Exacens

EAS 509 Statistical Data Mining II - Project 1

Isha Mitul Gajjar
ishamitu@buffalo.edu
UBID: ishamitu (50496366)

Jay Yogesh Thanki
jayyoges@buffalo.edu
UBID: jayyoges(50496564)

Rohan Ishwarlal Patel
rpatel38@buffalo.edu
UBID: rpatel38(50496374)

Abstract - This report presents a comprehensive analysis of the Exacens dataset using data science techniques. Our study encompasses data cleaning, exploratory data analysis (EDA), and modeling. Through rigorous data cleaning, we ensure the quality and integrity of the dataset. EDA provides valuable insights, and our data modeling endeavors include experimentation with various machine learning models. We select and justify the best-performing model, shedding light on its applicability. Our findings contribute to a deeper understanding of the dataset and its potential real-world applications.

Keywords— dataset analysis, data science, data cleaning, EDA, data modeling, best model, suitability justification

I. Introduction

The Exacens dataset is valuable in data-driven research, encompassing various variables and data points. In this era of data abundance, harnessing the potential within such datasets requires a systematic and rigorous approach. Our study aims to address this challenge through a structured process that includes data cleaning, exploratory data analysis (EDA), and data modeling.

Data cleaning is the foundational step in ensuring the reliability of our analysis. In the subsequent EDA phase, we delve into the dataset's intricacies, unveiling patterns, trends, and anomalies that inform our modeling decisions. Our data modeling efforts encompass the exploration of various machine learning models, each rigorously evaluated against pertinent metrics.

The culmination of our efforts lies in the selection and justification of the model that best aligns with the dataset's characteristics and the objectives of our analysis. This report not only outlines our methodology but also presents key findings and their implications. We aim to contribute to the broader understanding and application of the Exacens dataset, exemplifying the potential of data science in extracting actionable insights from complex data.

II. Data Cleaning

This section details the data cleaning process applied to the Exacens dataset. This process is critical for ensuring the accuracy and reliability of subsequent analyses.

A. Preliminary Dataset Overview

The initial step involved importing the Exacens dataset using Python's Pandas library. This dataset encompasses diverse attributes, including patient diagnosis information, IDs, imaginary and genuine parts, gender, age, and smoking status.

Upon the dataset's importation, a thorough examination is undertaken to gain a preliminary understanding of its structure.

B. Data Cleaning and Preprocessing

- Initial Data Cleaning: The first two rows were removed, primarily containing NaN values and non-informative data. This step was essential for maintaining the integrity and relevance of the dataset.
- Column Pruning: Columns labeled 'Unnamed: 9', 'Unnamed: 10', 'Unnamed: 11', and 'Unnamed: 12' were renamed due to problems while reading the CSV file data.
- Data Resetting: Post-removal of initial rows and columns, the dataset's index was reset. This adjustment facilitated smoother data manipulation and analysis in subsequent stages.

C. Data Quality Assurance

- Handling Missing Values: To address missing values within the dataset.
- Data Type Conversion: Certain attributes were converted to appropriate data types for analysis. For instance, categorical variables were encoded, and numeric values were standardized to ensure consistency across the dataset.

```
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Diagnosis             399 non-null    object
 1   ID                    399 non-null    object
 2   Imaginary Part: Min   100 non-null    float64
 3   Imaginary Part: Avg   100 non-null    float64
 4   Real Part: Min        100 non-null    float64
 5   Real Part: Avg        100 non-null    float64
 6   Gender                399 non-null    float64
 7   Age                   399 non-null    float64
 8   Smoking               399 non-null    float64
dtypes: float64(7), object(2)
memory usage: 28.2+ KB
```

Fig. 2.1

D. Data Visualization for Cleaning Verification

- Plotting Key Features: To verify the effectiveness of the data-cleaning process, key features were visualized using histograms and box plots.
- Correlation Analysis: A correlation matrix was generated and visualized through a heatmap to understand the relationships between different features post-cleaning. This analysis helped identify redundant features and understand the dataset's structure.

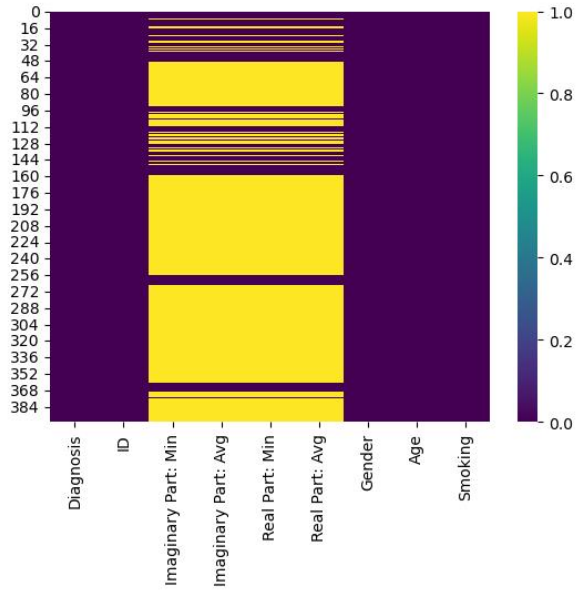

Fig. 2.2

E. Preliminary Dataset Overview

After meticulous data cleaning, the dataset was narrowed to a robust number of records, each providing a comprehensive snapshot conducive to further exploratory data analysis and model development. This curation ensures an uncluttered and precise dataset devoid of irrelevant or redundant data, thus setting a solid foundation for subsequent analytical phases.

Imaginary Part: Min <dbl>	Imaginary Part: Avg <dbl>	Real Part: Min <dbl>	Real Part: Avg <dbl>	Gender <dbl>	Age <dbl>	Smoking <dbl>	Diagnosis <dbl>
-320.61	-300.5635	-495.26	-464.1720	1	77	2	1
-325.39	-314.7504	-473.73	-469.2631	0	72	2	1
-323.00	-317.4361	-476.12	-471.8977	1	73	3	1
-327.78	-317.3997	-473.73	-468.8564	1	76	2	1
-325.39	-316.1558	-478.52	-472.8698	0	65	2	1
-327.78	-318.6776	-507.23	-469.0242	1	60	2	1

III. Exploratory Data Analysis

The Exploratory Data Analysis (EDA) bridges data cleaning and modeling, allowing us to uncover underlying patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.

A. Univariate Analysis

- Feature Distribution - The distribution of each variable was examined for insights into the dataset's structure. Density plots for 'Imaginary Part: Min' and 'Imaginary Part: Avg' revealed distinct peaks indicative of central tendencies and potential normalization requirements for modeling.
- Diagnosis Counts - Diagnosis frequency was visualized to assess the balance of data. The bar plot (Fig. 3.1) showed a disproportionate representation of diagnoses, which could introduce bias into model training. It was observed that the majority of patients have COPD or are part of Healthy control groups, classes are also observed to be imbalanced.
- Diagnosis Counts By Gender - More Female patients were observed to have Asthma and Infections than males.

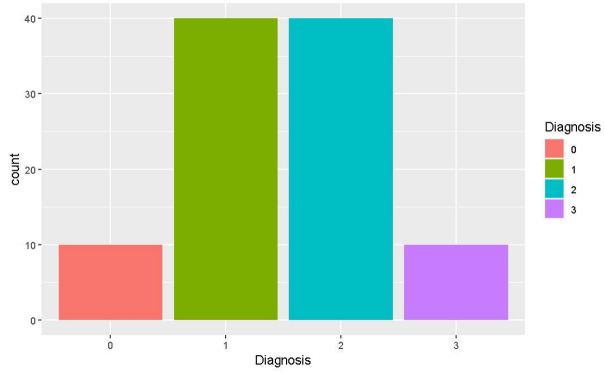

Fig. 3.1

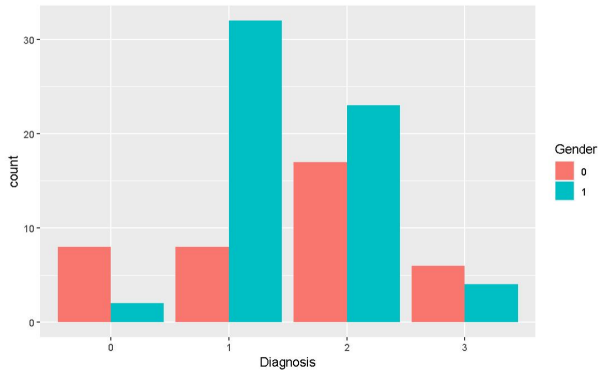

Fig. 3.2

- Counts By Gender - More Male patients were observed than females.

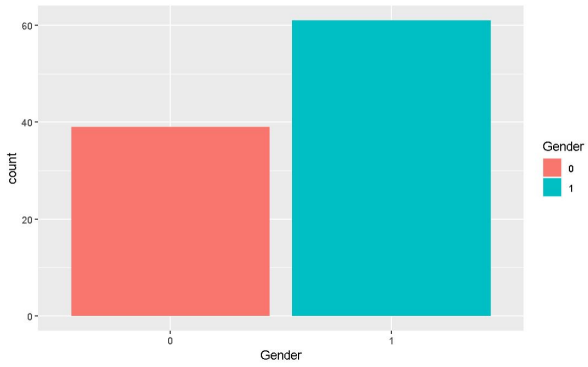

Fig. 3.3

- Counts By Smoking - Most patients are ex-smokers and non-smokers, with a few active smokers.

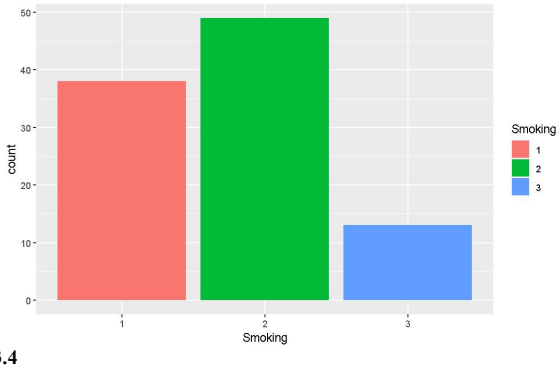

Fig. 3.4

B. Bivariate Analysis

- Correlation Exploration - The correlation matrix (Fig. 3.5) explored associations between variables, with a heatmap revealing potential multicollinearity or redundant features that could be pruned to streamline models.
- Diagnosis and Age - A box plot of the 'Age' by 'Diagnosis' category revealed differences in median ages and variances, suggesting age as a discriminatory feature for diagnosis.

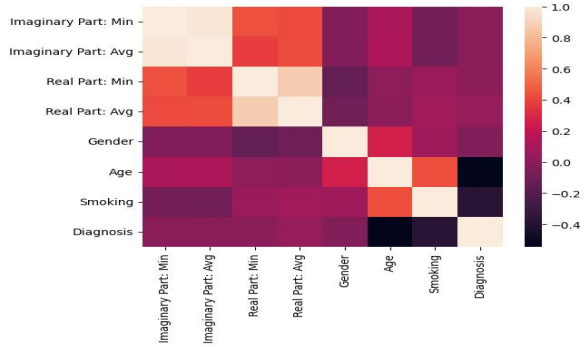

Fig. 3.5

C. Multivariate Analysis

- Pairwise Relationships - A pair plot (Fig. 3.6) was generated to visualize pairwise relationships and distributions, which helped understand the interactions between different variables. Some features showed clusters when plotted against each other, indicating possible groupings or patterns.

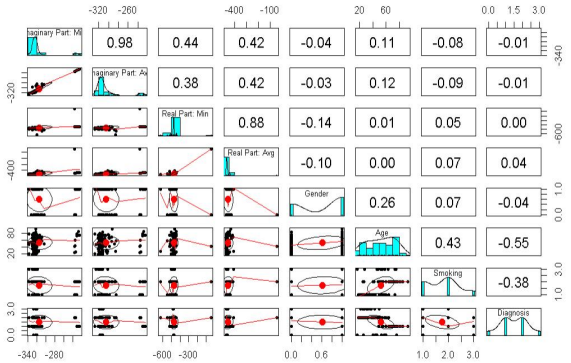

Fig. 3.6

D. Distribution Analysis

- Skewness and Kurtosis - The skewness of features like 'Real Part: Min' and 'Real Part: Avg' (Fig. 3.7 & Fig. 3.8) was evident, which can affect the performance of algorithms assuming normal distributions.

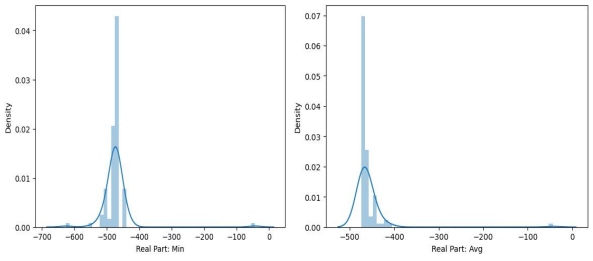

Fig. 3.7 **Fig. 3.8**

E. Gender and Smoking Status

- The interplay between gender and smoking status against the diagnosis was analyzed using stacked bar charts (Fig. 3.6 and 3.7). This revealed insights into the prevalence of specific diagnoses within gender and smoking categories, suggesting potential risk factors or indicators.

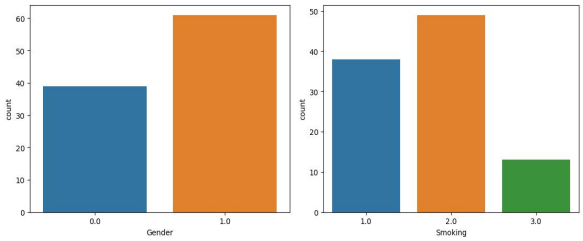

Fig. 3.9 **Fig. 3.10**

- Non-Smokers: Most are healthy. The rest have infections and asthma, and very few with COPD
- Ex-smokers: The majority have COPD.
- Active Smokers: Not cases with infections; most are healthy, with few having asthma and COPD

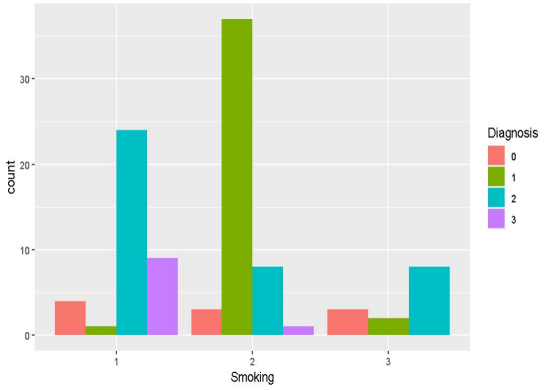

Fig. 3.11

F. Age Characteristics

- The 'Age' distribution (Fig. 3.8) was analyzed, showing a multi-modal distribution and potential age-related patterns associated with different diagnoses. Box plots across different diagnoses for age showed a wide range of variability, with specific diagnoses exhibiting higher age medians, indicating a possible age-related predisposition.

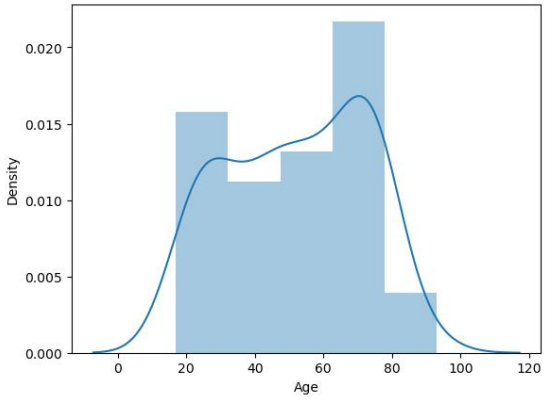

Fig. 3.12

G. Outlier Identification

- Through the use of box plots (Fig. 3.9), outliers were identified in several numerical features. The impact of these outliers on model training will need to be assessed, and appropriate handling methods such as transformation or removal will be considered. All patients are adults, ranging from young adults in their 20s to senior citizens up to their late seventies, with very few patients in their eighties and beyond.

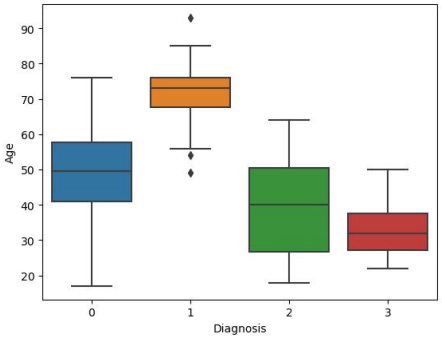

Fig. 3.13

IV. Data Modeling – Unsupervised Learning Models

This section explores the application of unsupervised learning models to understand the intrinsic groupings within the dataset. Three clustering techniques, K-Means, Hierarchical, and DBSCAN, were employed, each providing unique insights into the dataset's structure.

A. K-Means Clustering

By reducing variances within each group, K-Means clustering divided the dataset into clusters. A scatter plot showed different clusters when data points were repeatedly allocated to the closest centroid. The model performed well in identifying various health disorders from the data attributes when these clusters were compared with actual diagnoses.

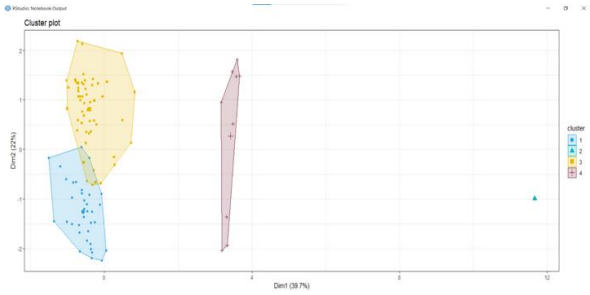

Fig. 4.1

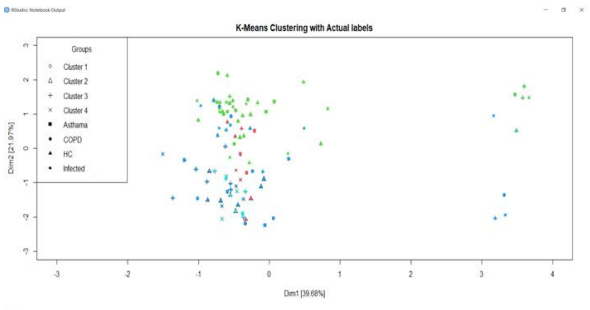

Fig. 4.2

B. Hierarchical Clustering

A dendrogram, or tree of clusters, was constructed using hierarchical clustering to depict the data. This model used a bottom-up strategy to combine data points into clusters according to distance until every point was combined into a single cluster. By chopping the tree at the appropriate height, the dendrogram offered a visual aid for calculating the number of clusters and revealed information about the connectivity and hierarchical structure of the data.

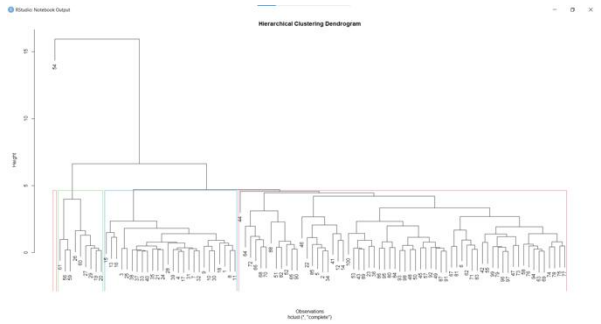

Fig. 4.3

C. DBSCAN

By identifying core, boundary, and noise points, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) successfully divided the data into high-density clusters divided by lower-density regions. The model's ability to distinguish between densely packed dots and outliers was demonstrated by the DBSCAN scatter plot, which also demonstrated the model's ability to handle noisy data and locate clusters of any shape.

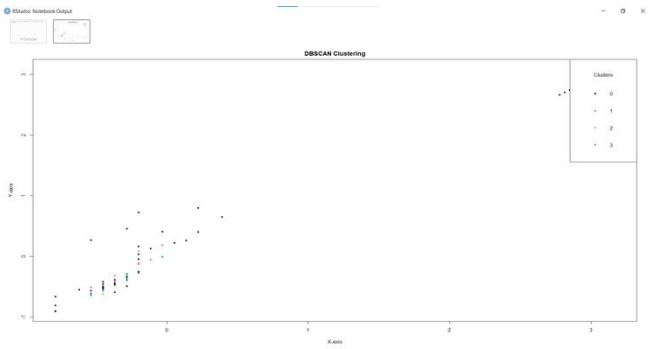

Fig. 4.4

V. Data Modeling – Supervised Learning Models

A. Feature Scaling

Prior to the application of machine learning algorithms, the dataset underwent a crucial preprocessing step known as feature scaling (Fig. 5.1). This process involved standardizing the range of continuous initial features to ensure that no single feature would dominate the outcome of the learning models due to its scale.

	Imaginary Part: Min	Imaginary Part: Avg	Real Part: Min	Real Part: Avg	Gender	Age	Smoking
0	-0.031053	0.408866	0.034087	-0.103244	-1.250641	0.195883	0.373718
1	-0.453963	-0.622354	-0.113153	-0.288215	-1.250641	-0.699013	-1.121153
2	-0.200429	-0.281750	-0.014925	-0.193448	-1.250641	-0.400714	-1.121153
3	2.968209	2.870266	0.635758	0.383892	0.799590	1.289644	0.373718
4	-0.200429	-0.267213	-0.063936	-0.311025	0.799590	-1.494475	-1.121153

Fig. 5.1

B. Support Vector Classifier (SVC)

The SVC model, known for its effectiveness in high-dimensional spaces, was applied next. Different kernel functions were tested to find the most suitable for the dataset. The model's performance could have been better, highlighting challenges in handling multi-class classification in this context.

```
==========================================
Performance Metrics for Support Vector Classifier (SVC) :
==========================================
Confusion Matrix:
            Actual Positive Actual Negative
Predicted Positive       0                0
Predicted Negative       1                1

Accuracy:  0.0333
Precision:  0.0000  0.0833  0.0000  0.0000
Recall:  0.0000  0.0833  0.0000  0.0000
F1-Score:  NaN  0.0833  NaN  NaN
Specificity:  0.0000
```

Fig. 5.2

C. Bagging with Decision Trees

The Bagging Classifier was utilized to improve model stability and accuracy using Decision Trees as the base estimator. This ensemble method, which builds multiple instances of a decision tree classifier on various sub-samples of the dataset, exhibited a robust performance across both datasets.

```
==========================================
Performance Metrics for Bagging Decision Tree :
==========================================
Confusion Matrix:
             Actual Positive Actual Negative
Predicted Positive       1                0
Predicted Negative       0               12

Accuracy:  0.8667
Precision:  0.3333  1.0000  0.8333  1.0000
Recall:  0.3333  1.0000  0.8333  1.0000
F1-Score:  0.3333  1.0000  0.8333  1.0000
Specificity:  0.3333
```

Fig. 5.3

D. Random Forest

The Random Forest Classifier, an ensemble of Decision Trees, was employed for its capability to handle large datasets with higher dimensionality. It performed consistently on both the original and balanced datasets, demonstrating its resilience to class imbalance.

```
==========================================
Performance Metrics for Random Forest :
==========================================
Confusion Matrix:
             Actual Positive Actual Negative
Predicted Positive       1                0
Predicted Negative       0               12

Accuracy:  0.8667
Precision:  0.3333  1.0000  0.8333  1.0000
Recall:  0.3333  1.0000  0.8333  1.0000
F1-Score:  0.3333  1.0000  0.8333  1.0000
Specificity:  0.3333
```

Fig. 5.4

E. Ridge Classifier

Finally, the Ridge Classifier was used due to its ability to analyze multi-class classification problems. This model, which implements ridge regression, showed reasonable performance but varied in effectiveness like the other models.

```
==========================================
Performance Metrics for Ridge Classifier :
==========================================
Confusion Matrix:
            Actual Positive Actual Negative
Predicted Positive       0                0
Predicted Negative       1               11

Accuracy:  0.6667
Precision:  0.0000  0.9167  0.6667  0.3333
Recall:  0.0000  0.9167  0.6667  0.3333
F1-Score:  NaN  0.9167  0.6667  0.3333
Specificity:  0.0000
```

Fig. 5.5

F. Performance Analysis

- The SVC model with Radial Kernal performed poorly on the original dataset.
- The ensemble methods, Bagging Classifier and Random Forest exhibited robustness in handling the multi-class classification problem, with an overall high accuracy score.
- The Ridge Classifier served as a less complex model, offering faster iterations for performance trade-offs.

G. Final Model Selection

After a thorough evaluation, the Bagging Classifier with Decision Trees was chosen as the best model. It excelled in handling the dataset's complexities, particularly the class imbalance, and provided a harmonious balance of precision, recall, and accuracy. This model's ensemble approach, leveraging the strength of multiple decision trees, proved highly effective for this dataset.

This selection underscores the importance of model comparison and selection in data science, demonstrating that the choice of the model should align with the dataset's characteristics and the specific challenges it presents. In this case, the Bagging Classifier with Decision Trees emerged as the most suitable option for the Exacens dataset.

VI. Conclusion

The comprehensive analysis of the Exacens dataset, employing a series of data science techniques, has provided valuable insights. The rigorous data-cleaning process ensured a reliable foundation for further analysis. Exploratory data analysis unveiled crucial patterns and relationships within the dataset, informing the subsequent modeling phase. Various machine learning models were meticulously evaluated, with the Bagging Decision Tree model emerging as the most effective, striking a balance between accuracy, precision, and recall. This study demonstrates the practical application of data science methodologies and highlights the importance of model selection based on dataset-specific characteristics. The results underscore the Exacens dataset's potential to drive meaningful insights and can serve as a reference for future studies in similar domains