National College of Ireland

Project Submission Sheet

Student Name:	Shivansh Bhatnagar		
Student ID:	X23237252		
Programme:	Msc. In Data Analytics (MSCDAD_A)	**Year:**	Sept 24-25
Module:	Data Mining & Machine Learning		
Lecturer:	M. Bradford		
Submission Due Date:	13 Dec 2024		
Project Title:	A COMPREHENSIVE ANALYSIS TECHNIQUES OF DATA MINING AND MACHINE LEARNING PREDICTIVE ACCURACY IN HEALTHCARE, AIRBNB HOTEL LISTING, AND CHATGPT APP REVIEWS TEXT ANALYSIS.		
Word Count:	4242		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Shivansh Bhatnagar

Date: 13 Dec 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only

Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

[Insert Module Name]

[Insert Title of your assignment]

Your Name/Student Number	Course	Date
Shivansh Bhatnagar / x23237252	Msc Data Analytics (MSCDAD_A)	13 Dec 2024

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgment

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
Grammarly	To improve paragraph grammatically.	https://www.grammarly.com/
Microsoft Copilot	To check evaluation and performance of my classification and regression report.	https://copilot.microsoft.com/onboarding

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

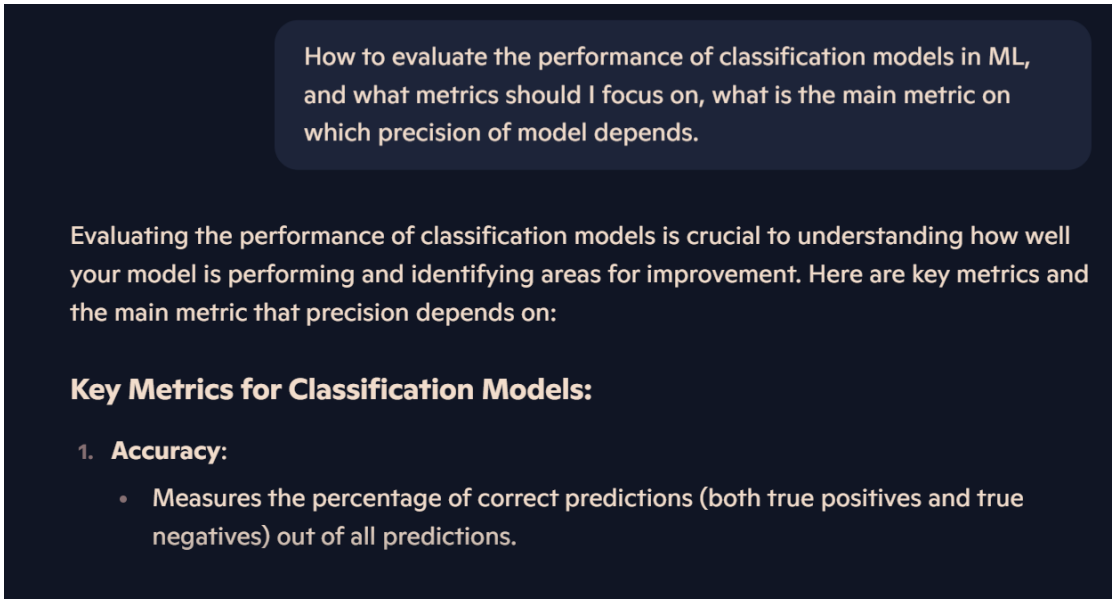
Microsoft Copilot	
To improve paragraph grammatically.	
How to evaluate the performance of classification models, and what metrics should I focus on, what is the main metric on which precision of model depends.	How to Choose the Right Metric 1. Imbalanced Dataset: Focus on precision, recall, F1 score, or PR-AUC. 2. Balanced Dataset: Accuracy or ROC-AUC can be reliable. 3. Domain-Specific Goals: ○ High precision for fraud detection (minimize false

	alarms). ○ High recall for disease diagnosis (minimize missed cases).

Evidence of AI Usage

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence:



Additional Evidence:

3. **Recall (Sensitivity or True Positive Rate):**
   - Measures the proportion of true positive predictions out of all actual positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- Important when the cost of false negatives is high.

4. **F1 Score:**
   - The harmonic mean of precision and recall, providing a balance between the two.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Useful when you need to balance precision and recall.

5. **Confusion Matrix:**
   - A table that summarizes the performance of a classification model by showing the true positives, true negatives, false positives, and false negatives.

A COMPREHENSIVE ANALYSIS TECHNIQUES OF DATA MINING AND MACHINE LEARNING PREDICTIVE ACCURACY IN HEALTHCARE, AIRBNB HOTEL LISTING, AND CHATGPT APP REVIEWS TEXT ANALYSIS.

Shivansh Bhatnagar
x23237252
National College of Ireland, Dublin
x23237252@student.ncirl.ie

Abstract

The project investigates data from three distinct domains with machine learning techniques: Airbnb bookings, Patient health data analysis and ChatGPT reviews text analysis.The aim is to predict the features affecting room price in Barcelona, identifying stroke risk factors, and sentiment analysis of chatgpt reviews. The Airbnb dataset consisted of 19,833 records covering neighborhood details, price of room and Amenities information.XgBoost yielded 74.24% accuracy in predicting the room cost in Barcelona.For healthcare,I have considered the dataset containing 237,630 related to the demographics and health conditions of patients achieved 81% accuracy with Random Forest Classifier. The ChatGPT review dataset comprises 291,878 entries and evaluated sentiment analysis(Positive, Neutral, and Negative categories based on text reviews) and user rating with logistic regression and random forest excelled with 80% and 84%, respectively. *Overall, the* ***project highlights the applicability of machine learning models in various fields, thereby making them applicable for predictive analytics in real-world applications.***

Keywords: Machine Learning, Predictive Accuracy, Ensemble Methods, Random Forest, Gradient Boosting
Healthcare Analytics, Environmental Monitoring, Data Imbalance, Text Analysis, Sentiment Analysis, Prediction, Regression, Classification.

I. INTRODUCTION

Machine learning, as a field of applied computer science, has the potential to drive insights in today's data-driven world. This report focuses on the comparison of several machine learning algorithms intended for three different datasets: Barcelona Airbnb hotel room price predictions, patient health data prediction and ChatGPT review rating prediction. The significance of this lies in the identification of which algorithms is well suited to predict Airbnb hotel room price prediction, patients' chances of heart attack and sentiments of users of ChatGPT by analysing ratings and comments. The insights gained by businesses, healthcare professionals, and Hospitality sectors are tremendous.

A. Objectives

The primary objectives of this project are as follows:

- To Evaluate the performance of different machine learning algorithms, such as Random Forest, Logistic Regression, Decision Tree, Support Vector Classifier, Gradient Boosting, XGBoost and K-Nearest Neighbours, on the Airbnb room price prediction, patient health and ChatGPT user review datasets.
- To identify the most accurate model for each dataset in terms of the performance metrics such as accuracy, R2, precision, recall, F1-score, Mae, RMSE, MSE, MAPE, ROC curve, Cohen's Kappa and confusion matrix analysis.
- To use the different algorithms to explore their strengths and weaknesses, especially as regards handling imbalanced classes-a common challenge in real-world datasets.

B. Prediction Problem Statements(Research Questions)

- Airbnb Room Price Prediction:
  The objective is to predict the factors on which room price of Airbnb impacts by analysing previous trends of rates, locality and various amenities. With accurate identification of top factors on which hotel room cost depends, hotels owner can see our prediction to construct new hotels, to increase revenue, and improve customer service through predictive insights.
- Patient Health Risk Predictive Modelling

This is a prediction task: a patient's risk of experiencing myocardial infarction according to various indicators, including demographics and health, such as age, BMI, and medical history. Reliable predictions help doctors identify patients early and initiate preventive treatment.

- ChatGPT Sentiment/Rating Prediction
  This is a prediction task to analyse the sentiments and ratings of users. It will help to identify common issues or strengths to refine application.

C. Report structure

The paper is structured to include a comprehensive understanding of methodology, results, and future directions towards better improvement of machine learning applications in these domains.

- The motivation behind the research as well as the specific questions being researched form the introduction.
- Next section, "Related Works," consists of a review of previous literature concerning the subject on machine learning applications related to the fields.
- "Data Mining Methodology" describes the process adopted for preprocessing of data, feature selection, and modelling.
- In "Evaluation", each of the models is evaluated for various measures of performance. Results have been discussed in detail.
- The report concludes with a deep summary of key findings and some suggestions for future work which may mitigate the limitations arising during analysis.

II. RELATED WORK

Machine learning (ML) techniques have been applied extensively in various domains, including healthcare, the hospitality and AI industry, to enhance prediction accuracy and decision-making processes. Many researchers have explored the use of a variety of machine learning algorithms on datasets with structured and unstructured data, discussing the merits and demerits of the models in different contexts .

- In the healthcare sector, many studies have applied machine learning to predict medical conditions, especially heart disease prediction. For example, [1] applied Logistic Regression algorithm, Random Forest models , SVC, SVM and Gradient Boosting to predict heart attacks with an accuracy of over 80% but had difficulties with imbalanced datasets, where the minority class, which is patients with heart attacks, [2] was underrepresented. To this limitation, [3] improved on the techniques, We have dropped some rows with 0 counts to make data somewhat balanced and used statistical techniques which include SMOTE for balancing the dataset. Although this research enhanced the prediction accuracy of minority classes, it failed to address the interpretation of models applied, which remains an essential consideration in health care.
- Machine learning models have been widely studied in the hospitality industry, especially in terms of Airbnb Hotel Room Price Prediction on Various Factors . Previous research highlights factors like location, facilities, safety, cancellation policy, booking policy that influencing prices. This Analysis promoted regression models to determine key price factors, showcases insights to cost determination. A study [5] applied XGBoost regression, Decision Tree Regression and Random Forest Regression to predict the factors on which hotel room price depends based on previous trends including location of hotel, size, Facility and Amenities of room.[6] The XGBoost model performed excellently with an accuracy of 74.2%. [7] extends the above research by comparing Decision Tree, XGBoost and Random Forest regressor and showed that XGBoost Regression gives a better interpretability. In general, such studies pay little attention to pre-data techniques such as scaling of feature and handling missing values which will affect the performance of ML model.
- Text analysis using SVM, Logistic regression had been widely performed for sentiment analysis classifications. My Project will extends the methods to classify review/scores pattern in user sentiments. In ChatGPT Sentiment/Review Analysis, ML models have been used to predict Ratings, an

increasingly important aspect due to the growing demand for Artificial Intelligence. [8] used Logistic Regression for Sentiment Analysis and Random Forest for Rating Prediction with high accuracy over 80% each. The Random Forest model performed much better than Logistic Regression model in aspect of accuracy and F1-scores, especially for identifying high ratings. The Logistic Regression struggles with neutral sentiment, indicating room for improvement in handling this class.

- In most predictive modelling tasks, especially those related to rare events like heart attacks, the majority class hugely outnumbers the data, and thus biased predictions occur. discussed several strategies for this problem, including resampling techniques such as SMOTE resample balancing and ensemble methods; [9] however, these techniques usually have trade-offs in terms of computational cost and model interpretability. In addition, studies like have applied Gradient Boosting models to improve the accuracy of prediction in different domains. The authors found that Random Forest obtained top notch performance in terms of accuracy and recall but was not able to achieve high precision when dealing with highly imbalanced data. [10] This shows the trade-off between different performance metrics and highlights the need for testing several models when solving predictive tasks. In addition, KNN, which is commonly employed for classification, [11] also produced encouraging results in some works. For example, applied KNN to predict Heart attack prediction in the healthcare industry [16]. However, the model failed to perform well when there was a high presence of irrelevant features or noise, which is a common scenario in real-world data. Ensemble methods like Gradient Boosting have been more popular because they allow combining many weak models into a strong predictor, thus diminishing the impact of overfitting [18]. demonstrated that Random Forest performed better than other algorithms such as SVC in predicting heart attacks with an accuracy of 81.7% in Patients dataset. However, complexity and lack of interpretability are major challenges especially in domains where understanding the decision-making capability of the model is important.

Strengths and Limitations

The reviewed studies reveal several strengths and limitations in the application of machine learning techniques:

Strengths:

- Utility of Ensemble Methods: Random Forest, SVM and XGBoost are modelling approaches that work well with diverse datasets and attain a high degree of prediction ability including complex domains like health care and hospitality monitoring.
- Improved Performance Techniques: The model robustness and overall performance are increased with methods such as SMOTE for balancing datasets and ensemble techniques.

Limitations:

- Lack of Interpretability: This is while models like KNN and Gradient Boosting are accurate, but they don't have interpretability, which is necessary in fields including healthcare where predicting what will happen is decisive.
- Imbalanced Datasets: Often times the majority class is overshadowed by the minority class, resulting in biased predictions. SMOTE is a technique that need to fine tune and are not applicable to all cases.
- Limited Generalizability: Findings are therefore limited by small sample sizes or restricted to specific regional focuses.

Relevance to Work

This body of related work substantially informs the approach taken in this research. This study contributes to existing findings and works to fill some of the gaps that have been identified in the literature by assessing several machine learning algorithms over various domains. For example, imbalanced datasets are dealt with through appropriate techniques such as resampling. The models tested range from Random Forest, SVC, and gradient Boosting, to Logistic Regression to see how each one would perform with the different datasets. It is also in relation to other existing research because it brings along a much more detailed comparison of the performance metrics including precision,

recall, and F1-score and brings insights on how these models handle feature selection and data preprocessing.

All gaps highlighted in prior research are also covered in this study, that is the lack of interpretability and the need to standardize preprocessing. So, it has ensured systematic preprocessing of data while combining these models in the current study, allowing for a much more robust, comprehensive analysis of the performance of all models across different domains.

III. DATA MINING METHODOLOGY

For Solving project, I have used the CRISP-DM, or Cross-Industry Standard Process for Data Mining methodology. The process that will be discussed in this report is comprised of six phases: Business Understanding(Problem Statement), Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment based on three different data sets.

Fig 1: CRISP-DM

A. Data Selection

The datasets selected have been relevant to real-world problems, such as hotel room price prediction, healthcare and Artificial Intelligence App Reviews; and showing significant potential caliber for application of machine learning techniques to derive some actionable insights [12].

The selected datasets for analysis represent the below insights and applications:

1. Airbnb Room Price Prediction Factors. This data set provides a description of the room price with regard to features like longitude, latitude, Amenities, room type, cleanliness and safety. The information is primarily numerical and categorical.
   Link of Dataset:- https://www.kaggle.com/datasets/zakariaeyoussefi/barcelona-airbnb-listings-inside-airbnb/data
2. Patient Data: This dataset consists of information regarding patients' health. This includes demographic information and health indicators like BMI and history of heart attack. The dataset contains both numerical and categorical data.
   Link of Dataset:- https://www.kaggle.com/datasets/tarekmuhammed/patients-data-for-medical-field
3. ChatGPT user review dataset: This dataset includes user reviews, scores and application versions. Using a text-based dataset, I deployed machine learning models to classify sentiments and predict high/low ratings. Logistic Regression and Random Forest models were evaluated for their performance.
   Link of Dataset:- https://www.kaggle.com/datasets/ashishkumarak/chatgpt-reviews-daily-updated

B. Data Preprocessing

Data preprocessing is an important step in the Machine Learning & data mining process that ensures quality data used for analysis. The following steps were undertaken for each dataset:

- Importing All Libraries and Dataset:

This is first step to perform any analysis in python. Firstly, dataset is imported using pandas library, consequently all libraries which is required in whole process are also imported, for example matplotlib, seaborn, sklearn, numpy, wordcloud.

- EDA(Exploratory Data Analysis):

EDA is used to inspect and understand complete dataset (columns, data types, summary statistics) and it also helps to examine relationships between variables. All Three datasets are described using python features like (describe, info, columns, dtypes, and shape). Multivariate/Bivariate is performed to check relationships and distributions of variables.

-In Airbnb Room Booking Dataset, all the above mentioned steps have been performed followed by exploration steps to check top 10 types of property and room types in hotel to work more thoroughly.

-In Patient Dataset after performing above stated basic exploration, I visualized general health distribution and BMI by sex.

-In the ChatGPT User Review Dataset, I inspected the distribution of variables available, scores, and pair plot of variables to check and understand data and their relationships [19].

After Performing these steps on all datasets, a correlation Matrix was built which helped to check relationships between all numerical variables and the target variable, according to which I selected features and variables to drop.

- Data Cleaning and Preprocessing:

This implementation handled missing values, Duplicate entries in each dataset which were identified in prior stages. Normalization & Standardization of data to make it scalable and normalize values. There is an encoding stage which converts categorical column to numerical.

-In case of the Airbnb room price prediction dataset, missing values were filled by a median of each numerical column, dropped duplicate and NA rows. And Remove $ sign from my "Price" column and convert it into float, because it is our target variable. Than performed transformation by performing label encoding to transform categorical columns. [20] And then print encoding information to see, which number is allotted to which value, refer below Screenshot of my output.

```
Encoding Information:

city: {'Badalona': np.int64(0), 'Barcelona': np.int64(1), 'Barcelonaneta': np.int64(2), 'Bcn': np.int64(3),
property_type: {'Aparthotel': np.int64(0), 'Apartment': np.int64(1), 'Barn': np.int64(2), 'Bed and breakfas
room_type: {'Entire home/apt': np.int64(0), 'Private room': np.int64(1), 'Shared room': np.int64(2)}
instant_bookable: {'f': np.int64(0), 't': np.int64(1)}
```

-In the case of the heart disease dataset, 20% sampling was done, and few rows containing Hadheartattack=0 was dropped due to much high count. After Dropping these rows, a new sample remained with 10,526 instances, after which a Correlation Matrix was generated. Based on this I dropped further irrelevant columns consisting of bad correlations. Next Step of label encoding was performed on the categorical columns for formation of numerical that proved a perfect fit in machine learning algorithms. Moreover, I utilized a standard scaler for important features like BMI, height, weight which previously had scaling problems resulting in normalized features which contributed equally to Data Mining [16].

-In the ChatGPT review dataset, I dropped rows containing NA values. As per the fixed requirement, rows with less than 40 words per row were dropped and new shape of dataset was (10559,9). Further, new variable named sentiment was defined which was divided on the basis of score threshold (Score>3=Positive, Score=3=Neutral, Score<3=Negetive). Additionally, label encoding and TfidfVectorizer were also implemented in the main text column named 'content'. It helped in the conversion of text to numerical features by weighting words according to the repetitions and importance. [13]

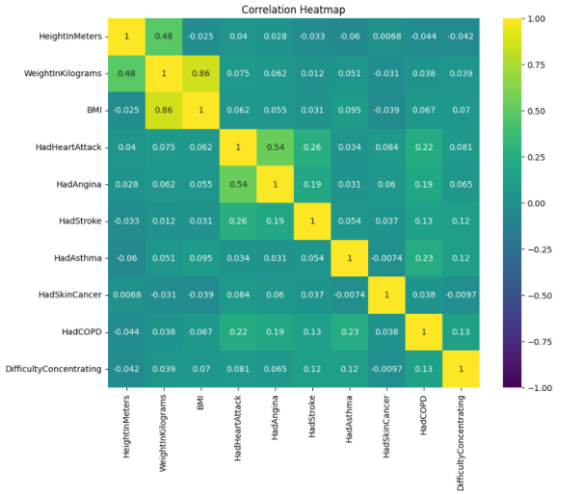


Fig2: Correlation Heatmap of Heart Attack Data

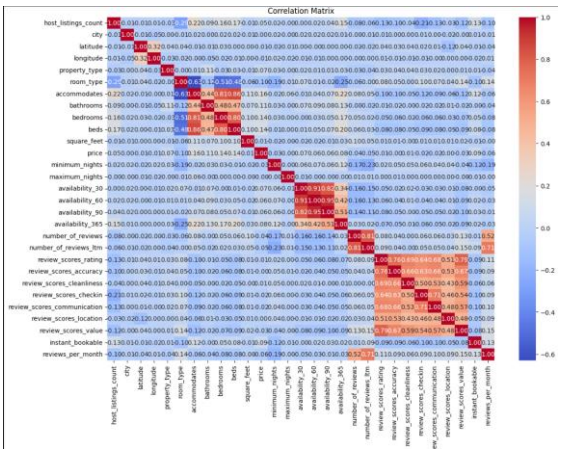


Fig3: Correlation Heatmap of Airbnb Dataset

- Feature Selection

This filter was set to include only the most relevant features for modelling by excluding target variables. For example:

- In Airbnb price prediction data, "price" was dropped in creating a feature set.
- In the patient's data, "HadHeartAttack" was excluded from the features. Used SMOTEN sampling technique to balance minority class to balance minority class by oversampling. Used

RandomUnderSampler to Reduce Majority class and balance dataset [4].

– In ChatGPT user Review Dataset, "content", "score" were dropped from the feature list.

After the feature selection stage, I performed train_test_split to divide the original data into test and train sizes using sklearn model library. I have considered a 20% test and 80% train ratio in all datasets.

C. Modelling Process

This phase of involved the selection of appropriate machine learning algorithms according to the nature of each dataset and the problem to be solved.

– Airbnb Hotel Price Prediction:

a. In order to predict price based on top factors employed several classifiers including XGBoost, Decision Tree Regressor and Random Forest Regressor. Later analysed their performances based on accuracy and classification reports. I selected these models because it has specialty to tackle non linear relationships and performs robust environment to manage feature efficiently.

b. XGBoost regressor performed best among these models in terms R2 value but overall Random Forest gave best result because it contains less errors(MSE,MAE).

– Patients Data:

a. To predict risk prediction amongst age categories Random Forest, Logistic Regression, Decision Tree, Support Vector Classifier, Gradient Boosting Classifier and K-NN algorithm were utilized in this dataset [17]. I implemented all these models to show which Classification model performs best in this type of dataset and can be used in future purpose in healthcare.

b. The repeated use of SVC, Random Forest Classifier and Gradient Boosting was due to the complex interaction in data by using ensemble methods [14].

c. Each model has been trained using a training set derived from 80-20 split as discussed earlier. The best-Performed model among all classifiers was SVC with 81.7% accuracy, 0.53 Cohen's Kappa score and Specificity of 86.67% which proved this model good fit for this dataset,[15] the resultant confusion matrix has been shown below:

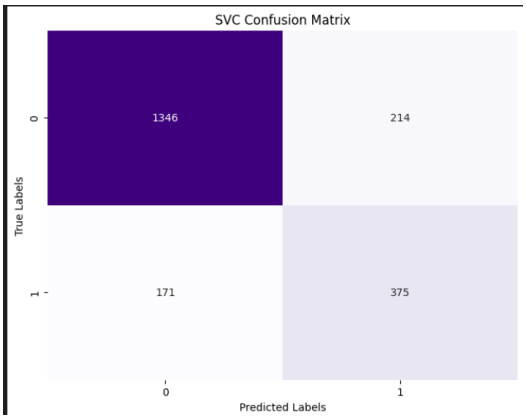


Fig4: Confusion matrix of SVC

– ChatGPT User Review Dataset:

a. Random Forest classifier and Logistic Regression were used in the implementation of Sentiment Analysis and Rating Predictions.

b. Random Forest Classifier was used continually because it handles non-linear relationships well. Logistic regression is used for sentiment analysis because of its simplicity and effectivity with the relationship of variables and can deal with complex interaction in data by using ensemble methods [14].

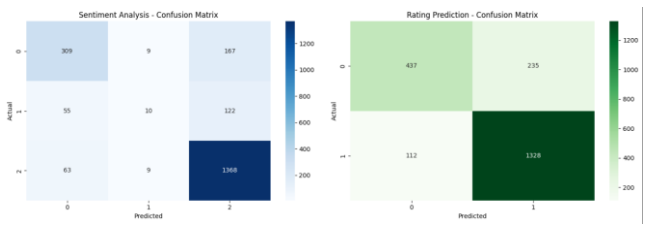


Fig5: Confusion Matrix of ChatGPT Dataset

IV. EVALUATION AND MODEL COMPARISON

– ***Airbnb Room Price Prediction Dataset***

After Performing all the steps mentioned in the Methodology section, I generated several model evaluation factors such as

R2, Mean Square Error(MSE), Mean Absolute Error(MAE) and Mean Absolute Percentage Error(MAPE) using the sklearn module.R2 should be close to 1 as much as possible, MAPE should be under 20% and MSE/MAE should be as low as possible to make both model and prediction ideal. Results of XGboost, Decision Tree, and Random Forest are as follows. Dataset has class imbalances resulting in higher error values.XGBoost regressor was top performer with 74.2 in terms of R2 however, overall Random Forest offers the best trade off between accuracy and errors, making it the most reliable choice for this dataset. Accuracy among these models.

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R² Score	Mean Absolute Percentage Error (MAPE)
XGBoost Regression	53.098121	43269.613313	0.742445	66.553147
Decision Tree Regressor	48.932845	49157.459480	0.707399	56.801206
Random Forest Regressor	52.828356	44545.764089	0.734849	63.129742

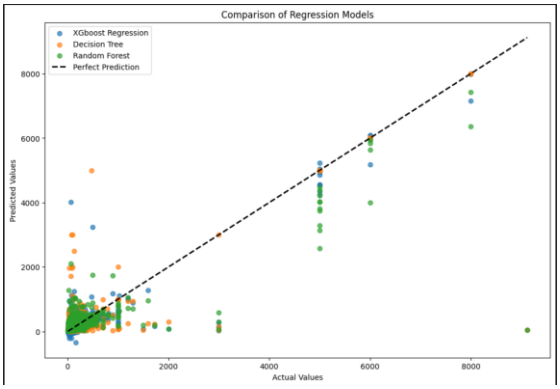


Fig 6: Model Comparisons of Airbnb Data

– ***Patient Heart attack Risk Prediction***

After performing above steps, various classification models were critical to ensure a comprehensive evaluation of performance and help identify the most effective algorithm for predicting heart attacks based on the dataset's characteristics. Each algorithm signifies different strengths and weaknesses. Moreover, I utilized multiple evaluation metrices to validate model prediction strength such as Accuracy, R2, Cohen's Kappa, RMSE, RSS, Sensitivity, Specificity, F-Measure. According to which, Support Vector Classifier(SVC) made the best prediction. Below table explained the strengths and weakness of each employed model:-

Model	Strengths	Weaknesses
Random Forest	High specificity (85.6%) and robust performance with good accuracy (81.1%). Handles non-linearity well.	Lower sensitivity (68.3%) may miss positive cases. Slightly higher RMSE (0.43).
Logistic Regression	Balanced sensitivity (69.4%) and specificity (85.2%). Simple, interpretable, and efficient.	Struggles with non-linear relationships. Similar RMSE (0.43) as Random Forest.
Decision Tree	Easy to interpret.	Lowest accuracy (73.6%), sensitivity (63.3%), and high RMSE (0.51), prone to overfitting.
Support Vector Classifier	Highest accuracy (81.7%) and specificity (86.2%), suitable for high-dimensional data.	Computationally expensive and slightly lower sensitivity (68.6%).
Gradient Boosting	Balanced performance with good F-Measure (66.3%). Performs well on imbalanced data.	Marginally lower accuracy (81.0%) and high complexity, slightly lower specificity (84.1%).
K-Nearest Neighbors	Simplicity and effectiveness on small datasets.	Lower accuracy (76.0%), sensitivity (67.8%), and specificity (78.9%), sensitive to noise.

	Model	Dataset	Accuracy	R2	Cohen's Kappa	RMSE	RSS	Sensitivity	Specificity	F-Measure
0	Random Forest	Patient Heart Risk	0.811491	0.018407	0.523664	0.434176	397	0.683150	0.856410	0.652668
1	Logistic Regression	Patient Heart Risk	0.811491	0.018407	0.526943	0.434176	397	0.694139	0.852564	0.656277
2	Decision Tree	Patient Heart Risk	0.736467	-0.372253	0.371830	0.513354	555	0.633700	0.772436	0.554932
3	Support Vector Classifier	Patient Heart Risk	0.817189	0.048077	0.535917	0.427564	385	0.686813	0.862821	0.660793
4	Gradient Boosting Classifier	Patient Heart Risk	0.810066	0.010989	0.532225	0.435814	400	0.721612	0.841026	0.663300
5	K-Nearest Neighbors	Patient Heart Risk	0.760209	-0.248626	0.427798	0.489685	505	0.677656	0.789103	0.594378

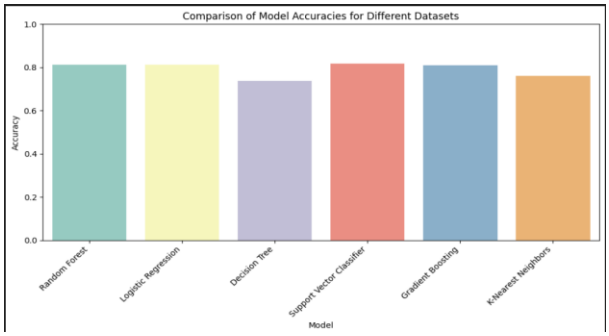
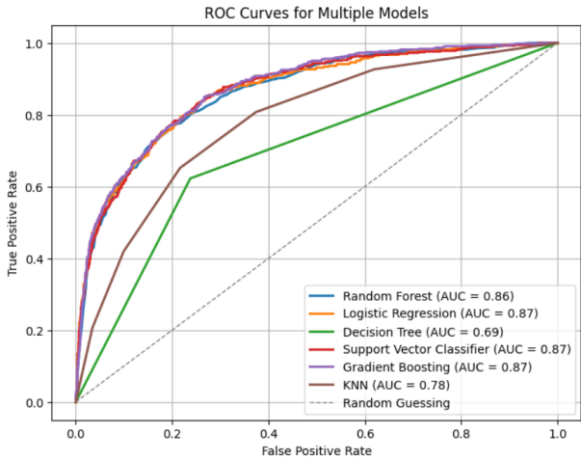


Fig7: Showing Performance of Models on Patient Dataset.

To validate the performance of the models I implemented ROC curve. Underwhich, AUC(Area under curve) measured the overall performance. Ideal AUC must be over 0.70, diagonal represents 0.50 which means 50%.

As per ROC-AUC graph below, Logistic Regression, SVC, and Gradient Boosting are Performing equally well, While Decision Tree performance was questionable being under 0.7.



– ***ChatGPT USER REVIEW DATASET(TEXT ANALYSIS)***

Following the previously outlined steps further evaluation stages including Accuracy, Precision, Recall, f1-score, support were carried out. For sentiment analysis, Logistic Regression is adequate but struggles with the imbalanced data.[21] For rating prediction, random forest proved to be a better fit, providing overall high accuracy and balanced performance [22].

Sentiment Analysis (Logistic Regression)

	precision	recall	f1-score	support
Negative	0.72	0.64	0.68	485
Neutral	0.36	0.05	0.09	187
Positive	0.83	0.95	0.88	1440
accuracy			0.80	2112
macro avg	0.64	0.55	0.55	2112
weighted avg	0.76	0.80	0.77	2112

High/Low Rating Prediction (Random Forest)

	precision	recall	f1-score	support
Low Rating	0.80	0.65	0.72	672
High Rating	0.85	0.92	0.88	1440
accuracy			0.84	2112
macro avg	0.82	0.79	0.80	2112
weighted avg	0.83	0.84	0.83	2112

.

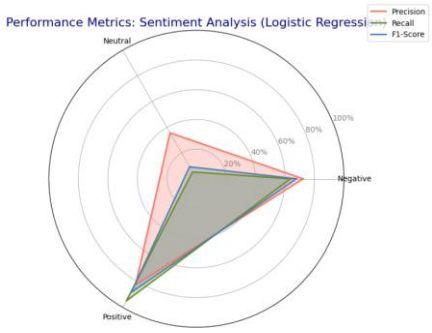


Fig8: Performance Metrics of Text Analysis

V. CONCLUSIONS AND FUTURE WORK

Summary of Findings

Diverse machine learning models were executed to analyze three distinct datasets. Resulting in considerable differences in the model performances among datasets and algorithms.

- Airbnb Price Prediction Dataset:The Random Forest classifier performed accurately with 73.4% result and predicted top 10 Features according to which price fluctuate.

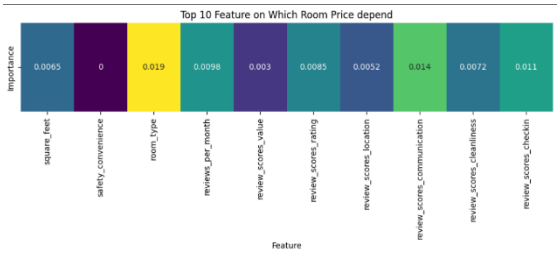


Fig9: Prediction based on ML model on Airbnb room price prediction dataset

- Stroke Risk Prediction Dataset: Support Vector Classifier achieves optimal performance with 81.7% Accuracy. Below is the visualization of prediction using “svc_pred= svc.predict(patients_X_test)” for the age category.The chart shows the Deployment of predicted stroke (MI) risk distribution across age categories. Younger age groups (e.g., 18–34) are predominantly in red, indicating lower risk, while older groups (55 and above) dominate the "MI" category (green), indicating a higher predicted risk. This implies that heart attack risk increases with age.

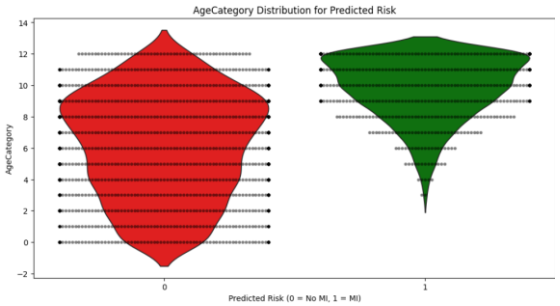


Fig10: Prediction of Heart attack risk

- CHATGPT user review dataset:- Random forest demonstrates superior performance in predicting rating and sentiments with an accuracy of 84%. Analysis revealed there were increased positive, reviews in comparison to previous year trends.

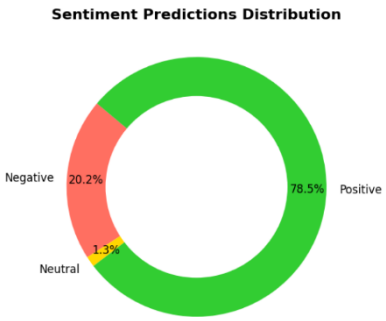


Fig11: Sentiment Prediction Distribution

Limitations:

Issues concerning data quality, such as missing values and class imbalances—especially in the case of patients' data may have impacted the performance of models. Computational constraints limited the ability to perform extensive hyperparameter tuning or to investigate more complex ensemble methods. [16]

Future Directions

Future studies may include the availability of more datasets containing diversity in features or larger sample sizes that

enhance the generalization ability of the study. This approach shall employ sophisticated machine learning techniques such as deep learning or ensemble methods yielding more optimal results, especially for complex tasks like health prediction. Similarly, analyzing minority classes in patients' health data sets by using class imbalance handling techniques, such as SMOTE (Synthetic Minority Over-sampling Technique). Finally, further study in feature engineering and selection may unveil hidden patterns in the data, resulting in improved model accuracy and reliability in different applications.

VI. References

[1] Mall, S., 2024, May. Heart Attack Prediction using Machine Learning Techniques. In 2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1778-1783). IEEE.

[2] Rose, J.S., Bruntha, P.M., Selvadass, S. and Rajath, M.V., 2023, March. Heart Attack Prediction using Machine Learning Techniques. In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 210-213). IEEE.

[3] Verbiest, N., Ramentol, E., Cornelis, C. and Herrera, F., 2014. Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. Applied Soft Computing, 22, pp.511-517.

[4] Dablain, D., Krawczyk, B. and Chawla, N.V., 2022. DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. IEEE Transactions on Neural Networks and Learning Systems, 34(9), pp.6390-6404.

[5] Yang Yuchao. Daily Rental Price Prediction of Shared Accommodation Based on Tree Ensemble Model [D]. Northeastern University of Finance and DOI:10.27006/d.cnki.gdbcu.2022.000714.
Economics, 2022.

[6] Jiang Yujie. Online Short-term Rental Price Prediction [D]. Guizhou University of Finance and Economics,2019.DOI:10.27731/d.cnki.ggzcj.2019. 000153.

[7] Zhao Yunfei, Lou Feng, Cheng Yuan. Construction of Macroeconomic Leading Index System Based on Random Forest Algorithm [J]. Investigation & Research World, 2024, (04): 3-15. DOI:10.13778/j.cnki.11-3705/c.2024.04.001.

[8] Nadeem, M. W., Ghamdi, M. A. A., Hussain, M., Khan, M. A., Khan, K. M., Almotiri, S. H., & Butt, S. A. (2020). Brain tumor analysis empowered with deep learning: A review, taxonomy, and future challenges. Brain sciences, 10(2), 118.

[9] Anwar, A., & Majid, M. (2021). An analysis of machine learning models for heart attack prediction. International Journal of Engineering and Technology, 13(2), 119-126.

[10] Zhu, C., Li, X., Wang, Y., Zhang, H., & Zhou, J. (2021). An overview of machine learning applications in cardiovascular disease prediction. Biomedical Engineering Online, 20(1), 1-18.

[11] Ghorbani, R., Ghousi, R., Makui, A. and Atashi, A., 2020. A new hybrid predictive model to predict the early mortality risk in intensive care units on a highly imbalanced dataset. IEEE Access, 8, pp.141066-141079.

[12] Fiorentini, N. and Losa, M., 2020. Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures, 5(7), p.61.

[13] Sahin, E.K., 2020. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Applied Sciences, 2(7), p.1308.

[14] Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, pp.1937-1967.

[15] Rahman, M. and Kumar, V., 2020, November. Machine learning based customer churn prediction in banking. In 2020 4th international conference on electronics, communication and aerospace technology (ICECA) (pp. 1196-1201). IEEE.

[16] Imron, M.A. and Prasetyo, B., 2020. Improving algorithm accuracy k-nearest neighbor using z-score normalization and particle swarm optimization to predict customer churn. Journal of Soft Computing Exploration, 1(1), pp.56-62.

[17] Mueen, A., & Mohiuddin, A. K. (2020). Prediction of heart disease using machine learning algorithms: a review. Journal of Medical Systems, 44(6), 111.

[18] Y. Indulkar and A. Patil, "Comparative Study of Machine Learning Algorithms for Twitter Sentiment Analysis," 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2021, pp. 295-299.

[19] Yu Zhou, Yanxiang Tong, Ruihang Gu and Harald Gall: Combining text mining and data mining for bug report classification, Journal of Software, Evolusion and Process, Vol.28, pp.150–176 (2016)

[20] Chen, S., Ngai, E.W., Ku, Y., Xu, Z., Gou, X. and Zhang, C., 2023. Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction. Decision Support Systems, 170, p.113959.

[21] MikuWatanabe,YutaroKashiwa,BinLin,ToshikiHirao,Ken'IchiYamaguchi, andHajimuIida.2024. LikeReviewsByChatGPT?.InEASE2024.375–380.

[22] D. Carabantes, J. L. González-Geraldo, y G. Jover, "ChatGPT could be the reviewer of your next scientific paper. Evidence on the limits of AI assisted academic reviews", EPI, vol. 32, no 5, 2023.