Medical Diagnosis of Exasens Dataset

Statistical Learning II | Project 1

Sushmitha Manjunatha *Sahithya Arveti Nagaraju* *Mounika Pasupuleti* *Saahithi Chippa*
UBID:50560530 *UBID: 50559752* *UBID: 50560583* *UBID: 50559876 smanjuna*
sarvetin *Mpasupul* *schippa*
smanjuna@buffalo.edu *sarvetin@buffalo.edu* *mpasupul@buffalo.edu* *schippa@buffalo.edu*

ABSTRACT:

Asthma and COPD are the current global respiratory diseases that affect lung function, quality of life, and further mortality rates. This project deep dives into the Exasens dataset in order to find commonly occurring patterns as well as the risk factors responsible for the development of respiratory diseases, concentrating on certain variables like age, the history of smoking and other physiological factors. This research study provided us with meaningful relationships that enhance prediction and treatment strategies by processing the data, exploratory analysis, clustering, and classification along with visualization. The findings will inform better health resource allocation and improve outcomes in respiratory health management.

INTRODUCTION

Respiratory diseases such as COPD, asthma, and respiratory infections are among the biggest concerns in healthcare systems globally, with millions of cases worldwide. While early and appropriate diagnosis is very important in the treatment of the disease, many of the conventional techniques rely on symptom assessment, medical history, and expensive or invasive tests. This can be highly inconvenient for the patients, time-consuming, and costly. Realizing these challenges, now the researchers are emphasizing developing new diagnostic techniques which should be less costly and fast.

The Exasens dataset used in this project is from UCI repository, which was created as part of a research collaboration between the Research Center Borstel and Biomat Bank Nord based in Germany. The permittivity biosensor involved is used to measure the dielectric properties of saliva specimens from patients who were diagnosed with COPD, asthma, respiratory infections, and healthy controls (HC). These dielectric properties of saliva, which have been determined to be correlated with various physiological conditions, include both real and imaginary components of the permittivity; hence, it is a good candidate for possible biomarkers that will result in respiratory disease classification.

In addition to the observed dielectric characteristics of saliva, this dataset contains demographic data including age, gender, smoking status, and diagnosis. This dataset gives the great opportunity to develop machine learning models that can classify individuals into one of four categories: COPD, asthma, infections, or healthy controls. It includes 399 occurrences and four main features. Furthermore, missing values introduce another level of complexity that must be addressed during data preprocessing.

Our project aims to use this dataset and investigate the possibility of using machine learning algorithms in order to provide a diagnosis of the respiratory conditions based on measures of saliva permittivity. We will explore a couple of classification methods, see how they all work, and evaluate how relevant each feature is for distinguishing between the groups. Ultimately, this project seeks to contribute to the discussion of non-invasive diagnostic tools that could lead to more efficient, affordable, and accessible healthcare solutions for individuals suffering from respiratory conditions.

DATASET:

The dataset used in this project is “EXASENS”, sourced from the UCI repository. It consists of six main columns, each with sub-columns as follows:

1. Diagnosis (COPD-HC-Asthma-Infected)
2. ID
3. Age
4. Gender (1= male, 0 = female)
5. Smoking Status (1 = Non-smoker, 2 = Ex-smoker, 3 = Active-smoker)
6. Saliva Permittivity
    a) Imaginary part (Min(I) = Absolute minimum value, Avg(I) = Average)
    b) Real part (Min(I) = Absolute minimum value, Avg(I) = Average)

The dataset was loaded using read.csv() with na.strings set to handle blank and "NA" values as missing. This made it easy to identify and process incomplete data entries.

DATA PREPROCESSING:

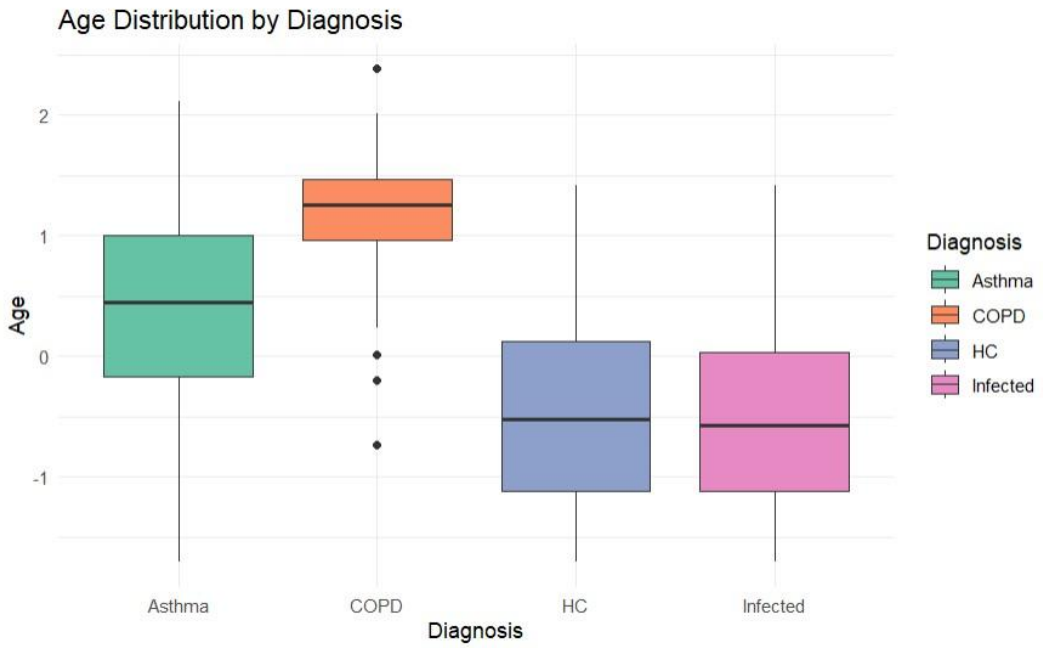
The process of transforming unstructured data into a structured format, ensuring that it is ready for analysis and modeling. This step is crucial because it removes inconsistencies in the data, fills in missing information, and arranges everything in a way that makes it easier to work with, especially for creating models and insights.

- **Handling Missing Values:** Our data had 1,196 missing values across a number of columns. The size of the dataset is so small that we aren't filtering out any data. We instead used matrix completion by softImpute method for missing values, where we created a low-rank approximation of the matrix and filled in the missing entries. The rank of the approximation was set to 3, then a regularization parameter was added ($\lambda = 0.1$) in order to avoid overfitting. This approach reduced noise and ensured that a cleaner dataset for modeling.
- **Dropping Unnecessary Columns:** Columns that did not add value to the analysis, such as those with only null values or irrelevant information, were removed from the dataset. This step helped reduce dimensionality and focus on essential features. In our dataset, the Gender, Smoking label and ID columns were removed since they are extraneous to our analysis. This approach ensured that we concentrated only on essential features.
- **Data Type Conversion:** Some columns were initially stored as character data, such as numeric-like values in the Imaginary_Part, and Real_Part columns. These were converted to appropriate numeric types.
    - Categorical columns, such as Diagnosis, Gender and Smoking, were converted to factors to treat them as categorical variables, aiding in more accurate encoding for analysis.
- **Scaling Numeric Features:** Columns Imaginary_Part_Min, Imaginary_Part_Avg, Real_Part_Min, Real_Part_Avg, and Age, were standardized using z-scaling. This step involved centering each column around a mean of 0 and scaling it to a standard deviation of 1. Standardization ensured that each feature contributed equally during model training and prevented features with larger ranges from dominating the analysis.
- **Feature Engineering:** Additional columns were created by combining existing features. New features, Imaginary_Difference and Real_Difference, represented the difference between

Imaginary_Part_Avg and Imaginary_Part_Min, and between Real_Part_Avg and Real_Part_Min, respectively. These features provided additional insights and added value to the dataset.

EXPLORATORY DATA ANALYSIS:

1. Distribution of Age by Diagnosis



Box plots Insight:

- Boxplots show the age profiles relative to the situation corresponding to the diagnosis.
- The box indicates the interquartile range IQR which indicates the 50 percent of data while the line within each of the boxes indicates median value.
- 'Whiskers the lower and upper quartiles and extend up to 1.5 * IQR from the lower quartile and upper quartile respectively. The points beyond this range are termed outliers, and are shown as dots.

Asthma:

- The optimal age depiction shows a fairly wider band although the IQR seems to show a betterbalanced ratio around median.
- The symmetrical distribution of the whiskers indicates a distribution which is slightly skewed.

COPD:

- The distribution of the age bracket among COPD patients is clustered into a section indicating a fairly unbroadened elderly pack.
- Median indicates that concentration is quite close to the upper quartile indicating a somewhat highest ages for concentration.

- Younger ages that are younger than the COPD average patients age essay some younger ages into outliers.

HC (Healthy Control):

- More regularly the HC group placed median within the IQR band which was more evenly distributed suggesting symmetry of the distribution.
- The Whiskers together with IQR which was wider pointed out the amount of heterogeneity in this case and so there is a wide span of age present in this group.

Infected:

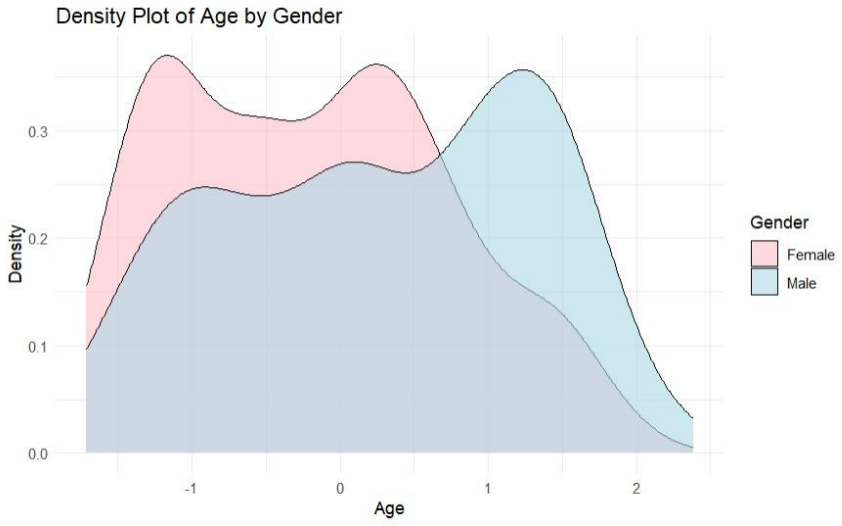
- This group shows wider variability in age than COPD but less than HC.
- The median is slightly below the IQR center, hinting at a minor skew towards younger ages.

Interpretation:

- Age distribution patterns differ across diagnoses, with COPD showing older, more clustered ages, while HC and Infected groups are more spread out.
- Asthma shows a moderate spread and younger age skew compared to COPD.
- Outliers in the COPD group represent younger individuals, less typical for this condition.

This analysis highlights how age factors differ across health conditions, which may be significant for understanding age-related impacts in these groups. The box plot effectively shows these differences in age distribution.

2. Density plot of Age by Gender



Density Plot Insight:

- The density plot estimates the distribution of age for male and female and thus age distribution within genders has been depicted in the form of a smooth curve.
- In the figure, the x-axis labels age, the y-axis measures density which indicates how many individuals are there for different ages within each gender population.
- Alpha levels are used to visually represent the overlying areas.

Relative Age Distribution:

- The pink shaded area represents females, and the light blue shaded area represents males.
- Areas of colour overlap illustrate areas of convergence or resemblance of age distributions of males and females.

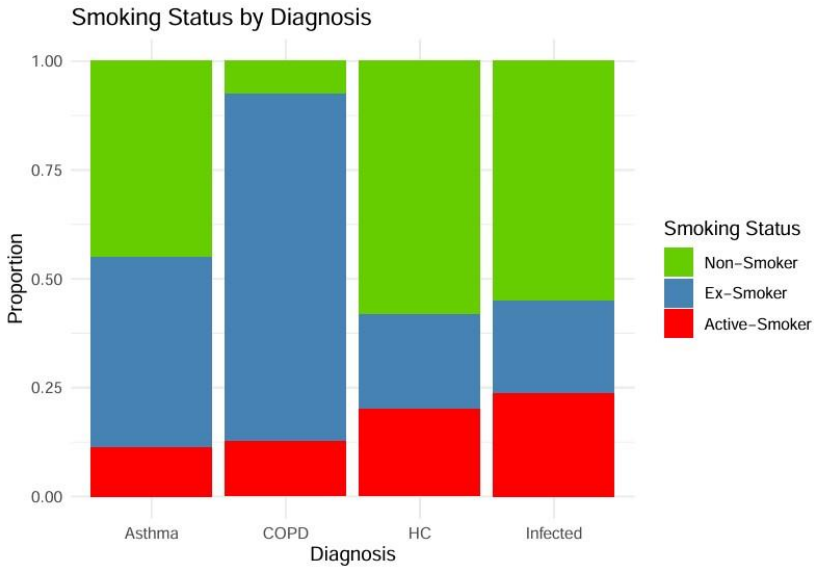
Key Findings:

- It can be noted that both males and females have several modes, meaning there are a number of age classes that appear most prominent in each gender.
- Males tend to have their age distribution relatively more diffuse across different age ranges as compared to females meaning that they have relatively more heterogeneity.
- Particularly, in the central region (around zero in the X-axis) there is a high degree of coincidence implying that the two genders comprise a high concentration of individuals residing in that region.
- The tail ends of each distribution highlight that extreme age ranges have relatively low representation for both males and females.

Interpretation:

This density plot provides a detailed view of how age varies between males and females. The shape and overlap indicate the distribution differences and similarities, offering insights into whether age patterns are consistent or diverse across genders within your dataset.

3. Smoking Status by Diagnosis



Stacked bar plot Insight:

This bar plot shows the proportion of smoking statuses (Non-Smoker, Ex-Smoker, and Active-Smoker) across different diagnosis categories: Asthma, COPD, HC (likely Healthy Controls), and Infected.
Smoking Status:

- Green represents Non-Smokers.
- Blue represents Ex-Smokers.

- Red represents Active-Smokers**.**

1. **Proportions by Diagnosis:**

    - Asthma: Non-smokers – the majority followed by Ex-Smokers' and few ActiveSmokers'.
    - COPD: The majority are Ex-Smokers, followed Non-Smokers with Active-Smokers being in a small percentage. This is evident as history of smoking is often seen in COPD patients.
    - HC (Healthy Controls): Non-Smokers have the highest proportion, with Ex-Smokers next and a small proportion of Active-Smokers.
    - Infected: A majority of Non-Smokers also dominated this group with Ex-smokers and Active smokers also following.

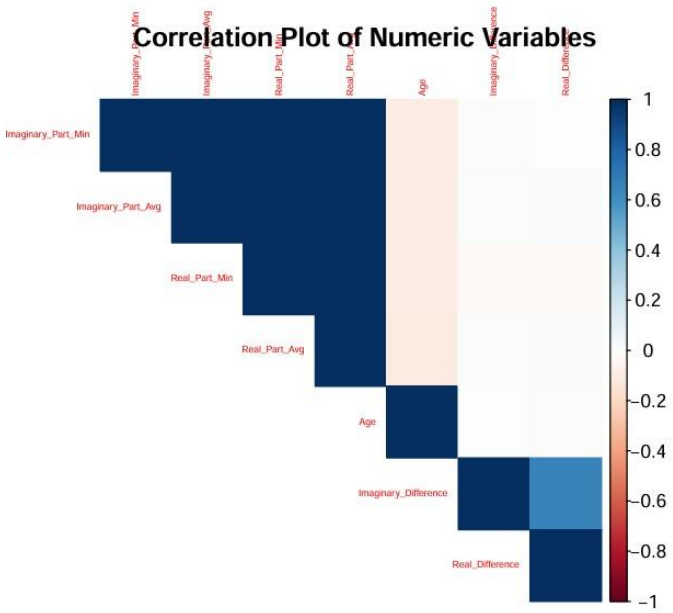
2. **Interpretation:**

    - Non-Smokers on the other hand seem to be common in the Asthma, HC and Infected groups.
    - For the COPD group, Ex-Smokers appear to have a high percentage which is understandable since smoking dominates as being a high risk factor for COPD in females.
    - Active-Smokers though lower in proportion in all categories are however most evident in the HC and Infected groups than in the other diagnoses.

3. **Purpose:**

    - This plot visually highlights the distribution of smoking statuses across different health conditions, which can be helpful for understanding the relationship between smoking habits and respiratory diseases.

4. **Correlation Plot for Numeric Variables**



Correlation heatmap insight:
Correlation heatmap of numeric variables, which visually represents the correlation coefficients between different pairs of variables in a dataset. Here's an explanation of the main elements:

Variables: Each variable is listed along both axes of the heatmap. In this case, the variables include

- Imaginary_Part_Min, Imaginary_Part_Avg, Imaginary_Difference
- Real_Part_Min, Real_Part_Avg, Real_Difference
- Age

Color Scale: The color bar on the right represents the correlation values, which range from -1 to 1.

- Positive Correlations (closer to 1) are shown in darker blue, indicating a strong positive linear relationship between two variables.
- Negative Correlations (closer to -1) are shown in red, indicating a strong negative linear relationship.
- No Correlation (around 0) is shown in white or light colors.

Diagonal Line: The diagonal from the top left to the bottom right shows a correlation of 1 (dark blue) for each variable with itself.

Off-Diagonal Values: These represent the correlation between different variables.

Attributes and Their Correlations:

1. **Imaginary_Part_Min**
    - Strongly **positively correlated** with **Imaginary_Part_Avg** and **Real_Part_Min** (dark blue cells), suggesting these variables increase together.
    - Weak or no correlation with **Imaginary_Difference** and **Real_Difference** (lighter cells), indicating these do not vary in a related way.
2. **Imaginary_Part_Avg**
    - Strong **positive correlation** with **Real_Part_Avg** and **Imaginary_Part_Min** (dark blue cells).
    - Very low or no correlation with **Imaginary_Difference** and **Real_Difference** (lighter colors), suggesting minimal association with these attributes.
3. **Real_Part_Min**
    - Strong **positive correlation** with **Imaginary_Part_Min** and **Real_Part_Avg** (dark blue cells).
    - Low or no correlation with **Imaginary_Difference** and **Real_Difference**, indicating these variables behave independently.
4. **Real_Part_Avg**
    - Strongly **positively correlated** with **Imaginary_Part_Avg** and **Real_Part_Min** (dark blue cells), meaning these variables tend to increase together.
    - Very weak or no correlation with **Imaginary_Difference** and **Real_Difference**.
5. **Age**
    - Generally shows **little to no correlation** with most other variables (light cells), implying it behaves independently of the others.

- Slight positive correlation with **Real_Part_Avg** and **Imaginary_Part_Avg**, though these are not very strong.

6. **Imaginary_Difference**

- **Negatively correlated** with **Real_Difference** (red cell), meaning as one difference increases, the other tends to decrease.
- Weak or no correlation with most other variables, which shows that this variable doesn't have a strong association with the minimum or average values of the real or imaginary parts.

7. **Real_Difference**

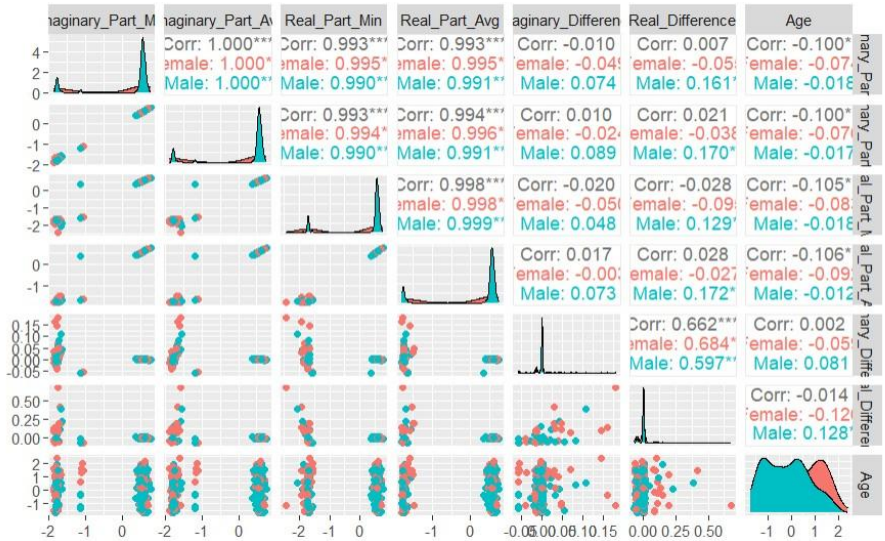
- Strong **negative correlation** with **Imaginary_Difference** (red cell).
- Shows very low or no correlation with other variables, indicating it varies independently of the minimum and average values in this dataset.

Purpose: This correlation plot helps identify relationships among numeric variables. Variables with high positive or negative correlations might have underlying relationships, while values near zero suggest weak or no linear association.

Summary:

- **Strong Positive Correlations** (dark blue): Between the "Min" and "Avg" attributes of real and imaginary parts (e.g., **Imaginary_Part_Min** with **Real_Part_Min**).
- **Strong Negative Correlations** (red): Between **Real_Difference** and **Imaginary_Difference**.
- **Weak or No Correlations** (light colors): Observed with **Age** and between difference variables and the min/avg attributes.

5. Pairplot



This image is a pair plot (scatter plot matrix) that provides insights into relationships between pairs of variables in a dataset. The pair plot displays the correlation coefficients and distributions for each pair of variables, with separate indicators for different groups (labeled as "Male" and "Female" in this plot).

Key Components

1. **Variables:**
    - The variables in the dataset include Imaginary_Part_Min, Imaginary_Part_Avg, Real_Part_Min, Real_Part_Avg, Imaginary_Difference, Real_Difference, and Age.
2. **Diagonals (Histograms):**
    - Along the diagonal, each variable's distribution is shown as a histogram. The distributions are split by gender (represented by different colors: red for females and blue for males), allowing us to see the distribution patterns within each group for each variable.
3. **Off-Diagonal Scatter Plots:**
    - The scatter plots show the relationships between each pair of variables, separated by color for males and females. This helps visualize the data points' spread and potential clustering by gender for different variable pairs.
    - Patterns in the scatter plots can reveal whether there is a linear, nonlinear, or no association between variables.
4. **Correlation Coefficients:**
    - The correlation coefficients are displayed in each off-diagonal cell. The values indicate the strength and direction of the linear relationships between pairs of variables, with:
        - Positive values indicating a positive correlation.
        - Negative values indicating a negative correlation.
    - Statistical significance is marked with asterisks: * ($p < 0.05$), ** ($p < 0.01$), etc., indicating the reliability of the correlation.
5. **Separate Correlations by Gender:**
    - The correlation values are provided separately for females and males, making it easy to compare the strength of relationships for each group.

Insights

1. **Strong Positive Correlations:**
    - Imaginary_Part_Min and Imaginary_Part_Avg have a very high positive correlation across all data (Corr: 1.000) and separately for males and females, suggesting a strong linear relationship between these two variables.
    - Real_Part_Min and Real_Part_Avg also have a very high positive correlation across both genders (Corr: 0.998), indicating that as one increases, the other does too.
2. **Weak or Negligible Correlations:**
    - Variables like Imaginary_Difference, Real_Difference, and Age generally show weak or no correlation with other variables. For instance, Imaginary_Part_Min and Imaginary_Difference have a very low correlation (Corr: -0.010).
    - Age also shows weak correlations with most variables, which could imply that it is largely independent in this dataset.

3. Differences Between Genders:

- In some variable pairs, there are notable differences in correlations between males and females:
  - For example, Real_Difference and Age have a higher positive correlation for males (Corr: 0.128*) than for females, where the correlation is nearly zero (Corr: -0.012).
- These gender differences might suggest that certain relationships hold more strongly within one gender than the other.

4. Distributions:

- The histograms along the diagonal give insight into how each variable is distributed within each gender.
- For example, in the Age histogram, the blue (male) distribution has a slight shift compared to the red (female) distribution, indicating that the age distribution varies between genders.

Summary:

This pair plot provides a comprehensive view of:

- How variables are linearly related to each other.
- Differences in correlation strengths between genders.
- The distribution of each variable within each gender.

This visualization is useful for identifying potential patterns, dependencies, and gender-based differences in the data.

PCA

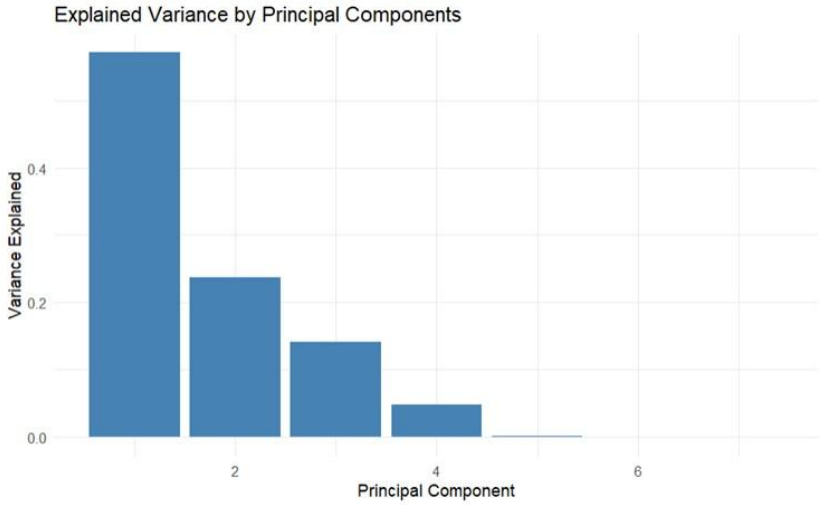
1. Importance of Components Table

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5       PC6       PC7
Standard deviation     1.9998 1.2897 0.9932 0.5821 0.11048 7.481e-16 8.658e-17
Proportion of Variance 0.5713 0.2376 0.1409 0.0484 0.00174 0.000e+00 0.000e+00
Cumulative Proportion  0.5713 0.8089 0.9498 0.9983 1.00000 1.000e+00 1.000e+00
```

Standard Deviation:

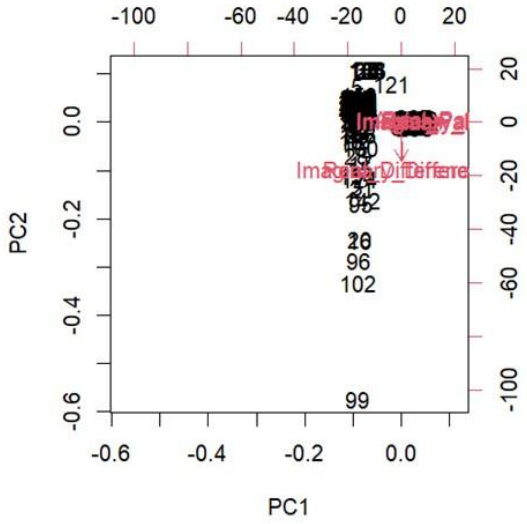
- This parameter indicates the amount of variation captured by each principal component. For example, PC1, with the largest standard deviation, captures the most variation in the data, while PC7 captures almost none. PC2 accounts for nearly one-third of the total variance, focusing primarily on essential elements that contribute significantly to the dataset's structure. Proportion of Variance:
- This row demonstrates the percentage of the total variance explained by each component. PC1 explains 57.13% of the variance, followed by PC2 with 23.76%, and so on. The last three components (PC5, PC6, and PC7) contribute almost no variance, suggesting they are not essential for understanding the data structure. Cumulative Proportion:
- This measure accumulates the variance explained up to each principal component. By the fourth component (PC4), approximately 99.83% of the variance is accounted for, meaning the first four components capture nearly all the information embedded in the data.

2. Variance Bar Plot



- This bar plot provides a visual representation of the variance explained by each principal component.
- PC1 has the highest explanatory power, followed by PC2, PC3, and PC4. After PC4, there is a steep drop-off in variance explained, indicating that components beyond PC4 contribute minimal new information. This suggests that the first four components are sufficient for capturing the key structure in the data.
- This plot suggests that the first few components (primarily the first four) capture most of the meaningful variation in the data, making them the most informative for dimensionality reduction.

3. PC1 vs. PC2 Scatter Plot:

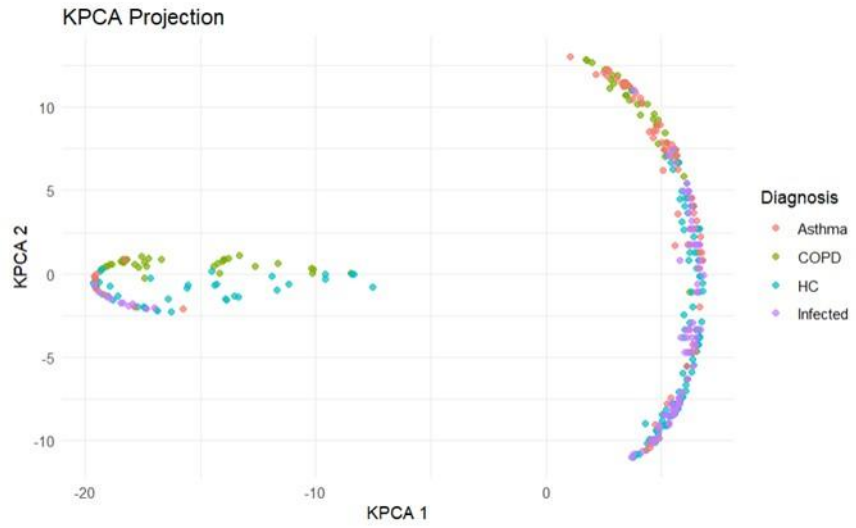


- This scatter plot represents the data projected onto the first two principal components (PC1 and PC2).

- Each point corresponds to an observation in the dataset, now represented in terms of PC1 and PC2 values. This type of plot is commonly used to visualize clustering or spread in data along the two most significant axes of variance.
- The labels seem cluttered, making it challenging to distinguish individual points, but overall, it shows that the data points are concentrated within a small range along both PC1 and PC2 axes, suggesting that they capture the primary variation in the data.

In summary, this PCA analysis indicates that the first four components capture nearly all the variance, and thus, the dataset could be effectively reduced to four dimensions without losing significant information. The scatter plot further illustrates the distribution of data along the first two dimensions, helping in visualizing potential clustering or patterns.

KPCA



This plot shows the result of Kernel Principal Component Analysis (KPCA) on a dataset where the data points are projected onto two principal components (KPCA 1 and KPCA 2). KPCA is a non-linear dimensionality reduction technique, often used to capture complex structures in data that cannot be represented well with linear transformations (like in traditional PCA).

1. **Axes (KPCA 1 and KPCA 2)**: These are the two principal components derived from the kernel transformation. They represent directions in the transformed feature space that capture the most variance in the data in a non-linear manner.
2. **Data Points and Colors**: Each point represents an individual data sample, with colors indicating different diagnosis groups:
    - Asthma (pink)
    - COPD (green)
    - HC (Healthy Controls) (blue)
    - Infected (purple)
3. **Clustering Pattern**: The points appear grouped in clusters, suggesting that KPCA has managed to separate the data samples based on diagnosis types to some extent.
    - The elongated arc of points stretching along the KPCA 1 axis may indicate that this component captures a prominent separation in the data, particularly between HC and other groups.

- The concentration of points to the left could imply some diagnostic similarity between certain groups.

In summary, this KPCA plot shows that different diagnosis types (Asthma, COPD, HC, Infected) can be partially separated in the KPCA-transformed feature space. This could imply that there are distinct underlying patterns in the data, which KPCA helps to reveal.

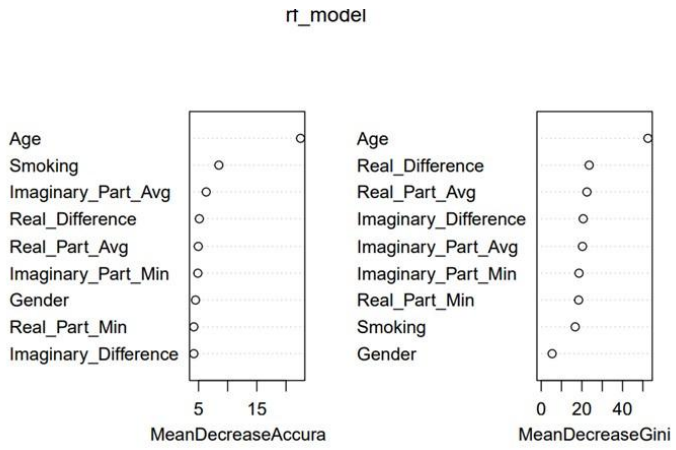
MODEL TRAINING:

The goal of this project is to deploy multiple machine learning models to classify medical data into discrete diagnostic categories. Each model's training process and contribution to the classification problem are described in detail in the sections that follow.

1. Random Forest

Because of its resilience and capacity to manage high-dimensional data and feature interactions, the Random Forest model was selected. It works by constructing a group of decision trees, each of which has been trained using a bootstrapped dataset sample. Three characteristics (mtry = 3) were taken into account at each split of the 100 trees employed in this research.

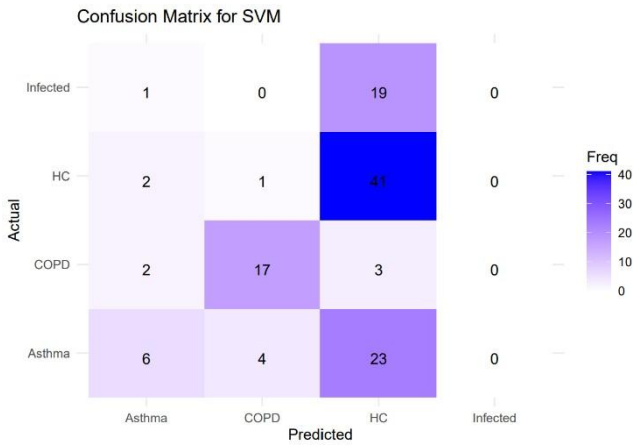
The model uses the diversity of trees to enhance generalization as it learns patterns that differentiate between diagnostic categories during training. A confusion matrix, which displayed the model's classification accuracy throughout the test data, was used to assess its performance after training. As seen in the figure below, feature importance plots identified important factors that are essential for differentiating diagnostic outcomes, including age, smoking, and imaginary difference.



2. Support Vector Machine (SVM)

A linear kernel, which is perfect for issues where data is linearly separable in the feature space, was used to train the SVM model. The trade-off between eliminating classification mistakes and maximizing the margin was managed by the regularization parameter (cost = 1).

SVM is especially well-suited for this task because it can generate a hyperplane to divide several diagnostic classifications. As seen in the picture below, the model was assessed by predicting test data labels and contrasting them with actual values using a confusion matrix.



The findings demonstrated that SVM effectively caught the connections between diagnostic labels and characteristics, particularly when class borders were evident.

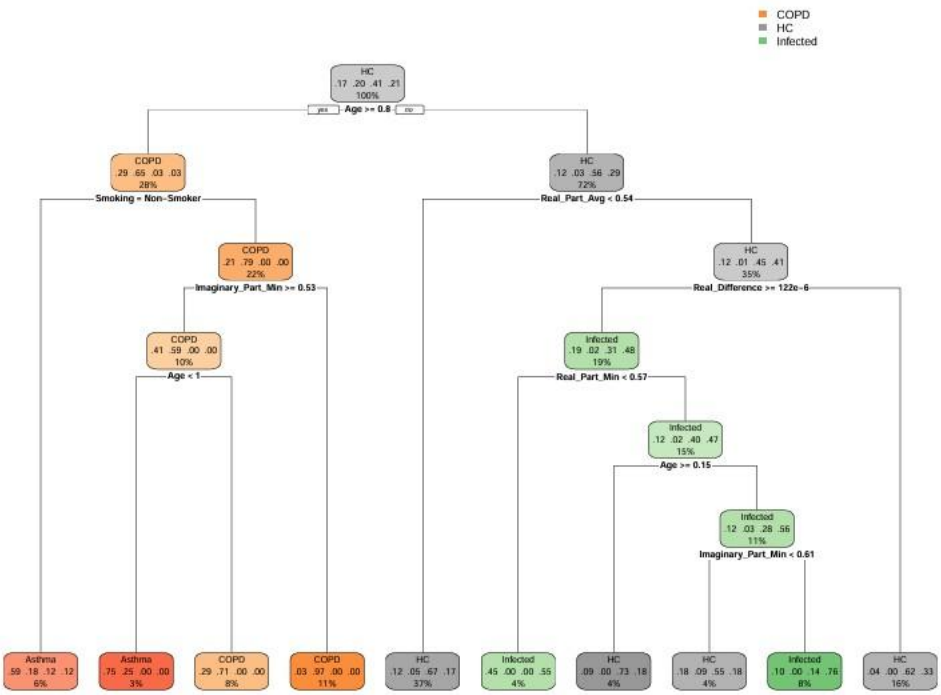
3. Multinomial Logistic Regression

This model is an extension of logistic regression for multi-class classification, and it is implemented using the nnet package. By representing the log-odds as a linear mixture of predictors, it calculates the likelihood that each observation falls into one of the diagnostic categories.

In order to reduce residual deviation, iterative optimization was used to modify the parameters after the model was fitted to the training data. A confusion matrix was used to evaluate the performance, demonstrating the model's capacity to identify properly in the majority of instances. The model's useful baseline performance, despite its simplicity, demonstrated how well the diagnostic classes are explained by linear connections between predictors and outcomes.

4. Decision Tree

The dataset is recursively divided by this model according to feature values that optimize class separation at every node. Features like Age, Real_Part_Min, and Imaginary_Difference emerged as crucial split points in the decision-making process, which was facilitated by its clear and understandable structure. This model's decision tree is displayed below,



Plotting the decision tree's structure as seen above allowed for evaluation. This model is a clear tool for medical practitioners, especially when it comes to comprehending the principles underlying classification.

5. KMeans Clustering

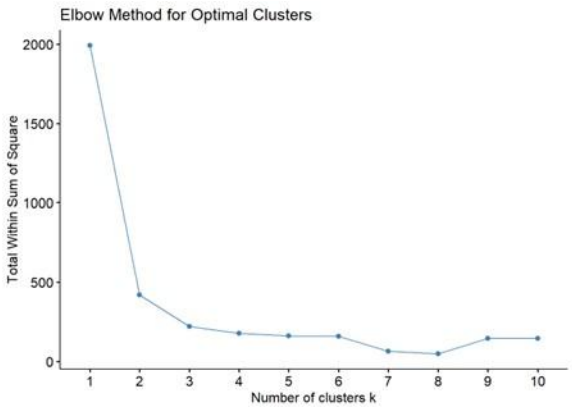
While primarily a classification task, unsupervised learning through KMeans Clustering was also applied to explore inherent patterns in the data.

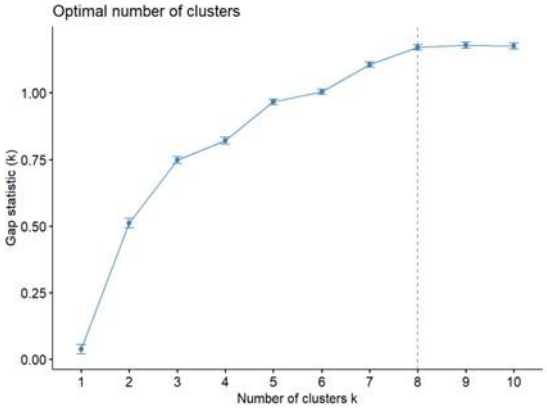
Determining the Optimal Number of Clusters:

To ensure effective clustering, it was crucial to determine the optimal number of clusters (k). Two widely used methods were employed:

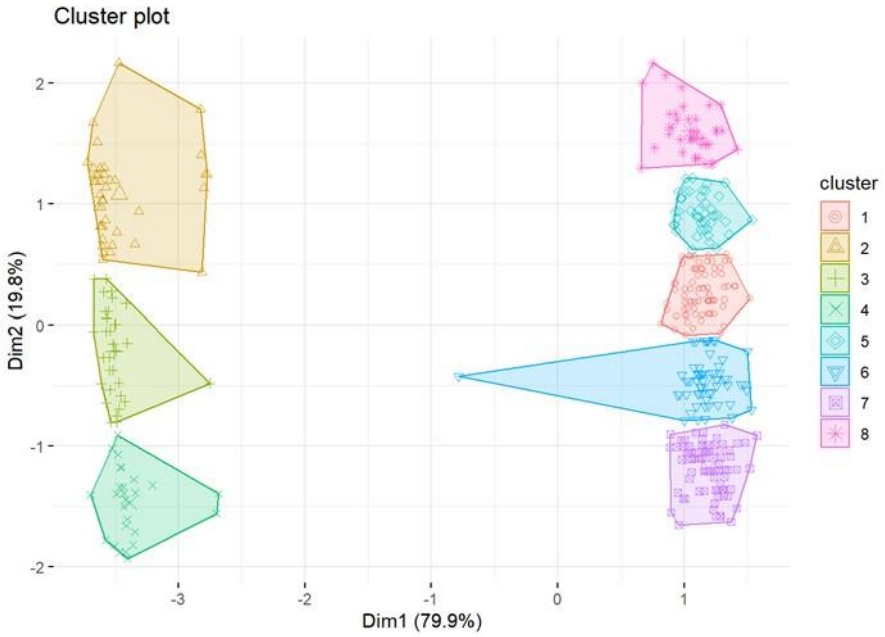
- Elbow Method: This technique involves plotting the total within-cluster sum of squares (WCSS) against different values of k. Since the data points will be closer to the cluster centroids, when the number of clusters increases, WCSS decreases. The point corresponding to the "elbow," where the rate of decrease begins to slow down significantly, is the choice for k.

- Gap Statistics: This method considers the WCSS of the real data versus that of a reference data set generated under the null hypothesis-a uniform distribution. Therefore, the optimal k is that at which the gap statistic is the highest. The above indicates that, using the chosen number of clusters, one gets the best separation as compared to random distributions.



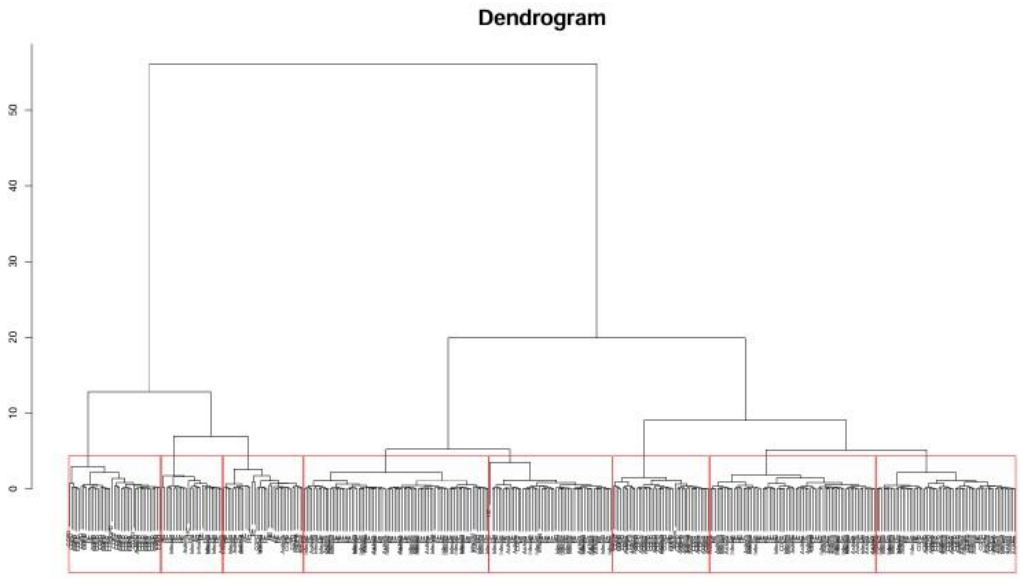


Both methods pointed to 8 as the best number of clusters. The model was then visualized as shown in the figure below, to analyze how well these clusters aligned with the actual diagnostic labels. Although clustering itself doesn't use labels directly, it can offer a way to check class separability and spot some subgroups in the data.



6. Hierarchical Clustering

Hierarchical clustering is a way of understanding the intrinsic structure of the data by grouping similar data points without using the labels. A distance matrix was computed on the numeric features using Ward's method, which, at each step of the clustering process, minimized the within-cluster variance. A dendrogram is a tree-like diagram showing how data points merged into clusters at different levels of similarity. This allowed one to intuitively understand the hierarchical relationships among the data points.



This dendrogram, therefore, revealed the presence of eight optimal clusters same as that provided by both Elbow Method and Gap Statistics. Highlighted in the original image in red boxes, eight clear separations that may correspond to diagnostic categories were evident in the dendrogram. By noting at which heights clusters merged, the dendrogram provided insights into the extent of the similarity among data points, giving a complementary outlook to KMeans clustering and, at the same time, validating the separability of classes within the dataset.

EVALUATION AND INSIGHTS:

Classification Models Evaluation:

Each model's performance is measured based on accuracy and Mean Squared Error (MSE). The results are as follows:

Model	Mean Squared Error	Accuracy
Random Forest	46.22%	53.78%
Support Vector Machine	45.4%	54.6%
Multinomial Logistic Regression	50.4%	49.6%
Decision Tree	50.4%	49.6%

The SVM model achieved the highest accuracy at 54.6%, with the lowest MSE among the models. This indicates that SVM outperformed the other models in classifying the diagnostic labels.

Each model's confusion matrix is also used to evaluate performance in terms of precision, recall, and overall accuracy. Models like Random Forest and SVM delivered the best performance, showing high classification accuracy for all diagnostic categories. Decision Trees and Logistic Regression, while slightly less accurate, offered interpretability and insights into feature importance. The combination of these models allowed for a robust analysis, providing both predictive accuracy and interpretability, which are crucial in a healthcare context for supporting diagnostic decision-making.

These results reflect relatively modest accuracy across all models, most likely due to the kind of complexity and the overlapping nature of the diagnostic categories in the dataset.

K-Means Clustering Evaluation

The within-cluster sum of squares (WSS) and the average silhouette width is evaluated for K-means Clustering performed on the dataset and the results are as follows:

- Within-Cluster Sum of Squares (WSS): 43.9189
- Average Silhouette Width: 0.5045

WSS measures the compactness of the clusters with lower values indicating more compact clusters. The score of 43.9 for this model suggests that the clustering model created reasonably compact clusters.

Silhouette width gives an idea about the separation between the samples in the case of clusters. Values closer to 1 indicates the samples fall into well-defined clusters. The average silhouette width of 0.5045 suggests that the clusters are moderately well-defined but might have some overlap.

To evaluate accuracy of clustering, we have mapped each cluster to its most common true label, converting cluster labels to match the actual diagnostic categories. The confusion matrix and accuracy score indicate how well K-means clustering aligned with actual diagnosis labels:

- Cluster-to-Label Accuracy: **56.1%**

The accuracy of 56.1% is slightly better than the supervised models' performance, suggesting that the clustering model was able to capture some underlying patterns in the data without the guidance of labeled examples. However, this accuracy is based on the frequency of the labels in the respective cluster, this could not be completely efficient.

The diagnostic labels in the clusters reveal notable groupings, as displayed in the cluster-to-diagnosis table:

Cluster	Asthma	COPD	HC	Infected
1	19	1	37	20
2	1	36	3	0
3	7	4	18	3
4	2	0	18	7
5	18	16	6	4
6	7	2	34	18
7	9	0	43	27
8	17	20	1	1

Clusters generally correspond to one or more dominant diagnoses, with some clusters clearly associated with specific conditions. However, overlap across clusters indicates that not all diagnoses are distinct enough to be easily separable, highlighting the complexity of the dataset.

CONCLUSIONS

We have constructed four classifiers and applied K-means clustering on the data. SVM emerged as the most effective classifier for predicting diagnosis, likely due to its ability to capture non-linear patterns, followed by Random Forest. Logistic regression and decision trees offered lower but still informative accuracies.

PCA and KPCA showed that a few principal components effectively summarized data, hence justifying the use of dimensionality reduction for visualization and simplification of models.

K-means clusters could approximately be mapped into the diagnostic labels. Although the accuracy was much higher than that of the classification models, it indicated that some underlying pattern could be learnt by the clustering; however, it is an approximate estimation.

REFERENCES:

1. *Exasens [Dataset]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5M03M.*

2. *Pouya Soltani Zarrin, Niels Roeckendorf , Christian Wenger , May 28, 2020, "Exasens: a novel dataset for the classification of saliva samples of COPD patients", IEEE Dataport, doi: https://dx.doi.org/10.21227/7t0z-pd65.*

3. *https://www.sciencedirect.com/science/article/pii/S0010482523007606.*