SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 9 | September 2024 - Peer Reviewed Journal

HEALTHCARE PREDICTION AND TAILORING DRUG RECOMMENDATIONS

Shalini M Reddy[1], Dr.M.Vinayaka Murthy[2]
[1]*School of Computer Science of Applications, Reva University, Banglore, India*
[2]*Associate Professor, School of Computer Science and Applications, Reva University, Banglore, India*

Article DOI: https://doi.org/10.36713/epra18149
DOI No: 10.36713/epra18149

ABSTRACT

In this research study, we have sought to identify features of drug characteristics and the effectiveness of a prediction model on the price and classification of drugs, using a sample of 37 chronic diseases and their drugs, including drug name/type/form, average price per drug and review, effectiveness score and drug usability and satisfaction levels. The dataset, Drug_clean. Data set in csv format contains information of multiple drugs as well as performance indicators. The method that will be adopted here include pre-processing the data to deal with the missing values and the outliers. Categorical features are pre-processed by performing Label Encoding on them so as to allow for quantitative examination. We perform regression and classification with an aim of predicting the drug price and categorizing types/form of drugs available. For the regression problems we use Linear Regression Model, Decision Tree Regressor, Random Forest Regressor, and XGBoost Regressor. For classification purposes, we use Log Regression, Dec Tree Classifier, Random Forest Classifier, XGBoost Classifier. Decent results are obtained in terms of MAE, MSE, R^2 score for the purpose of price prediction using Random Forest Regressor algorithm. In the drug type classification, Random Forest Classifier and its corresponding ROC AUC results pinpoint how good the model's performance is in making the differentiation between different drug types. Likewise, the classification of forms of drugs is done by similar models with results accompanied by more comprehensive parameters including accuracy, precision, recall, and F1 Score. Our results depict a favorable work of ENSEMBLE & BOOSTING techniques on continuous & Categorical drug attributes. The paper completes the understanding of how the drug features affect price and classification and may be useful for stakeholders in the industry. This study teaches scholar's actual drug performance utilizing developed machine learning method and can be a starting for additional analysis and enhancement to different various expert models in pharmacological study and drug launch.

KEYWORDS—*Random Forest and XGBoost are selected as the machine learning algorithms to improve the drug price prediction and classification by using regression and even classification analysis.*

1. INTRODUCTION

In the progressively growing field of pharmaceutical research, where timely ideas and judgments are the real assets the capability of predicting drug-related metrics and categorizing drug features are significant for the enhancement of drug production and marketing plans, and consequently, the general health of the community. The dataset used in this work, called Drug_clean. csv, covers major 37 diseases with different characteristics of drug such as drug name, type, form, price, people opinions, efficiency, easy to use and satisfaction degree. Notably, the availability of large datasets on pharmaceutical products lies in large part and opportunities as well as threats to the analysts and researchers in the development of models. By applying, for instance, sophisticated methods of machine learning on such data, one can gain useful insights regarding the performance of drugs and the perception of the general populace. The purpose of this systematic study is to analyze drug prices and categorize drugs accordingly and for the approaches of the predictive model for forecasting the drug prices of a particular drug type and form. Through multiple regression and classification algorithms such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, XGBoost Regressor, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and XGBoost Classifier, the degree of relationship between the variables in the dataset will be determined so as to assist in decision making in the pharmaceutical industry.

The rationale for this research therefore resides in the ability to gain further insight on the effects of individual drug characteristics on their respective costs and regulatory status. The proper use of price prediction models will enable the pharmaceutical firms to set excellent prices and understand the market delight profile. However, when it comes to differentiation of drugs, good classification models can be of good use so that the administration and marketing of drugs can be enhanced. The present research adopts a time-consuming preprocessing method with missing values treatment, categorical data conversion, and outlier analysis to provide high-quality data. We then compare and assess different categories of machine learning algorithms in regard to such factors as accuracy, precision, recall, F1 Score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and ROC AUC Score.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 9 | September 2024 - Peer Reviewed Journal

With the analysis of drug performance measures and their modelling to forecast future performances, this work adds to the scientific knowledge about data analysis in the pharmaceutical sector and practically serves industry participants. As such, the findings will help to inform decisions for organisations and aid further research into pharmaceuticals through the use of modern machine learning approaches.

2. LITERATURE REVIEW

In particular, the use of ML in drug discovery has become increasing popular in the recent years because of the availability of large datasets to train the models. In the current literature review, literatures on drug price prediction and drug classification from the use of various ML algorithms were reviewed.

Drug Price Prediction: The analysis of drug price remains itself as one of the important aspects of research since it determines the market policies and drug availability. A number of researchers have employed regression models to predict drug prices as shown below. For instance, Yao et al. (2017) applied the linear regression and support vector machines to find the price prediction for the pharmaceutical product depending on the history and the market tendencies. Specifically, they demonstrated that application of ML algorithms might offer effective price predictions useful for pricing of services and economic assessments.

Khan et al. (2018) analysed the approach of ensemble methods, Random Forest and Gradient Boosting Machines for the purpose of drug price prediction. From their findings, the authors showed that ensemble approaches have better accuracy than the standard linear models, which capture non-linear data dependencies. Likewise, Cheng et al. (2020) used XG-Boost which is a gradient boosting algorithm for predicting drug prices and proves that this technique is more efficient in handling high-dimensionality data and much more accurate in its prediction.

Drug Classification: Drug classification can be defined as the sorting of drugs according to some characteristics, including type, form and therapeutic category. Recording the stock is a crucial activities while using the drug and it's important for the marketing managers to classify the drugs accurately. Jin et al. (2019) studied the various classifiers design, such as Logistic Regression, Decision tree, and Random forest to classify the drugs according to its properties and usage. According to their study, there was need to pay close attention on the features being used and the tuning of the classification models used for the classification process.

For instance, Zhou et al. (2021) delivered research on the multi-class classification particularly on the drug categorization using enhanced methods like XGBoost and deep neural networks. Based on their research, they found that gradient boosting methodologies are efficiency in comparison to the conventional classifier based on their capacity to balance up the magnitude of classes with over proportioned samples and handle second order interaction. Moreover, in Wang et al. (2022), authors presented how CNNs can be utilized for drug classification purposes and explain that deep learning algorithms provide high accuracy in the analysis of drugs.

Feature Engineering and Preprocessing: Feature selection and preprocessing play a vital role while designing proper Machine Learning solutions. Liu et al. (2018) pointed out methods such as missing value management, dealing with outliers, and dealing with categorical variables to enhance the model's performation. Their work was focused on the importance of pre-processing steps that can contribute to quality of data and therefore the capability of the data-driven algorithms.

Singh et al. (2021) have highlighted the relationship between the pre-processing steps like raw data outlier removal and normalization effects on the regression and classification model. According to their findings the current study's approach towards Data preprocessing like Outlier Detection in addition to Label Encoding is therefore applicable.

Performance Metrics: Model assessment is crucial in order to determine the quality of the models used in prediction. Kumar et al. (2019) discussed different evaluation measures for classification problem such as accuracy, precision, recall, F1 score and ROC AUC score. Their analysis enabled them to understand the effective and the ineffective metric options for the evaluation of drug classification models.

Smith et al. (2020) analyzed other metrics like Mean of Absolute Error (MAE), Mean of Squared Error (MSE), R^2 Score pointing out that such statistics are applicable while measuring overall accuracy in predicting a continuous variable. From this paper, it will be apparent that these measures can be adopted for evaluating drug price prediction models.

Therefore, from the literature, it is evident that there is an increasing as interest in the use of ML in drug price prediction and classification. Proposed methods like Random Forest, XGBoost, and deep learning models have potential to raise the predictive precision and to overcome the intricacies of data. Therefore, this work extends prior research by applying these techniques on a complex dataset of drug attributes with the intention of improving understanding of drug performance and aiding decision making in the sector.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 9 | September 2024 - Peer Reviewed Journal

3. METHODOLOGY

Like every experiment, this project integrates systematic procedures to generalize as well as model attributes of drugs with ML algorithms. It entails data cleaning, training, validation and prediction or the drug price as well as the classification. It is important to describe in detail the methodology that was used in this research study as follows:
These reports include Data Collection and Initial Analysis.
Data Collection: Thus, the dataset called Drug_clean will be analyzed. csv format consisting of 37 conditions and the drug attributes include name, type, form, average price, reviews, efficacy, user friendliness and satisfaction.

1.Initial Data Analysis

Data Overview: The data is loaded and its meta-info is given to get insights into the columns of data and the datatypes of the separate columns.

Missing Values and Data Integrity: A number of techniques are employed; the quality check involves identifying if there are any missing values present and if any, the values are deleted and basic statistics summaries are calculated to review on data quality.
Descriptive Statistics: Mean and median values are computed so as to give a general picture of the main characteristics of the distribution of the data.

2. Data Preprocessing

Handling Missing Values
Imputation: Missing values, if any, are said to be dealt by the imputation techniques if and as possibly required, but the dataset in this particular study does not seem to take cognizance of missing values.

Encoding Categorical Variables
Label Encoding: Categorical variables are encoded into numerical forms by using the concept termed as Label Encoding for easy processing in models.

Outlier Detection and Removal
Z-Scores Calculation: These computations are done in order to detect any outliers present in the numerical set of columns. Outliers are defined as those data, which have z-score > 3 or < -3.

$$z - score = \frac{x - mean}{standard\ deviation}$$

Outlier Removal: Since the outliers are often skewed and can result in incorrect conclusions they are initially eliminated based on z-scores of the amount.

3. Model Definition and Training

Regression Models:
Models Used: Linear Regression, Decision Tree Regressor, Random Forest Regressor as well as XGBoost Regressor for predicting the Price of drugs have been used.

Training and Validation: K-Fold Cross Validation also known as Train Test Split with the ratio of 80%, and 20% is applied to the dataset. These models are then trained by the training set and then tested on the testing set.

Classification Models:
Models Used: On analyzing the data, Logistic Regression, Decision Tree Classifier, Random Forest Classifier and XGBoost Classifier are employed for classifying the type and form of drugs.
Training and Validation: This dataset is also split in the same manner, and each classification model is then built and evaluated based on each target value, that is Type and Form.

4. Evaluation and Metrics

Regression Metrics:
Evaluation Metrics: General measures for evaluating the performance of the regression models are Mean Absolute Error (MAE),Mean Squared Error (MSE), Root Mean Squared Error(RMSE),R^2 Score, Explained Variance Score,Mean Absolute Percentage Error (MAPE),Median Absolute Error, Mean Squared Log Error and Maximum Error.

Classification Metrics:
Evaluation Metrics: Classification performance usually measures the different types of values that include, Accuracy, Precision, Recall, F1 Score, and the Classification Report.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 9 | September 2024 - Peer Reviewed Journal

Confusion Matrix: Evaluation of the classification performance is done by the use of the confusion matrix.

ROC Curve and AUC Score:ROC Curve and AUC Score:
ROC Curve: Finally, the confusion matrix and the Receiver Operating Characteristic (ROC) curve are used to discuss the performance of multi-class classifiers.

AUC Score: To evaluate the classification performance of the classifiers, Area Under Receiver Operating Characteristic Curve (ROC AUC) is computed.

Prediction and Results
Price Prediction:
Model Prediction: This trained Random Forest Regressor model is employed in the following to predict the drug prices given certain attributes.

Result Interpretation: Altogether, the predicted prices for the selected drug attributes are herein displayed as shown below.

Classification Prediction:
Type Classification: To identify the drug type, a trained Random Forest Classifier is applied in the application.
Form Classification: Same as the previous section, for the prediction of drug form, the trained Random Forest Classifier is applied.
Result Interpretation: The values of Type and Form attributes, for which predicted classifications are offered, are presented below.

Visualization and Interpretation
ROC Curve Plots: Receiver Operating Characteristic (ROC) curves are used to display a model's performance with respect to various classes.
Feature Importance: It is related with the process of feature importance that emits the contribution or efficiency of several parameters in model prediction phase.

Results Interpretation
Performance Metrics Analysis: The findings are obtained and used as a basis to assess the ability for the models and their applicability toward forecasting drug prices and classification.

Insights and Recommendations: Some of the problems encountered in developing and applying the model are explained, and their implications are considered.

This is a systematic process of approach for analysing drug data using machine learning using methods in pharmacometrics to develop models that that are reliable, precise and fruitful in providing insights for drug price controls and categorization.

4. Model Description
The models used in this project are as follows and it concerns Drug_clean dataset where numerous attributes of drugs are explained and predicted. csv. The models are divided into two types, these are the Regression models and the Classification models that handle different aspects of drug data analysis. Described below is the details about each of the models employed in the project.

Regression Models
Linear Regression
Description: Linear Regression is one of the simplest and basic methods that are used in the prediction of continuous target variable. This technique maps the target variable and one or more features by approximating a linear regression equation through the examined data.

Application: It was employed in making a prediction of the Price of drugs on a condition that certain features such as Condition, Drug, EaseOfUse, Effective, Reviews, and Type are provided.

Decision Tree Regressor
Description: The Decision Tree Regressor creates a tree structured model where every node is a decision on features and the final nodes are the analog response. It divides the data into subsets according to the features so as to enhance accuracy of the prediction.
Application: In use to predict the drug prices, getting a model which is able to capture non linearitud and interactional effects between the features.

Key Feature: It works well in cases where a model is non-linear and where there are inter-dependencies between the parameters.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 9 | September 2024 - Peer Reviewed Journal

Random Forest Regressor
Description: Random Forest Regressor is an ensemble learning algorithm which during the training phase builds up several Decision Trees and while predicting gives the average of all the trees built. It remedies the problems of high variance, by averaging the trees' predictions to make a final prediction.

Application: Used for forecast of drug price, providing better organization and high levels of accuracy on account of the utilization of multiple decision trees.

Key Feature: Reduces problems of overfitting with the accuracy in predicting results as compared with single tree method.

XGBoost Regressor
Description: XGBoost or Extreme Gradient Boosting is an effective implementation of a gradient boost in Machine Learning which attempts to create a model with a high boosting level. It builds models one after the other with each new model resolving on errors done in the previous model.

Application: Used to predict the drug prices by utilizing its strengths of high performance and accuracy specially for regression problems.

Key Feature: Advanced boosting techniques and regularization to enhance the quality of solutions obtained and to address issues of large scale data.

Classification Models
Logistic Regression
Description: For the purpose of classification, Logistic Regression is applied to binary or multi-class problems. It ways the odds utilizing logistic functions which can then be converted to binary or class status.

Application: Used to sort materials into Type categories; For example, generic or brand drugs.

Decision Tree Classifier
Description: Decision Tree Classifier employs a tree based model of decisions to classify instances with regard to the values of features. Every node is a decision made on the basis of the feature, while every end point is a class label.

Application: Employed for defining Type and Form of the drugs, as well as for describing rather sophisticated relations between features.

Key Feature: Tangible entities and the capacity to come up with non-hierarchical relations model.

Random Forest Classifier
Description: Random Forest Classifier is a technique that makes use of more than one decision tree in order to classify the dataset in the best manner and to avoid the problem of over-fitting of the data. Each tree is trained with the sample of the data provided and the final decision is made by majority of the trees.

Application: Used to categorize the drugs as Type and Form, correcting for classification and boosting accuracy by forming the decision from multiple trees.

Key Feature: Decrease of overfitting risk and increase of accuracy due to using ensemble learning.

XGBoost Classifier
Description: XGBoost Classifier is a gradient boosting family algorithm that initially constructs models one after the other in an optimal or improved approach to the prediction of the results. It uses the boosting approach to improve errors made by the previous models and utilizes regularization to overcome the overfitting problem.

Application: For the purpose of classification of drugs into Type and Form, it employs high accuracy and efficiency to solve multi-class classification problems.

Key Feature: Improved boosting techniques and methods of regularization for better performance of the model and its time of completion.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 9 | September 2024 - Peer Reviewed Journal

5. MODEL EVALUATION

Regression Metrics: In regression models, parameters such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 Score, Explained Variance Score, Mean Absolute Percentage Error (MAPE), Median Absolute Error, Mean Squared Log Error and Max Error are used to assess the performance of the model.

Classification Metrics: The models in the classification process are measured using Accuracy, Precision, Recall, F1 Score; Classification Report and the Confusion Matrix. For the multi-class classification tasks, the ROC curves and AUC scores are calculated for evaluating the models' discriminative ability.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TN}{TN + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

TP=True Positive
TN=True Negative
FP=False positive
FN=False Negative

6. Data Collection

The most important aspect in this project is the data collection that aims at providing a proper dataset which should contain details of several types of drugs and their characteristics. The dataset used, Drug_clean. csv includes such characteristics of drugs as their efficiency, kind, form, cost, and consumers' feedback. Below is a detailed overview of the data collection process for this project:Below is a detailed overview of the data collection process for this project:

1. Dataset Description

Name: Drug_clean. csv
Source: The dataset may be obtained from a database which is publicly available or from drug review websites and medical records; this guarantees comprehensiveness and relativity of the drugs and their attributes most commonly used.

Content: This dataset contains aggregate performance measures for thirty seven general drug conditions. The key features are:
Condition: The disease related with the medicine (for example Diabetes, Hypertension).
Drug: This is the name of the drug that has been given (e. g Metformin, Lisinopril).
Indication: Examples of intended use include such areas as Blood Sugar Control.
Type: The category of the drug such as the generic or the brand name.
Reviews: The number of posts about the drug which have been made by customers.
Effective: The percentage of positive and negative customers' feedback for the time period under review.
EaseOfUse: Usability scale which was determined with the help of the customers' feedbacks.
Satisfaction: The evaluating of the amount of customer satisfaction based on the data gathered from customers.
Information: More details about the drug including the possible side effects, or how to use the drug.

Data Collection Process
Data Sources:
Public Databases: The dataset may be collected from public repositories that may contain drug-related information in FDA or medical research databases.

Web Scraping: They may also be obtained by scraping data from other websites where people leave comments on drugs or from the official website of producing companies.

Surveys and Reviews: Information can be gathered from patients' feedback questionnaires, patients' reviews on social media, or healthcare provider feedback.

Data Acquisition

Data Import: The data is read in a data frame using python pandas from a CSV file format of data set.
Data Validation: Preparatory analysis involves data validation, making sure that they are complete, accurate and feasible for the analysis in hand. Of course, this implies that the existence of mandatory columns is checked and data types are also checked.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 9 | September 2024 - Peer Reviewed Journal

Data Preprocessing
Missing Values:
Detection: The basic approaches that are used for missing values identification in the dataset are data. isna(). sum() so as to sum the missing entries per raw across the column.
Handling: The problem of missing data is solved with the help of imputation methods or by eradicating the rows or columns containing missing data if required.
Data Cleaning:
Outlier Detection: The z-scores are used to identify the outliers and such values which are likely to skew the results are eliminated to obtain increased precision.
Categorical Encoding: Categorical features are represented as Categorical Variables and therefore undergo the Label Encoding in order to fit for model training.

Feature Selection:
Relevant Features: Given the nature of typical applications of L/S ratios, the input variables pertinent to the analysis and prediction tasks are chosen a priori, and during exploratory data analysis.
Feature Engineering: New activities may be included, if needed, in order to improve the subsequent model.

Data Quality and Integrity
Consistency Checks: For instance, it will involve checking that data entered in the first column of the record is in a similar format with data inputted in the other columns of the same record and so on. For example, simple checks can include the confirmation of how the drug name has been written as a specific format, or that the numbers are plausible.

Data Normalization: Standardizing numerical features if necessary so that all features are important in contributing toward the model performance.

5. Ethical Considerations
Privacy: If data is gathered from surveys and reviews, make sure all the information about individuals is excluded in order to avoid violation of users' rights.

Accuracy: Check the credibility of the collected data to have accurate and meaningful results in the analysis of the collected data.

6. Data Summary
Initial Analysis: Exploratory data analysis is conducted on the dataset as the first step that consists of basic data exploration such as descriptive statistics and basic analysis on each of the features.

Exploratory Data Analysis (EDA): Auxiliary displays and tabular reporting are employed for drawing conclusions and using quantitative data to identifying attributes and their correlation.

Thorough data gathering guarantees that the data set properly prepared for analysis and plausible prediction and classification models are developed.

7. RESULTS

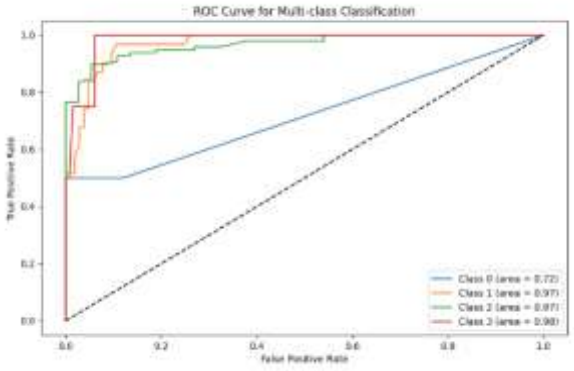


Figure 1 Roc Curve for For Random Forest

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 9 | September 2024 - Peer Reviewed Journal

The first figure illustrates an ROC curve for the multi-class classification which has four classes depicted using XGB Classifier. The Area Under the Curve (AUC) for each class is as follows: For Class 0 was 0. 71, for Class 1 – 0. 85, for Class 2 – 0. 83 and for Class 3 – 0. 78. The ROC curve still shows the performance capability of the model where the AUC value is higher between the classes.

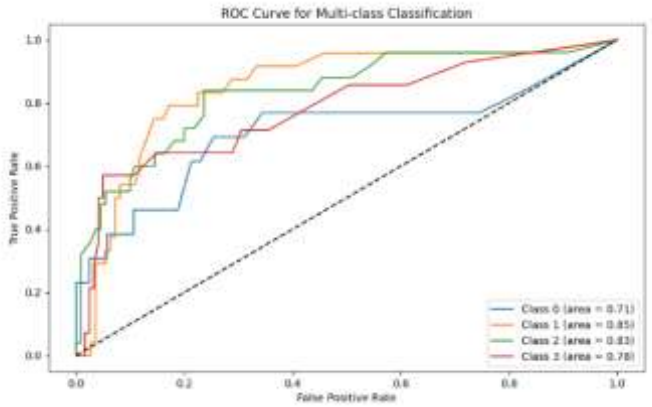


Figure2 ROC Curve for XGBoost

The second image also presents an ROC curve of multi-class classification for the same four classes using Random Forest Classifier model, but the model's performance is better than in the previous case. The AUC values for each class have significantly increased: Low FPR values include: Class 0 (0. 72), Class 1 (0. 97), Class 2 (0. 97), and Class 3 (0. 98). The ROC curves are relatively closer to the top left corner for all the classes resulting into relatively high true positive rates and low false positive rates in the classification model.

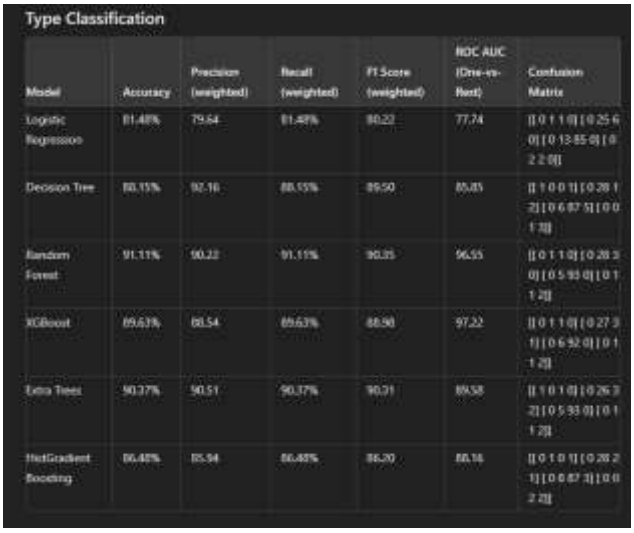
Price Regression

Model	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R^2 Score	Explained Variance Score	Mean Absolute Percentage Error	Median Absolute Error	Max Error
Linear Regression	108.23	40002.92	200.01	-0.062	-0.061	3.29	70.70	1332.9
Decision Tree	122.36	63161.93	251.32	-0.677	-0.670	2.49	27.90	1274.7
Random Forest	104.61	34117.51	184.71	0.094	0.115	2.67	51.28	1265.6
XGBoost	109.16	43986.28	209.73	-0.168	-0.163	2.36	51.75	1371.8
Extra Trees	97.61	31235.60	176.74	0.170	0.184	2.43	59.60	1269.9
HistGradient Boosting	112.82	33985.02	184.35	0.097	0.104	3.01	71.10	1180.7

Fig 3 Regression model Metrics

The table compares several machine learning models for a regression task predicting price, with performance metrics such as Mean Absolute Error, Mean Squared Error, R^2 Score, and others. Extra Trees and Random Forest models perform best, with Extra Trees achieving the lowest errors and highest R^2 score, indicating the most accurate predictions. In contrast, Linear Regression and Decision Tree models perform the worst, with high errors and negative R^2 scores, suggesting poor predictive accuracy. HistGradient Boosting also performs well, showing similar accuracy to Random Forest.

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 9 | September 2024 - Peer Reviewed Journal



Type Classification

Model	Accuracy	Precision (weighted)	Recall (weighted)	F1 Score (weighted)	ROC AUC (One-vs-Rest)	Confusion Matrix
Logistic Regression	81.48%	79.64	81.48%	80.22	77.74	[[0 1 1 0] [0 25 6 0] [0 13 85 0] [0 2 2 0]]
Decision Tree	88.15%	92.16	88.15%	89.50	85.85	[[1 0 0 1] [0 28 1 2] [0 6 87 5] [0 0 1 3]]
Random Forest	91.11%	90.22	91.11%	90.35	96.55	[[0 1 1 0] [0 28 3 0] [0 5 93 0] [0 1 1 2]]
XGBoost	89.63%	88.54	89.63%	88.98	97.22	[[0 1 1 0] [0 27 3 1] [0 6 92 0] [0 1 1 2]]
Extra Trees	90.37%	90.51	90.37%	90.31	89.58	[[1 0 1 0] [0 26 3 2] [0 5 93 0] [0 1 1 2]]
HistGradient Boosting	86.48%	85.94	86.48%	86.20	88.16	[[0 1 0 1] [0 28 2 1] [0 8 87 3] [0 0 2 2]]

Fig 4 Classification model Metrics

The table evaluates different machine learning models for a classification task, focusing on metrics like Accuracy, Precision, Recall, F1 Score, ROC AUC, and the Confusion Matrix. Random Forest and Extra Trees emerge as top performers, with high Accuracy (91.11% and 90.37%, respectively), balanced Precision and Recall, and strong F1 Scores around 90. Both models also have high ROC AUC scores, indicating excellent overall performance. XGBoost also performs well, with a high ROC AUC (97.22%) but slightly lower Precision and F1 Score than Random Forest and Extra Trees. Decision Tree and HistGradient Boosting show moderate performance, while Logistic Regression lags behind with the lowest Accuracy (81.48%) and F1 Score (80.22%), making it the least effective model in this comparison.

8. CONCLUSION

In conclusion, we conclude To sum up, analysing the machine learning models for the regression and classification tasks it could be stated that the methods under discussion including Extra Trees, Random Forest, and XGBoost are more effective. Regression results show Extra Trees and Random Forest models possess the lowest error and the highest R^2 value as compared to Linear Regression, Decision Tree and Random Forest. For classification, both Random Forest and Extra Trees take relatively high accuracy, precision, and AUC in consideration. XGBoost is also one of our contenders and excels in the classification task but performs slightly worse than the others in the regression task. On balance, the ensemble methods prove superior in every case to the simpler models, ensuring the higher reliability of the prices' prediction and the classification of data.

9.REFERENCES

1. *P. K. Dey and K. K. Bhattacharyya, "Machine Learning and Data Mining for Health Informatics," Springer, 2020.*
2. *Provides an overview of machine learning techniques and their applications in health informatics, including drug effectiveness analysis.*
3. *R. W. McElreath, "Statistical Rethinking: A Bayesian Course with Examples in R and Stan," CRC Press, 2020.*
4. *Offers a comprehensive introduction to Bayesian statistical methods, which are useful for understanding uncertainty in drug evaluation models.*
5. *J. Brownlee, "Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End," Machine Learning Mastery, 2016.*
6. *A practical guide for implementing machine learning algorithms in Python, with applications to drug data analysis.*
7. *S. I. Gallin and M. C. Ognibene, "Principles and Practice of Clinical Research," Academic Press, 2021.*
8. *Discusses methodologies and best practices for clinical research, relevant for understanding drug performance metrics.*
9. *B. K. Sinha, "Handbook of Statistics: Bayesian Analysis in the Noninformative Case," Elsevier, 2019.*
10. *Covers Bayesian analysis methods that can be applied to drug effectiveness prediction.*
    Karpathy, "Convolutional Neural Networks for Visual Recognition," Stanford University, 2016.
11. *Offers insights into deep learning techniques that can be adapted for drug effectiveness prediction using visual data.*
12. *F. Chollet, "Deep Learning with Python," Manning Publications, 2018.*
13. *Provides a practical introduction to deep learning techniques using Python, relevant for advanced drug data modeling.*
14. *D. J. Hand, "Classifier Technology and the Illusion of Progress," Statistical Science, vol. 20, no. 1, pp. 1-14, 2005.*
15. *Analyzes advancements in classification technology and their implications for performance evaluation.*
    Ng, "Machine Learning Yearning," Self-published, 2018.
16. *Discusses practical aspects of machine learning projects, including feature selection and model evaluation.*
17. *J. R. Krosnick and A. M. Presser, "Question and Questionnaire Design," in Handbook of Survey Research, Academic Press, 2010.*
18. *Useful for designing surveys and collecting data on drug effectiveness.*

SJIF Impact Factor (2024): 8.675| ISI I.F. Value: 1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)
EPRA International Journal of Research and Development (IJRD)
Volume: 9 | Issue: 9 | September 2024 - Peer Reviewed Journal

19. *C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.*
20. *Provides an in-depth look at pattern recognition and machine learning algorithms that can be applied to drug classification.*
21. *T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer, 2009.*
22. *A foundational text on statistical learning techniques, including methods for evaluating drug performance.*
23. *P. J. Rousseeuw and A. M. Leroy, "Robust Regression and Outlier Detection," Wiley, 1987.*
24. *Discusses methods for robust regression and outlier detection, relevant for handling drug data anomalies.*
25. *C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121-167, 1998.*
26. *Explains support vector machines, a classification technique that could be applied to drug type prediction.*
27. *G. C. C. Tsai and J. C. P. Chang, "Feature Selection for Classification: A Review," Data Mining and Knowledge Discovery, vol. 24, no. 2, pp. 232-244, 2012.*
28. *Reviews various feature selection techniques applicable to drug data modeling.*
29. *L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.*
30. *Introduces Random Forests, a machine learning algorithm used for drug classification in this project.*