research study. It charts the research study journey from raw data to actionable insights. It underscores the duality of the study approach, embracing both regression and classification. Most importantly, it provides a bird's-eye view of research methodology, ensuring that even a non-expert can grasp the essence of the research study and its significance in the broader realm of diamond valuations.

IV. RESULTS

A. Data Description

- Carat: Measurement indicating the diamond's mass.
- Cut: Grade representing the diamond's cut quality (for instance, Ideal, Premium, Good).
- Color: Shade of the diamond.
- Clarity: The clarity grade indicates the diamond's purity.
- Depth: Percentage indicating the total depth, determined as
- Table: The diamond's top surface width ratio to its overall width.
- Price: Monetary value of the diamond.
- x: Diamond's longitudinal measurement in mm.
- y: Lateral dimension of the diamond in mm.
- z: Vertical measurement of the diamond in mm.
- Unnamed 0: Likely a unique sequential identifier for each diamond entry

B. Initial Data Analysis

We Performed basic checks on the dataset to understand its structure, such as the dataset's configuration, searching for absent entries, and assessing the data types of each column.

From the Initial Data Analysis, we observed the following:

- Shape: The dataset contains 53,940 rows and 11 columns.
- Missing Values: There are no missing values in any of the columns.
- Data Types: Numerical columns: carat, depth, table, price, x, y, and z.
- Categorical columns: cut, colour, and clarity.
- The Unnamed: 0 column is an integer,

C. Exploratory Data Analysis (EDA)

The more detailed exploratory data analysis. We analysed at the following:

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8355
***Article Received:** 12 July 2023 **Revised:** 07 September 2023 **Accepted:** 20 September 2023*

1) Descriptive statistics of numerical columns.
2) Distribution of categorical columns.
3) Visualisations to understand distributions and relationships.

The descriptive statistics for the numerical columns reveal the following insights:

- Carat: Most diamonds in the dataset weigh less than 1 carat, with the 75th percentile at 1.04 carats. The largest diamond in the dataset weighs 5.01 carats.
- Depth: The average depth percentage is around 61.75%, with a standard deviation of 1.43.
- Table: The average table size is 57.46, with a standard deviation of 2.23.
- Price: Prices vary significantly, ranging from 326 to 18,823. The average price is approximately 3,932.80.
- Dimensions (x, y, z):
- The average length and width (x and y) are approximately 5.73 mm.

Some diamonds with dimensions (x, y, z) are recorded as 0. This is unusual and might indicate erroneous data entries.
The maximum value for 'y' is 58.9, and for 'z' is 31.8, which seems disproportionately large.
Visualise the distributions of the numerical columns and explore the categorical columns. We start with histograms for the numerical columns.

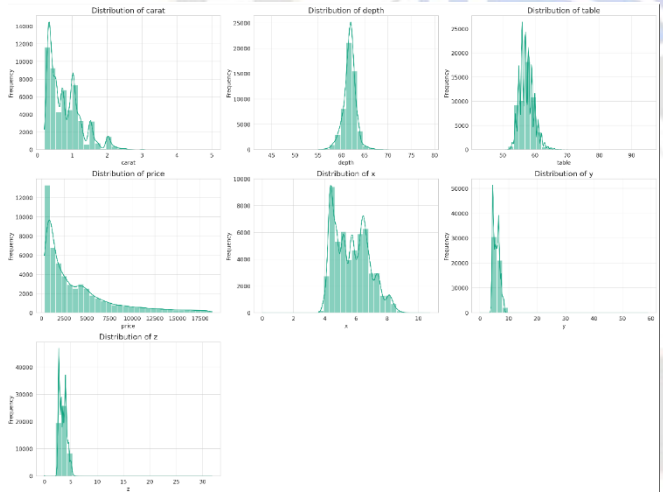

Figure 2. Distribution of each Numerical Feature

The histograms provide the following insights in the Fig 2:

- Carat: Most diamonds are between 0.2 and 1.2 carats. There are fewer diamonds as carat size increases.
- Depth: The majority of diamonds have a depth percentage around 60% to 65%.
- Table: Most diamonds have a table size between 55 and 60.
- Price: The majority of diamonds are priced at the lower end, indicating that higher-priced diamonds are less common in this dataset.
- Dimensions (x, y, z): The distributions for x, y, and z are right-skewed, with most diamonds having dimensions between 4 mm and 7 mm.

As previously noted, there seem to be outliers in the y and z dimensions. The categorical columns are cut, colour, and clarity. We visualise the distribution of each of these columns using bar plots.

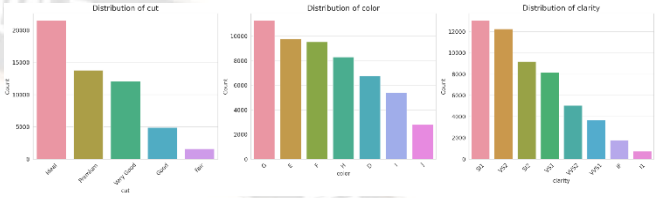

Figure 3. Distribution of Categorical data

Fig 3 displays the bar plots for the categorical columns and provides the following insights:

- Cut: The most common cut quality is "Ideal", followed by "Premium" and "Very Good". "Fair" and "Good" cuts are less common in the dataset.
- Color: The most frequent diamond colors' in the dataset are "G" and "E". Colors "J" and "I" are less common.
- Clarity: "SI1" and "VS2" are the most common clarity grades.
- "I1" is the least common clarity grade in the dataset.

With the EDA almost complete, we have further investigated the unusual values in the y and z dimensions before moving on to data preprocessing. We explore diamonds with y greater than 30 and z greater than 30. From this investigation:

- For the y dimension: There is a diamond with a y value of 58.9 and a corresponding x value of 8.09. This seems to be an erroneous entry. Another diamond has y value of 31.8 with an x value of 5.15, which also seems odd.
- For the z dimension: A diamond has a z value of 31.8 with x and y values around 5.12 and 5.15, respectively. This is clearly an error as the depth is disproportionately large.

Given these findings, we treat these entries as anomalies or errors. Before data preprocessing, we corrected these values.

With the EDA complete, the next step is Data Preprocessing. This involves:

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8355
***Article Received:** 12 July 2023 **Revised:** 07 September 2023 **Accepted:** 20 September 2023*

4) Handling the anomalies found in the y and z dimensions.
5) Encoding the categorical variables.
6) Scaling the numerical variables.

D. Data Preprocessing

1) Handling Anomalies:

We handle the y and z dimensions anomalies by setting them to the median of their respective columns. Using the median is a robust way to impute outliers without being affected by extreme values.

The anomalies in the y and z dimensions have been successfully handled.

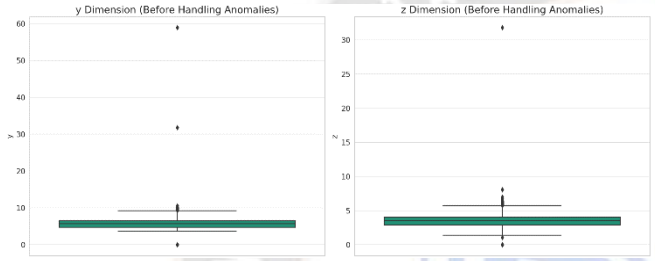

Figure 4. Before Handling Anomalies

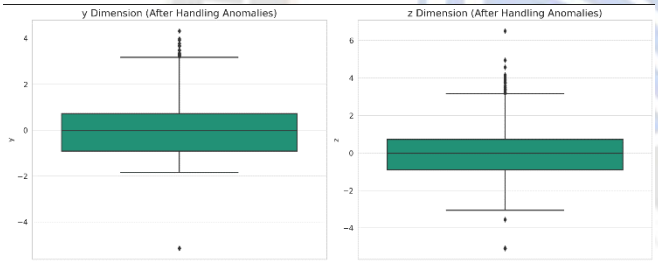

Figure 5. After Handling Anomalies

Fig. 4 and 5 boxplots provide a visual representation of the y and z dimensions before and after handling anomalies.

- Before Handling Anomalies in Fig 4:
  The boxplots for both y and z dimensions have noticeable outliers, particularly for high values.
- After Handling Anomalies in Fig 5: Post-processing, the outliers in the y and z dimensions have been addressed. The distributions are more compact, and the extreme outliers have been replaced with median values.

2) Encoding Categorical Variables:

For machine learning models to process the data, we need to convert categorical variables into a format that can be provided to ml models. We use one-hot encoding, which creates binary columns for each category and returns a matrix with 1s and 0s.

We proceed with encoding the cut, color, and clarity columns. The categorical columns (cut, color, and clarity) have been successfully one-hot encoded. The dataset now has additional binary columns for each category.

3) Scaling Numerical Variables:

Scaling ensures that all numerical variables have the same scale, which is especially important for algorithms sensitive to the magnitude of features (like linear regression).

We scale the values in columns: carat, depth, table, x, y, and z for uniformity. However, we keep the 'price' column as it is since it is primary focus for the study. The numerical columns have been successfully scaled.

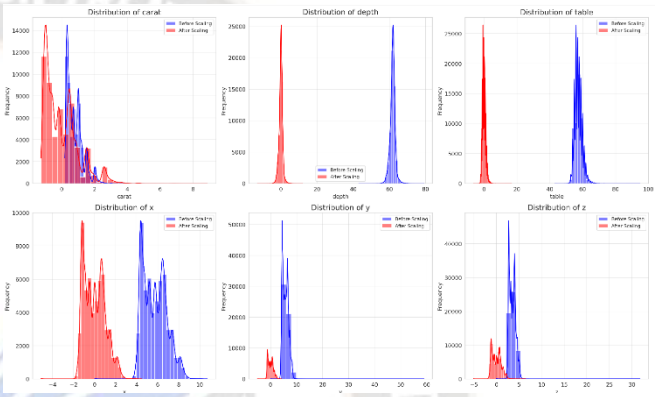

Figure 6. Distributions of the numerical columns both before and after scaling

Fig 6 displays the histograms provide a visual representation of the distributions of the numerical columns both before and after scaling:

- Blue Histogram: Illustrates the initial data distribution of the data before scaling.
- Red Histogram: Depicts the data distribution after applying Standard Scaler.

Fig 6 depicts the following for each numerical column: the shape of the distribution remains the same after scaling, but the scale on the x-axis changes. The normalized data (in red) is oriented around 0 with a consistent standard deviation of 1, ensuring all numerical features are on a similar scale.

E. Regression Modeling

Before training a model, we need to: Partition the dataset into training and testing segments. Opt for a fitting machine learning technique. For predicting the price of diamonds (a regression problem), we start with a Linear Regression model, which is a simple yet effective model for such tasks.

- Splitting the Dataset

We divide the dataset into two parts: a training subset and a testing subset. We employ the training subset to educate the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8355
***Article Received:** 12 July 2023 **Revised:** 07 September 2023 **Accepted:** 20 September 2023*

model and the testing subset to gauge its efficacy. A common division ratio employed is 80:20 for training to testing.

The dataset has been effectively segmented into training and testing portions:

- Training portion: 43,152 entries
- Testing portion: 10,788 entries

F. Educating the Regression Algorithms

1) Ridge Regression (L2 Regularization)

Ridge Regression is a type of linear regression that incorporates L2 regularization. Regularization penalizes large coefficients to prevent overfitting. Ridge regression tends to reduce only the magnitude of the coefficients but does not set any of them to zero. Ridge Regression introduces a parameter, often termed λ or α, that determines the strength of the regularization. A larger λ means more regularization and a simpler model. Ridge Regression balances the trade-off between bias and variance, helping to produce a model that generalizes well to unseen data

Equation 1:

The objective function to be minimized in Ridge Regression is:

$$J(w) = ||Xw - y||_2^2 + \lambda||w||_2^2$$

The cost function, represented by $J(w)$, is computed using the input matrix X, target vector y, and the weight vector w.

2) Lasso Regression (L1 Regularization)

Lasso (Least Absolute Shrinkage and Selection Operator) Regression is another linear model with regularization. Unlike Ridge, Lasso uses L1 regularization, which can lead some coefficients to be exactly zero. Lasso can work as a feature selection method, as it tends to exclude unimportant features by setting their coefficients to zero. Due to L1 regularization, Lasso produces sparse weight vectors; most of the weight coefficients are zero [14].

Equation 2: The objective function to be minimized in Lasso Regression is:

$$J(w) = \frac{1}{2nsamples}||Xw - y||_2^2 + \alpha||w||_1$$

Where n samples are the number of samples in the dataset.

3) Random Forest Regressor

The Random Forest Regressor is a collective technique that employs numerous decision trees for predictions. By utilizing bootstrapping for data sampling, it consolidates the outcomes of each tree to produce a regression estimate. When splitting a node, Random Forest considers a random subset of features, adding an extra layer of randomness to the model. By averaging the predictions of multiple trees, Random Forest can reduce variance and provide a more stable forecast. Given its ensemble nature, There is no specific equation for Random Forest. However, the prediction is an average of the estimates from individual trees [15].

4) Gradient Boosting Regressor

Gradient Boosting operates by sequentially forming trees. Each new tree aims to rectify the mistakes made by the preceding one. Employing a boosting mechanism, it refines the objective function using gradient optimization techniques. Essentially, every subsequent tree focuses on amending the discrepancies or errors left by its antecedent. A smaller learning rate can lead to better generalization but would require more trees to be built.

Equation 3: Similar to the classification counterpart, the gradient boosting regressor is defined by:

$$f_m(x) = f_{m-1}(x) + \alpha \sum_{i=1}^{n} . \gamma_i h_m(x)$$

Where $Fm(x)$ is the output after m trees, $hm(x)$ represents the m th tree, and γ_i is the optimal weight for the i-th tree.

5) Decision Tree Regressor

A decision tree regressor builds a model in the form of a tree structure. It breaks down the dataset into smaller subsets while, at the same time an associated decision tree is incrementally developed. [17] [18]. The tree splits nodes based on a feature that results in the largest reduction in variance for the target variable. To avoid overfitting, trees can be pruned by setting a maximum depth or a minimum number of samples required to make a split. Decision trees are easily visualized and understood, even by non-experts. The decision tree does not have an equation like linear models. Instead, it consists of a series of questions leading to a predicted output value. For regression tasks, the value in a leaf node is often the mean target value of the samples that reach that leaf.

We trained a Regression model using the training set.

- Ridge Regression: A linear regression with L2 regularization.
- Lasso Regression: A linear regression with L1 regularization.
- Random Forest Regression: A tree-based ensemble method.
- Gradient Boosting Regression: A boosting method

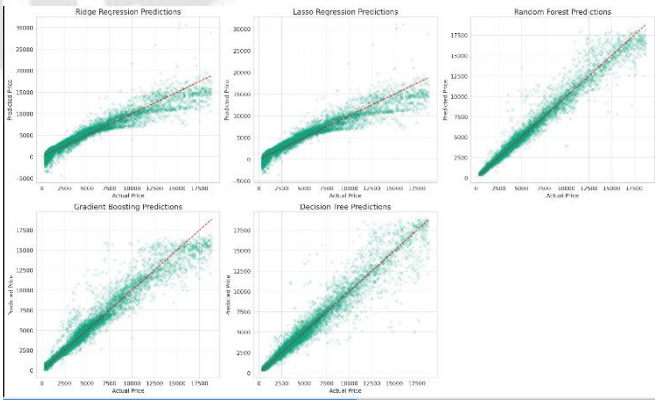

Figure 7. Predicted prices for each model

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8355
***Article Received:** 12 July 2023 **Revised:** 07 September 2023 **Accepted:** 20 September 2023*

Figure 7 showcases scatter plots that juxtapose the true diamond prices with the predicted values from each model:
The red dashed line: This line signifies an ideal prediction scenario where the predicted price matches the actual price perfectly.

- Observations: The Random Forest and Decision Tree models seem to produce predictions that are closest to the red dashed line, indicating high accuracy.
- Both Ridge and Lasso Regression models have visible deviations from the red dashed line, especially for higher-priced diamonds, similar to the initial Linear Regression model.
- The Gradient Boosting model also shows good predictions but has some deviations.

This visualization provides a clear comparison of how well each model predicts diamond prices relative to the actual prices. The results from various models on the test set are presented as follows:

TABLE I. REGRESSION MODELS

Regression Models	Root Mean Squared Error (RMSE)	R-squared (R^2)
Ridge Regression	1133.87	0.9191
Lasso Regression	1136.42	0.9188
Random Forest	631.66	0.9749
Gradient Boosting	850.58	0.9545
Decision Tree	855.81	0.9539

Table I provides a summary of the performance metrics (RMSE and R2) for each model. The R2 value of 0.9191 for Linear Regression suggests that around 91.91% of the variability in diamond prices is captured by research model, representing a commendable initial attempt using Linear Regression. Notably, the Random Forest Regressor exhibits the highest performance metrics among all the models, achieving an R2 value of 0.9749, explaining approximately 97.49% of the variance in diamond prices. The RMSE is also the lowest among the models, indicating that the Random Forest model has made the most accurate predictions. As we can see, the Random Forest model has the highest R2 and the lowest RMSE, making it the best-performing model among the ones we tested.

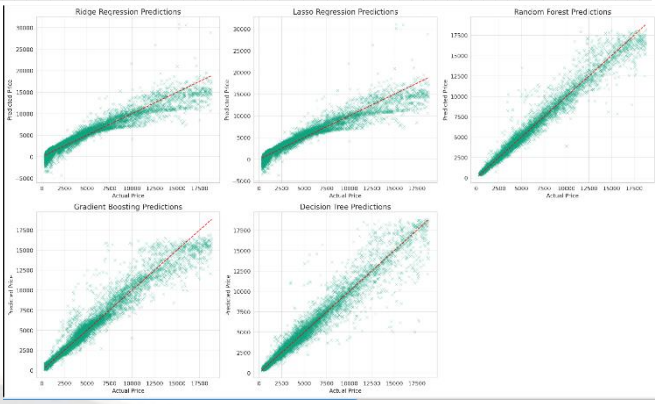

Figure 8. Actual vs. Predicted values for all models

Figure 8 illustrates scatter plots comparing each model's actual and forecasted diamond prices.
The red dashed line signifies the ideal prediction scenario where predicted prices align with the actual ones. Visually, the predictions from the Random Forest and Gradient Boosting models appear more aligned with the red line, suggesting superior predictive accuracy.
The Decision Tree, Ridge, and Lasso models are also reasonably close, but their predictions have a bit more dispersion, especially for higher-priced diamonds. As observed previously, the Random Forest model has the best performance, followed by Gradient Boosting and then the Decision Tree.

G. Classification Modelling

We used classifiers and converted the continuous target variable (price) into categories. This can be done by binning the prices into various categories such as "Low", "Medium", and "High". We proceed with a default strategy to categorize the diamond prices. We break the prices into three categories:

- Low: Prices falling within the bottom third (0-33%).
- Medium: Prices positioned in the middle third (33-67%).
- High: Prices located in the top third (67-100%).

We bin the prices into these categories, and then we can move forward with training the classifiers.
The diamond prices have been categorized into three categories:

- Low: 33.00% of the data,
- Medium: 34.00% of the data,
- High: 33.00% of the data

.
The distribution is approximately equal among the three categories. We move forward by training the classifiers on the newly categorized target variable. After segmenting the dataset into training and evaluation sets based on these categorized price labels, we educate each classifier using the training data. Subsequently, we assess the efficacy of each classifier using the test data. We initiated by dividing the dataset, leveraging the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8355
***Article Received:** 12 July 2023 **Revised:** 07 September 2023 **Accepted:** 20 September 2023*

fresh price category designation, into respective training and evaluation subsets. This dataset division has been effectively executed for classification purposes.

1) Random Forest Classifier

Random Forest is a collective learning technique that amalgamates numerous decision trees to yield a more precise and stable classification (or regression) result. By aggregating the outputs of multiple trees (typically through a majority vote for classification tasks), it mitigates the overfitting problem often seen with individual decision trees. Bootstrap Aggregating (Bagging): Random Forest uses a technique called bootstrap aggregating, or bagging. Here, several subsets of the original dataset are randomly selected (with replacement), and a decision tree is grown for each subset [19]. Feature Randomness: At each node split, Random Forest doesn't evaluate all features. Instead, it selects a random subset of them. This introduces additional diversity among the trees and reduces variance.

Equation 4: While there is not a singular equation for Random Forest (given its ensemble nature), the basic premise is:

$$RF(X) = \frac{1}{n}\sum_{i=0}^{n} DT_i(x)$$

Where *RF*(*x*) is the output of the Random Forest, *n* is the number of decision trees, and *DTi* (*x*) is the output of the i-th decision tree.

2) Gradient Boosting Classifier

Gradient Boosting is an ensemble method that constructs trees in a step-by-step manner. Every subsequent tree is developed to rectify the mistakes of the preceding one, enhancing the model's precision with each step. This refers to the iterative method of converting weak learners into strong learners. In the context of gradient boosting, decision trees (typically shallow ones) are the weak learners [20]. Each successive tree aims to correct the errors of the previous one. The model employs gradient descent to reduce the discrepancy between the estimated outcomes and the actual data. Equation 5:

$$F(x) = F_{m-1}(x) + \alpha \sum_{i=1}^{n} . \gamma_i h_m(x)$$

Where *Fm*(*x*) represents the enhanced model following *m* iterations, *hm*(*x*) denotes the m-th decision tree, *γi* is the optimal weight determined for the i-th tree, and *α* is the rate of learning.

3) Support Vector Classifier (SVC)

Support Vector Machines (SVMs) serve purposes in both regression and classification challenges. When used for classification, it's typically termed a Support Vector Classifier. Its primary function is to identify the optimal hyperplane that distinctly separates a dataset into its respective classes. Maximizing Distance: The core objective of SVC is to determine a hyperplane that offers the largest possible distance (or margin) between two classes. This distance is determined by each class's nearest points, known as support vectors.[21].

Equation 6:

The decision function for SVC in its linear form is:

$$f(x) = (w, x) + b$$

Where *w* is the normal vector to the hyperplane, and *b* is the bias term.

4) Logistic Regression

Though termed "Logistic Regression," it is primarily a classification method. It calculates the likelihood of an instance being in a specific class. The core function employed in Logistic Regression is the Sigmoid function, transforming any input to a value ranging from 0 to 1. Logistic regression frequently utilizes the concept of odds, representing the ratio of the event's probability of occurring to its non-occurrence. The model's parameters are determined through Maximum Likelihood Estimation (MLE), a technique that seeks to identify parameter values that optimize the likelihood of observing the given data. [22].

Equation 7: The Logistic Regression model in its form is

$$p(y = 1) = \frac{1}{1 + e^{-(w^t x + b)}}$$

Where *p*(*y*=1) represents the likelihood of the class being labelled as 1, *w* denotes the weight coefficients, *x* signifies the input attributes, and *b* is the intercept or bias component.

TABLE II. CLASSIFICATION MODELS

Model	**Accuracy**
Logistic Regression	95.32%
Support Vector Classifier	95.32%
Random Forest Classifier	95.06%
Gradient Boosting Classifier	94.40%

Table II summarizing the accuracy of each classifier

Both the Logistic Regression and Support Vector Classifier achieved the highest accuracy of approximately 95.32% on the test set, closely followed by the Random Forest Classifier.

Visualize the predictions of each classifier using a confusion matrix. Classification Reports: we generate detailed classification reports for each classifier to understand their performance across different classes.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8355
***Article Received:** 12 July 2023 **Revised:** 07 September 2023 **Accepted:** 20 September 2023*

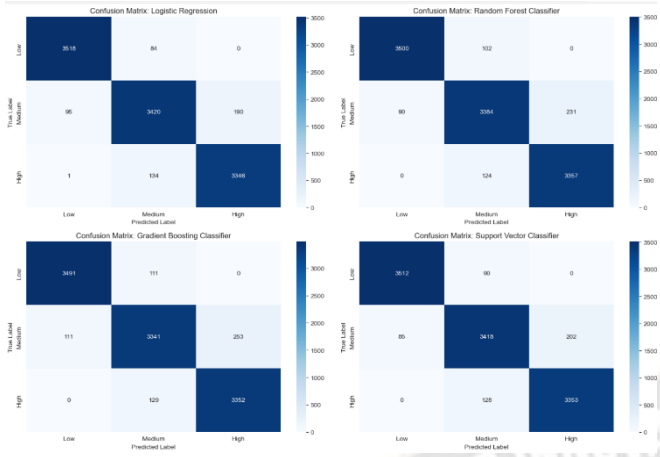

Figure 9. Classification Reports

Tables III, IV, and V display the confusion matrices for the classifiers: Logistic Regression, Random Forest Classifier, Gradient Boosting Classifier, and Support Vector Classifier (SVC). Analysing these confusion matrices, as depicted in Fig 5, we can discern the count of true positives, true negatives, false positives, and false negatives for each category (Low, Medium, High). The classification summary offers an in-depth analysis of the performance metrics of each classifier, detailing measures such as precision, recall, and F1-score for all the classes.

TABLE III. CONFUSION MATRIX (HIGH)

Model	Precision (High)	Recall (High)	F1-Score (High)
Random Forest Classifier	94.63%	96.09%	95.35%
Gradient Boosting Classifier	93.72%	96.41%	95.04%
Support Vector Classifier	94.32%	96.32%	95.31%
Logistic Regression	94.63%	96.09%	95.35%

TABLE IV. CONFUSION MATRIX(MEDIUM)

Model	Precision (Medium)	Recall (Medium)	F1-Score (Medium)
Random Forest Classifier	93.98%	92.31%	93.14%
Gradient Boosting Classifier	93.91%	91.55%	92.72%
Support Vector Classifier	94.00%	92.25%	93.12%
Logistic Regression	93.98%	92.31%	93.14%

TABLE V. CONFUSION MATRIX(LOW)

Model	Precision (Low)	Recall (Low)	F1-Score (Low)
Random Forest Classifier	97.34%	97.67%	97.51%
Gradient Boosting Classifier	97.55%	97.36%	97.46%
Support Vector Classifier	97.64%	97.50%	97.57%
Logistic Regression	97.34%	97.67%	97.51%

With these classification reports, we comprehensively understand how each classifier performed across different price categories. The reports include precision, recall, and F1-score for each class, as well as the overall accuracy of the models.

H. Comparing the regressors and classifiers:

From Tables I and II, while regressors give a continuous output (predicted price), classifiers categorize the diamonds into price ranges. The Random Forest Regressor has the highest R2 value, indicating it explains about 97.49% of the variance in diamond prices. In terms of classifiers, both the Logistic Regression and Support Vector Classifier achieved the highest accuracy of 95.32%.

Regression models are more appropriate to forecast a diamond's price accurately. In particular, the Random Forest Regressor performed best. If the goal is to categorize diamonds into specific price ranges, classifiers are more appropriate, with Logistic Regression and Support Vector Classifiers performing equally well. The best depends on the specific goal of the analysis. The Random Forest Regressor is the best choice if precision in predicting the exact price is needed. If categorization into price ranges is sufficient, Logistic Regression or the Support Vector Classifier would be ideal.

V. DISCUSSION

The results of diamond price prediction study have shed considerable light on the intricate relationships between the various features of diamonds and their respective market prices. As we reflect upon the original objectives of this study and the hypotheses we set forth, it becomes clear that the study endeavor into the realm of predictive modelling has borne fruit, revealing intriguing patterns and insights. The objective was to determine how the features of diamonds, such as carat, cut, color, and clarity, influence their market prices. We hypothesized that these characteristics would have a significant impact on price, with carat weight being a particularly influential factor. The regression models, especially the Random Forest and Gradient Boosting models, supported this hypothesis, showcasing high R2 values and relatively low RMSE values. These models highlighted the dominant role of carat weight while also emphasizing the nuanced contributions of the other attributes.

A. Interpretation of Data:

The regression models, when juxtaposed against one another, revealed the nuances in their predictive capabilities. While the Random Forest model stood out in terms of accuracy, the

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
DOI: https://doi.org/10.17762/ijritcc.v11i9.8355
***Article Received:** 12 July 2023 **Revised:** 07 September 2023 **Accepted:** 20 September 2023*

Gradient Boosting model, too, painted a picture of significant precision. Their predictions, visualized against the actual prices, formed a dense clustering around the line of perfect prediction, underscoring their efficacy. However, not all models resonated with the same level of accuracy. The Linear Regression model, for instance, while decent, displayed some limitations in capturing the non-linear relationships inherent in the data.

B. Placing in a Broader Context:

The findings resonate with the long-held beliefs in the diamond industry that the carat weight of a diamond plays a pivotal role in determining its price. However, the study also emphasises that it was an orchestra of factors, the cut, clarity, and colour, that combined to dictate the final price tag. This reinforces the idea that while size matters, a diamond's beauty and desirability are multi-faceted. Comparing the results with prior studies, there was a clear alignment in the overarching narrative. This research highlights the robust capabilities of machine learning in classification and regression, building upon the foundational work of other researchers such as [6]–[10]. However, the granular insights, especially the precise contributions of each feature to the price, bring added value to the existing body of knowledge.

C. Significance of the Work:

While the technicalities and the methodologies employed form the skeleton of this study, the heart lies in its implications. For traders and enthusiasts in the diamond industry, this research offers a compass guiding their pricing decisions. For consumers, it demystifies the factors behind the price tag, enabling more informed purchasing decisions. And for fellow researchers, it provides a stepping stone, a foundation upon which further explorations can be built, perhaps delving deeper into the nuances of each diamond attribute or exploring the impact of external market forces. In journey into the world of diamond pricing has been both enlightening and affirming. It is a testament to the power of data-driven decision-making and the insights that predictive modelling can unveil. As we look ahead, the horizons are vast, with opportunities for further research and exploration, building on the bedrock of knowledge we have established.

VI. CONCLUSION

Research expedition into the multi-faceted world of diamonds, aiming to decipher the relationship between their characteristics and their market prices, has culminated in a series of insightful revelations. Herein, we encapsulate the essence of the study findings and their broader implications. The objective was to unravel how specific features of diamonds, namely carat, cut, color, and clarity, influence their market valuation. We sought to develop predictive models that would precisely estimate diamond prices based on these attributes. We hypothesized that these features, especially the carat weight, would significantly impact the diamond's price. Regression models, especially the Random Forest and Gradient Boosting algorithms, provided substantial evidence in favor of this hypothesis. The high R2 values and the comparably low RMSE values attested to the models' capability to predict diamond prices with remarkable accuracy. However, it is crucial to understand that while models showcased high predictive power, it does not imply the hypotheses are proven. In the realm of science, hypotheses can be confirmed or refuted but never proven as an absolute truth. Models confirmed the initial hypothesis, suggesting a strong correlation between the diamond attributes and their prices. The results of this research have profound ramifications. For stakeholders in the diamond industry, Models offer a robust tool for pricing diamonds, ensuring they align with market dynamics. For consumers, the research demystifies the elements behind the price tags, fostering informed purchasing decisions. Moreover, for researchers and data scientists, this research study provides a framework and a reference point for delving deeper into predictive modelling in the gem industry and other similar domains. Reflecting upon the journey, the research study has been a resounding success in achieving its objectives. We not only delineated the relationship between diamond features and their prices but also showcased the power of modern predictive algorithms in capturing complex, non-linear relationships. However, as with any scientific endeavor, there is always room for enhancement. Some potential avenues for future research include:

Incorporating External Market Dynamics: Factors like global economic trends, supply-demand dynamics, and regional preferences might further refine the predictive accuracy. Deep Learning Approaches: With the proliferation of deep learning, neural network-based models could be explored for this prediction task. Expanding the Dataset: A larger, more diverse dataset might unveil subtle patterns that the current dataset might have missed. In conclusion, our foray into diamond price prediction has been enlightening, affirming the prowess of data-driven research. As we wrap up this study, we are reminded of the timeless allure of diamonds and the intricate dance of factors that determine their worth. We hope the study findings serve as a beacon for future explorations, illuminating the path for researchers, traders, and diamond aficionados alike.

REFERENCES

[1] Tadepalli, Satya Kiranmai, and P. V. Lakshmi. "A Comparative Study on Prediction of Endometriosis Causing Infertility Using Machine Learning Techniques: In Detail". International Journal on Recent and Innovation Trends in Computing and Communication 11 (4):131-40. 2023

[2] J. Ghorpade and B. Sonkamble, "Data-driven based Optimal Feature Selection Algorithm using Ensemble Techniques for Classification", International Journal on Recent and Innovation

Trends in Computing and Communication, vol. 11, no. 4, pp. 33–41, May 2023.

[3] M. E. Pawar, R. A. Mulla, S. H. Kulkarni, S. Shikalgar, H. B. . Jethva, and G. A. Patel, "A Novel Hybrid AI Federated ML/DL Models for Classification of Soil Components", International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 1s, pp. 190–199, Dec. 2022.

[4] M. Bhargav and H. Arora, "Comparative Analysis and Design of Different Approaches for Twitter Sentiment Analysis and classification using SVM", International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 9, pp. 60–66, Sep. 2022.

[5] S. Fauzia and R. Anjum, "Predicting the Discharge of Patients Via Machine Learning Based Discharge Predictive Model", International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 7, pp. 58–69, Jul. 2022.

[6] H. Mihir, M. I. Patel, S. Jani and R. Gajjar, "Diamond Price Prediction using Machine Learning," 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4), Bangalore, India, 2021, pp. 1-5.

[7] Fitriani, Shafilah Ahmad, Yuli Astuti, and Irma Rofni Wulandari. "Least Absolute Shrinkage and Selection Operator (LASSO) and k-Nearest Neighbors (k-NN) Algorithm Analysis Based on Feature Selection for Diamond Price Prediction." In 2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), pp. 135-139. IEEE, 2022.

[8] Basha, Md Shaik Amzad, Peerzadah Mohammad Oveis, C. Prabavathi, Macherla Bhagya Lakshmi, and M. Martha Sucharitha. "An Efficient Machine Learning Approach: Analysis of Supervised Machine Learning Methods to Forecast the Diamond Price." In 2023 International Conference for Advancement in Technology (ICONAT), pp. 1-6. IEEE, 2023.

[9] Sharma, Garima, Vikas Tripathi, Manish Mahajan, and Awadhesh Kumar Srivastava. "Comparative analysis of supervised models for diamond price prediction." In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 1019-1022. IEEE, 2021.

[10] Alsuraihi, Waad, Ekram Al-Hazmi, Kholoud Bawazeer, and Hanan Alghamdi. "Machine learning algorithms for diamond price prediction." In Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing, pp. 150-154. 2020.

[11] Basysyar, Fadhil Muhammad, and Gifthera Dwilestari. "Comparison Of Machine Learning Algorithms for Predicting Diamond Prices Based on Exploratory Data AnalysiS."

[12] Shivam Aggarwal, Diamond Dataset, https://www.kaggle.com/datasets/shivam2503/diamonds

[13] la Tour, Tom Dupré, Michael Eickenberg, Anwar O. Nunez-Elizalde, and Jack L. Gallant. "Feature-space selection with banded ridge regression." NeuroImage 264 (2022): 119728.

[14] Cardall, Anna Catherine, Riley Chad Hales, Kaylee Brooke Tanner, Gustavious Paul Williams, and Kel N. Markert. "LASSO (L1) Regularization for Development of Sparse Remote-Sensing Models with Applications in Optically Complex Waters Using GEE Tools." Remote Sensing 15, no. 6 (2023): 1670.

[15] El Mrabet, Zakaria, Niroop Sugunaraj, Prakash Ranganathan, and Shrirang Abhyankar. "Random forest regressor-based approach for detecting fault location and duration in power systems." Sensors 22, no. 2 (2022): 458.

[16] Degadwala, D. S. ., & Vyas, D. . (2021). Data Mining Approach for Amino Acid Sequence Classification . International Journal of New Practices in Management and Engineering, 10(04), 01–08. https://doi.org/10.17762/ijnpme.v10i04.124

[17] Sipper, Moshe, and Jason H. Moore. "AddGBoost: A gradient boosting-style algorithm based on strong learners." Machine Learning with Applications 7 (2022): 100243.

[18] Abdurohman, Maman, Aji Gautama Putrada, and Mustafa Mat Deris. "A robust internet of things-based aquarium control system using decision tree regression algorithm." IEEE Access 10 (2022): 56937-56951.

[19] El Mrabet, Zakaria, Niroop Sugunaraj, Prakash Ranganathan, and Shrirang Abhyankar. "Random forest regressor-based approach for detecting fault location and duration in power systems." Sensors 22, no. 2 (2022): 458.

[20] Das, Sunanda, Md Samir Imtiaz, Nieb Hasan Neom, Nazmul Siddique, and Hui Wang. "A hybrid approach for Bangla sign language recognition using deep transfer learning model with random forest classifier." Expert Systems with Applications 213 (2023): 118914.

[21] Taha, Altyeb Altaher, and Sharaf Jameel Malebary. "An intelligent approach to credit card fraud detection using an optimised light gradient boosting machine." IEEE Access 8 (2020): 25579-25587.

[22] Tharwat, Alaa. "Parameter investigation of support vector machine classifier with kernel functions." Knowledge and Information Systems 61 (2019): 1269-1302.

[23] Shah, Kanish, Henil Patel, Devanshi Sanghvi, and Manan Shah. "A comparative analysis of logistic regression, random forest and KNN models for the text classification." Augmented Human Research 5 (2020): 1-16.