**CLASSIFICATION: LUMPY SKIN DISEASE DETECTION**

**REGRESSION: DIAMOND PRICE PREDICTION**

*Machine Learning Group Project Report Submitted by*

**MOUNICA KAUMUDHI B 208W1A1273**

**LIKHITHA P 208W1A12B2**

**SRI SASHANK P 208W1A12B3**

****

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**V R SIDDHARTHA ENGINEERING COLLEGE**

**(AUTONOMOUS - AFFILIATED TO JNTU-K, KAKINADA)**

**Approved by AICTE &Accreted by NBA**

**KANURU, VIJAYAWADA-7**

**ACADEMIC YEAR**

**(2022-23)**

**1.fPROBLEM STATEMENT**

**CLASSIFICATION: LUMPY SKIN DISEASE DETECTION**

Skin types of diseases are most common among the globe, as people get skin disease due to inheritance, environmental factors. In many cases people ignore the impact of skin disease at the early stage. In the existing system, the skin disease is identified using biopsy process which is analyzed and medicinal prescribed manually by the physicians. To overcome this manual inspection and provide promising results in short period of time.

To address this issue, this process aims to develop an automated system that can accurately detect lumpy skin disease in individuals without the need for manual inspection. This system should be able to provide accurate results in a short period of time, making it easier for healthcare professionals to provide timely and effective treatment to patients.

**2. SCOPE OF THE PROJECT**

The scope of a classification-based lumpy skin disease detection project involves collecting a dataset of images or other data related to the disease. The data is preprocessed to handle missing values, outliers, and other anomalies, and prepared for analysis by transforming it into a format suitable for classification. An exploratory data analysis is conducted to identify patterns, trends, and relationships between the features and the target variable. Based on this analysis, the most relevant features are selected for classifying lumpy skin disease. A suitable classification model is then chosen, such as logistic regression, decision trees, or support vector machines, and trained on a portion of the data. The model's performance is evaluated on a separate portion of the data using metrics such as accuracy, precision, recall, and F1 score. The model is fine-tuned by adjusting the hyperparameters to improve its performance and generalization ability. Finally, the deployed model is used to detect lumpy skin disease, either as a standalone application or integrated into an existing platform. The project's overall goal is to develop an accurate and reliable classification model that can detect lumpy skin disease, with the aim of assisting veterinarians and farmers in controlling the spread of the disease.

**3. ARCHITECTURE/METHODOLOGY/ALGORITHM**

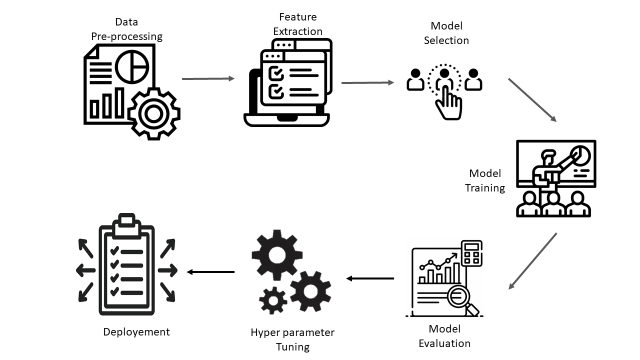


Fig 1 : Architecture Diagram

Data Input: This component represents the initial stage of your model where you will import your CSV file dataset into the system. The dataset will include features related to Lumpy Skin Disease such as age, breed, sex, clinical signs, and so on.

Data Preprocessing: This component involves cleaning and processing the raw data to prepare it for analysis. It may include tasks such as data cleaning, data transformation, handling missing data, and feature engineering.

Model Building: This component represents the main stage of your model where you will use three different algorithms, namely random forest, KNN, and decision tree classifier, to build predictive models for Lumpy Skin Disease. Each algorithm will be trained on a subset of the preprocessed data.

Model Evaluation: This component involves evaluating the performance of each algorithm in predicting Lumpy Skin Disease using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. Based on the evaluation results, you will select the best algorithm that yields the highest performance.

Output: This component represents the final stage of your model where you will generate an output in the form of a prediction of Lumpy Skin Disease based on the input data. The output will be generated using the selected algorithm with the highest performance.

**4. DATASET DESCRIPTION**

Assessing machine learning techniques in forecasting Lumpy Skin Disease occurrence based on meteorological and geospatial features.

These are the table attributes cloud cover (cld), diurnal temperature range (dtr), frost day frequency (frs), precipitation (pre), daily mean temperature (tmp), monthly average daily maximum (tmx) and minimum (tmn) temperature, vapour pressure (vap), Potential Evapo-transpiration (pet) and wet day frequency (wet),

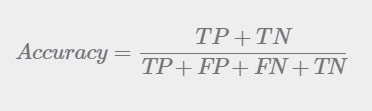
* cld: Cloud cover, which is the fraction of the sky covered by clouds.
* dtr: Diurnal temperature range, which is the difference between the maximum and minimum temperatures in a day.
* frs: Wind speed, which is the speed at which air is moving horizontally.
* pet: Potential evapotranspiration, which is the amount of water that could be evaporated and transpired by vegetation under ideal conditions.
* pre: Precipitation, which is the amount of rainfall or snowfall in a given period of time.
* tmn: Minimum temperature, which is the lowest temperature recorded in a day or other period.
* tmp: Average temperature, which is the mean temperature over a given period of time.
* vap: Vapor pressure, which is the pressure exerted by water vapor in the air.
* wet: Humidity or wet bulb temperature, which is the measure of how much moisture is in the air.
* elevation: Height above sea level, which is the vertical distance from a location to the mean sea level.

This disease is estimated based on the weather conditions in the specific region and country.This results the lumpy disease.

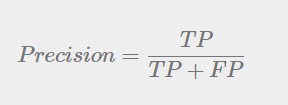
**5. EVALUATION MEASURES**

Accuracy, precision, recall, and F1-score are performance metrics commonly used to evaluate the performance of a classification model.

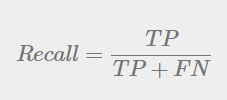
* Accuracy: The accuracy measures the proportion of correctly classified instances among the total instances in the dataset. It is calculated by dividing the number of correct predictions by the total number of predictions.



* Precision: Precision measures the proportion of true positive instances (correctly identified instances) among all instances predicted to be positive. In other words, precision indicates how well the model predicts the positive instances. Precision is calculated as the ratio of true positives to the sum of true positives and false positives.



* Recall: Recall measures the proportion of true positive instances among all actual positive instances in the dataset. In other words, recall indicates how well the model captures all positive instances in the dataset. Recall is calculated as the ratio of true positives to the sum of true positives and false negatives.



* F1-score: F1-score is the harmonic mean of precision and recall. It is a single score that combines the precision and recall metrics to provide an overall measure of the model's performance. The F1-score takes into account both precision and recall and is thus a good measure to evaluate the overall performance of the classification model.

**F1-score = 2 \* (precision \* recall) / (precision + recall).**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Random Forest** | **K-nearest neighbors** | **Decision Tree** |
| **Accuracy** | 1.0 | 0.986 | 0.986 |
| **Precision** | 1.0 | 0.987 | 0.987 |
| **Recall** | 1.0 | 0.905 | 0.905 |
| **F1 Score** | 1.0 | 0.945 | 0.945 |

Table 1 : Performance Measure

From above table, we can conclude that random forest algorithm have 100% accuracy results. When compared to KNN, Random Forest may perform better on large and complex datasets, as it can handle noise and reduce the risk of overfitting. In contrast, KNN may work better for small datasets with simple decision boundaries. When compared to Decision Tree, Random Forest can reduce the variance and bias of the model, which can result in higher accuracy and better generalization on unseen data. From the results we take random forest algorithm for disease detection.

**6. EXPERIMENTAL RESULTS**

The ROC curve is a graphical representation of the performance of a binary classifier, such as a random forest model, that shows the trade-off between sensitivity (true positive rate) and specificity (true negative rate) at different classification thresholds. (FPR, TPR): Any point on the ROC curve corresponds to a specific pair of FPR and TPR values, which can be used to evaluate the performance of the classifier at a particular threshold. The performance of the graph described as 1.0 and a straight line is shown.

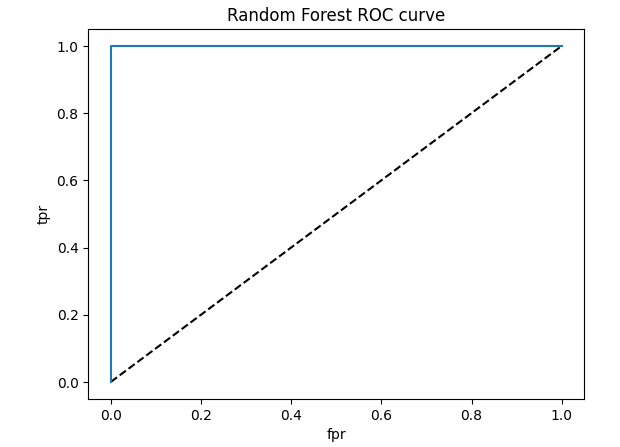


Fig. ‑ 2: Random Forest ROC Graph

The ROC curve is a graphical representation of the performance of a binary classifier, such as a random K-Nearest Neighbor(KNN), that shows the trade-off between sensitivity (true positive rate) and specificity (true negative rate) at different classification thresholds. (FPR, TPR): Any point on the ROC curve corresponds to a specific pair of FPR and TPR values, which can be used to evaluate the performance of the classifier at a particular threshold. The performance of the graph described and a slight variant (1.0,0.0) notable point

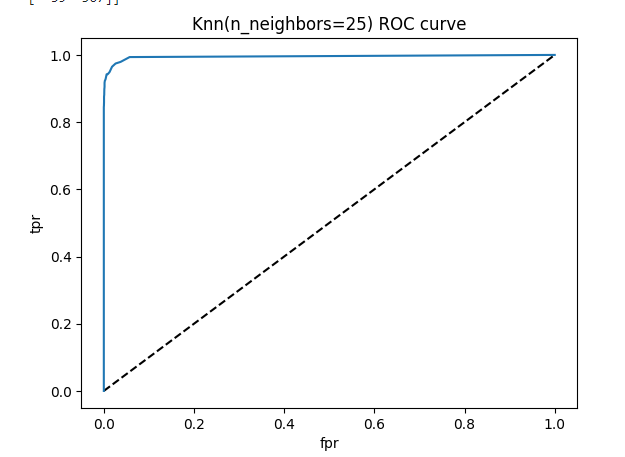
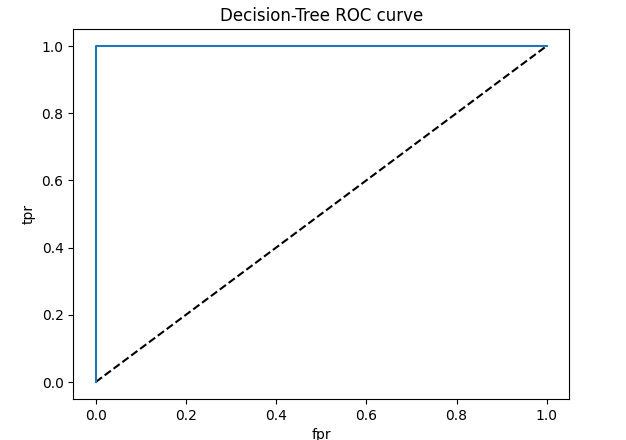


FIG. ‑3: KNN ROC GRAPH

The ROC curve is a graphical representation of the performance of a binary classifier, such as a random Decision Tree Classifier, that shows the trade-off between sensitivity (true positive rate) and specificity (true negative rate) at different classification thresholds. (FPR, TPR): Any point on the ROC curve corresponds to a specific pair of FPR and TPR values, which can be used to evaluate the performance of the classifier at a particular threshold. The performance of the graph described



FIG‑4: Decision Tree Classifier ROC Graph

ACCURACY PERFORMACE OF ALGORITHMS :

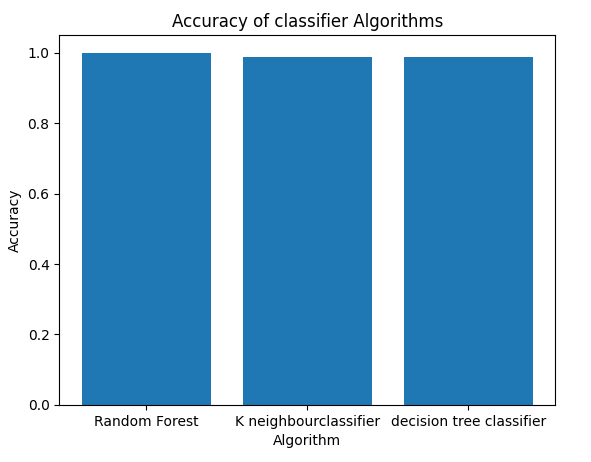


Fig 5 : Accuracy Graph

The accuracy performance of the algorithms : Random forest , KNN, Decision Tree Classifier are Respectively as 100%,98%,97% as shown in the above graph .

**CONCLUSION :**

It appears that the random forest algorithm performed the best out of the three algorithms tested, achieving 100% accuracy in predicting Lumpy Skin Disease. KNN and decision tree classifier also performed well, with accuracies of 98% and 97%, respectively. However the other metric such as recall , precision, f1 score, similarly shows the high percentage for the Random forest when compared to the KNN, Decision Tree Classifier . based on this we can conclude that the Random Forest model can be deployed into production to find the Lumpy Skin Disease.

**1.PROBLEM STATEMENT :**

**REGRESSION : DIAMOND PRICE PREDICTION**

The data is scrapped from Australian Diamond Importers website on 24th Feb 2022. It includes diamond features, like shape, size, colour, cut, clarity, and other features and the price in US dollars (excluding GST).The data is a good candidate for regression models to predict the price based on diamond features.

**2. SCOPE OF THE PROJECT**

The scope of a regression-based diamond price prediction project involves collecting a dataset of diamond prices and their associated characteristics, such as carat weight, cut, color, and clarity. The data is then preprocessed to handle missing values, outliers, and categorical variables, and prepared for analysis by transforming it into numerical format and scaling the features. Next, an exploratory data analysis is conducted to identify patterns, trends, and relationships between the features and the target variable. Based on this analysis, the most relevant features are selected for predicting diamond prices. A suitable regression model is then chosen, such as linear regression or polynomial regression, and trained on a portion of the data. The model's performance is evaluated on a separate portion of the data using metrics such as mean squared error, root mean squared error, or R-squared. The model is fine-tuned by adjusting the hyperparameters to improve its performance and generalization ability. Finally, the deployed model is used to predict diamond prices, either as a standalone application or integrated into an existing platform. The project's overall goal is to develop an accurate and reliable regression model that can predict diamond prices based on their characteristics, with the aim of assisting diamond buyers and sellers in making informed decisions.

**3. ARCHITECTURE/METHODOLOGY/ALGORITHM**

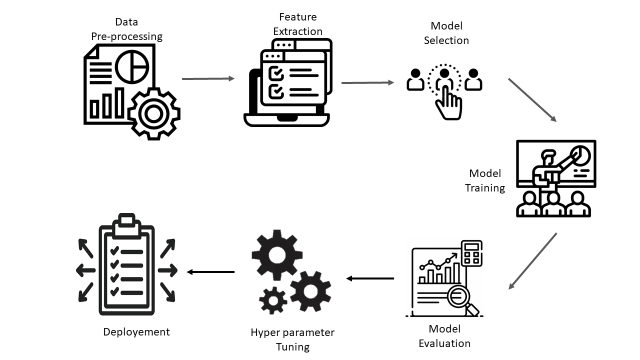


Fig 7 : Architecture Diagram

Data Input: This component represents the initial stage of your model where you will import your CSV file dataset into the system. The dataset will include features related to diamonds such as carat weight, cut, color, clarity, and so on.

Data Preprocessing: This component involves cleaning and processing the raw data to prepare it for analysis. It may include tasks such as data cleaning, data transformation, handling missing data, and feature engineering.

Model Building: This component represents the main stage of your model where you will use two different algorithms, namely random forest and linear regression, to build predictive models for diamond price prediction. Each algorithm will be trained on a subset of the preprocessed data.

Model Evaluation: This component involves evaluating the performance of each algorithm in predicting diamond prices using appropriate evaluation metrics such as mean squared error, mean absolute error, and R-squared. Based on the evaluation results, you will select the best algorithm that yields the highest performance.

Output: This component represents the final stage of your model where you will generate an output in the form of a predicted diamond price based on the input data. The output will be generated using the selected algorithm with the highest performance.

**4. DATASET DESCRIPTION**

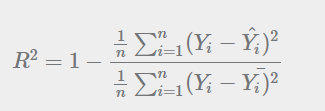
This classic dataset contains the prices and other attributes of almost 54,000 diamonds. There are 10 attributes included in the dataset including the target ie. price.

* **carat (0.2-5.01):** The carat is the diamond's physical weight measured in metric carats. One carat equals 0.20 gram and is subdivided into 100 points.
* **cut (Fair, Good, Very Good, Premium, Ideal):** The quality of the cut. The more precise the diamond is cut, the more captivating the diamond is to the eye thus of high grade.
* **color (from J (worst) to D (best):** The colour of gem-quality diamonds occurs in many hues. In the range from colourless to light yellow or light brown. Colouriess diamonds are the rarest. Other natural colours (blue, red,pink for example) are known as “fancy,” and their colour grading is different than from white colorless diamonds.
* **clarity (11 (worst), S12, S11, VS2, VS1, VVS2, VS, IF (best):** Diamonds can have internal characteristics known as inclusions or external characteristics known as blemishes. Diamonds without inclusions or blemishes are rare;however, most characteristics can only be seen with magnification.
* **depth (43-79):** Itis the total depth percentage which equals to z / mean(x, y) = 2 \* 2 / (x + y). The depth of themdiamond is its height (in millimetres) measured from the culet (bottom tip) to the table (flat, top surface) as referred in the labelled diagram above.
* **table (43-95):** Itis the width of the top of the diamond relative to widest point. It gives diamond stunning fire and brilliance by reflecting lights to all directions which when seen by an observer, seems lustrous.
* **price ($5326 - $18826):** Itis the price of the diamond in US dollars. Itis our very target column in the dataset.
* **x(0-10.74):** Length of the diamond (in mm).
* **y(0- 58.9):** Width of the diamond (in mm).
* **2(0-318):** depth of the diamond (in mm).

**5. EVALUATION MEASURES**

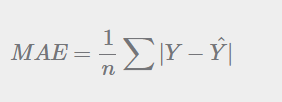
R-squared, MAE (Mean Absolute Error), and MSE (Mean Squared Error) are metrics used to evaluate the performance of regression models.

* R-squared : It is a statistical measure that represents the proportion of the variance in the dependent variable (y) that is explained by the independent variable(s) (x) in the regression model. R-squared can be calculated as the ratio of the explained variance to the total variance.



In the above equation, numerator is MSE and the denominator is the variance in 𝑌 values. We can use r2\_score function of sklearn.metrics to compute R squared value.

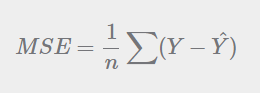
* MAE (Mean Absolute Error) :It is a measure of the average absolute difference between the actual values and the predicted values. MAE is calculated by taking the absolute difference between the actual and predicted values for each data point and then averaging the differences.



Here, 𝑌=Actual Output Values

Y^= Predicted Output Values

* MSE (Mean Squared Error): It is a measure of the average squared difference between the actual and predicted values. MSE is calculated by taking the square of the difference between the actual and predicted values for each data point and then averaging the squared differences.



𝑌=Actual Output Values

Y^ = Predicted Output Values.

|  |  |  |
| --- | --- | --- |
|  | **Linear Regression** | **Random Forest** |
| **Accuracy** | 0.9077 | 0.98164 |
| **R-Squared** | 0.9077 | 0.90770 |
| **MAE** | 807.426 | 807.4266 |
| **MSE** | 1477460.3 | 1477460.3 |

Table 2 : Performance Measure

On comparison analysis, we can conclude that random regression forest have high accuracy when compared to linear regression, as it shows 98% accuracy results. So the random forest regression can be used in production.

**CORRELATION GRAPH :**

Correlation coefficients: Each cell in the correlation graph will show a number, called the correlation coefficient, which represents the strength and direction of the linear relationship between two variables. Correlation coefficients range from -1 to 1, where -1 represents a perfect negative correlation (as one variable increases, the other decreases), 1 represents a perfect positive correlation (as one variable increases, the other also increases), and 0 represents no correlation. Positive correlations are shown in shades of green, while negative correlations are shown in shades of red.

Strong correlations: Look for cells in the correlation graph that are a darker shade of green or red, which indicate a stronger correlation between two variables. These variables may be good candidates to include in your model, as they may have a strong impact on diamond prices.

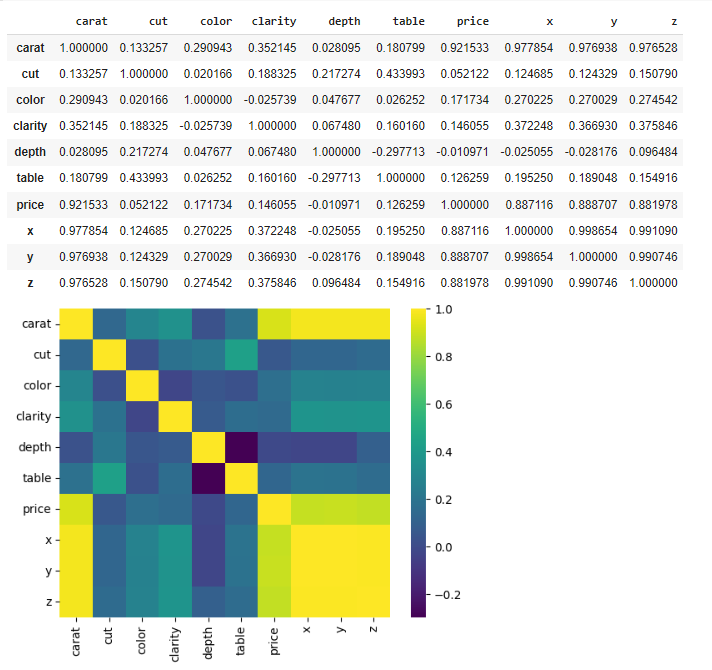


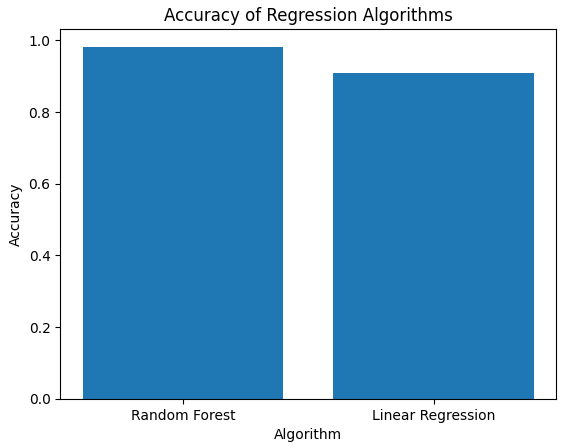
Fig 8 Correlation Matrix

Multicollinearity: If you see strong correlations between multiple features in the correlation graph, this may indicate multicollinearity, which is when two or more features are highly correlated with each other. Multicollinearity can cause issues in regression models, as it can make it difficult to determine the unique impact of each feature on the target variable. If you identify multicollinearity in your correlation graph, you may want to consider removing one of the highly correlated features from your model.

Outliers: If you see cells in the correlation graph that are a different color from the surrounding cells, this may indicate outliers, which are data points that fall outside the typical range of values for a variable. Outliers can have a significant impact on regression models, as they can pull the regression line away from the majority of the data points. If you identify outliers in your correlation graph, you may want to consider removing them from your dataset or transforming your variables to reduce their impact.

**6. EXPERIMENTAL RESULTS**

ACCURACY PERFORMACE OF ALGORITHMS :

****

**FIG ‑4** Accuracy of the models

**7. CONCLUSION**

It appears that the random forest regression algorithm performed the best out of the two algorithms tested, achieving 98% accuracy in predicting Diamond Price Prediction. Linear regression also performed well, with accuracie of 90% respectively. However the other metric such as Accuracy, R Square , MAE, MSE similarly shows the high percentage for the Random forest regression when compared to the Linear regression . based on this we can conclude that the Random Forest regression model can be deployed into production to find the Diamond Price prediction .