# **Business Machine Learning Project.**

**Group 4:**

**Team Members:**

Raja Sumanth Reddy Arram​

Vamsidhar Reddy Jannapally​

Shresta Chapala​  
Lakshmi Sumith Kumar Keerthipati​

**Introduction:**

​  
Asthma, a chronic respiratory disease, poses a significant global health challenge, affecting millions of individuals of diverse age groups worldwide. Despite its prevalence, the healthcare system currently lacks robust data models for predicting this common yet serious ailment. The absence of comprehensive patient data, coupled with a dearth of supportive resources for handling and presenting this data in a usable format, hinders the ability of medical professionals to detect asthma in its early stages or, ideally, predict its occurrence.

**Overview:**

The integration of technological advances, specifically through the application of machine learning within the healthcare domain, presents a transformative opportunity to address this challenge. By leveraging the power of machine learning, we aim to enhance early detection and prediction capabilities for asthma. This proactive approach enables hospitals, doctors, and patients to anticipate asthma attacks, fostering better emergency preparedness. The ultimate goal is to reduce emergency room visits, enhance overall patient well-being, and contribute to long-term health outcomes. Additionally, this innovative approach facilitates the tailoring of patient-focused treatment plans, ensuring a more personalized and effective healthcare strategy for individuals at risk of or already affected by asthma. Through the intersection of healthcare and technology, we aspire to create a paradigm shift in asthma management, ushering in an era of predictive and preventative healthcare measures that significantly improve patient outcomes and quality of life.

**Data Source:**

The project's foundation is a comprehensive dataset from Kaggle, consisting of over 639,590 patient records from the United States, spanning the years 2020 to 2022. The dataset features variables detailing symptoms, age groups, and treatment outcomes, providing a rich tapestry of data for analysis. The target variable for our predictive models is the presence of asthma, identified through a combination of health indicators and reported symptoms.

<https://www.kaggle.com/datasets/deepayanthakur/asthma-disease-prediction?resource=download>

**Data Description**

Our primary dataset from Kaggle comprises over 639,590 patient records, offering a substantial foundation for our analysis. This data set provided on Kaggle is a collection of health records and patient data curated for research and modelling purposes. This data contains the

* Symptoms
* Age groups
* Severity and treat outcomes of the patients which provides us with the valuable insights across all age.
* Severities and helps us in a great predictive analysis.
* Enhancing early detection, management, and overall efficient patient care.

Metadata: Presence of asthma, severity, and treatment outcomes.

**Data Preprocessing and Methodology:**

Data preprocessing involved rigorous cleansing, including the removal of null values and the reordering of data, formatted for compatibility with Tableau. Additional transformations were applied for analysis suitability, which will be outlined in this section, along with the methodologies applied for both Tableau visualization and machine learning model preparation.

**Descriptive Analysis: Tableau Visualization**

The decision tree model is a fundamental component of our descriptive analysis, providing an intuitive visualization of the data's underlying structure in predicting asthma. The Tableau platform enhances our ability to interpret complex model outputs by presenting them in an accessible graphical format.

**Variable Importance and Decision Trees:**

Within the decision tree, the variables 'Race' and 'General Health' are presented as critical in predicting asthma, as evidenced by their position at the top of the tree. This prominence within the tree structure indicates their strong influence on the model's outcomes. The Youden index value of 0.2015, associated with the decision tree, reveals a moderate diagnostic ability, suggesting the model has a reasonable capacity to differentiate between those with and without the condition.

The accompanying variable importance chart provides a quantitative visualization that complements the qualitative insights from the tree. Here, 'Race' is indicated as the most significant predictor, followed by 'General Health' and 'Dry Cough'. The length of the bars represents the relative importance of these variables, with longer bars indicating greater influence on the model's predictions.

**Decision Tree Visualization:**

The decision tree itself is depicted as a flowchart-like diagram, illustrating the sequence of binary decisions made based on the input variables that lead to a prediction. The root of the tree begins with 'Race', splitting the population into subsets, which are then further divided by 'General Health' and 'Dry Cough' at subsequent levels. These splits represent the decision rules derived from the data.

**Interpreting the Tree and Variable Importance:**

The tree's structure allows us to interpret the pathways of prediction and the interactions between variables. For instance, a particular path in the tree may show that individuals of a certain race and with specific general health status are more likely to have asthma, with further refinement provided by the presence or absence of a dry cough.

The variable importance chart informs us of which factors to prioritize in asthma screening and management. The prominence of race and general health in the model suggests potential areas for targeted interventions, public health messaging, and resource allocation.

The decision tree and variable importance visualization in Tableau serves as a powerful exploratory tool, uncovering the relationships between patient characteristics and the likelihood of asthma. This analysis aids healthcare professionals in understanding the complex interplay of demographic and health-related factors that contribute to asthma risk, facilitating more informed decision-making in clinical practice.

* **Decision tree:**

A screenshot of a computer

Description automatically generated

*Fig description: In the decision tree model the lift coincides with the model at 70% which means, both observed and expected were predicting the required values correctly from 70 percentile.*

**Variable Importance:**

Using Random Forest Regression, we See that ‘Runny Nose’ is the most important variable in early detection of Asthma.

A screenshot of a computer screen

Description automatically generated

* **Predictive Modeling:**

We employed several machines learning models, including Logistic Regression, Decision Tree, and Random Forest Regression, to analyze the dataset. Each model offered unique insights — Logistic Regression helped in understanding the linear relationships, while Decision Tree and Random Forest provided more nuanced views of complex interactions within the data. Through rigorous testing and validation, we assessed the effectiveness of these models in predicting asthma cases, with particular attention to their accuracy and reliability.

**A screenshot of a computer

Description automatically generated**

* **Key Findings from Visualizations:**

Bar charts from Tableau reveal that symptoms like sore throat and difficulty in breathing have varying prevalence across age categories, with older age groups experiencing higher incidences. The impact of demographic factors like age, sex, and race on asthma symptoms is also evident, showcasing the potential for tailored preventive care.

**Based on Soar Throat:**

**A screenshot of a graph

Description automatically generated**

It displays the distribution of a count metric related to sore throats across different age categories.

From left to right, the x-axis represents age groups ranging from 18-24 years to 80 years or older. The y-axis represents the count of sore throat cases recorded within each age group. The bars are colored to indicate whether the count of sore throats within each age category is high or low, but without a key or legend. The presence of "Yes" labels on bars from the "55-59" age group and older suggests that there may be a significant count or a threshold value being exceeded in these age groups.

**Affect of Dry Cough and Difficulty breath:**A screenshot of a computer

Description automatically generated

The bar chart illustrates the prevalence of dry cough and difficulty breathing across various age groups. Each age category displays two bars: orange for individuals with both symptoms and blue for those with only breathing difficulties. The data indicates that breathing difficulties are more common without a dry cough across all ages. Notably, the "60-64" age group exhibits the highest incidence of both conditions. This visualization underscores the importance of these symptoms in respiratory health monitoring.

**Impact of Age, Breathe, Sleep Time on Asthma:**

Fig titled "Impact of Age, Breathe, Sleep Time on Asthma". It shows data for different age categories on the x-axis, with the count of a certain variable (presumably relating to breathing difficulties) on the y-axis. The bars represent the number of observations or instances within each age category that correspond to the variable of interest, which seems to be related to symptoms or factors associated with asthma.

Each bar's height corresponds to the count within each age group, and there is a label at the top of each bar indicating the count value. The age categories range from 18-24 years old to 80 or older. The chart is designed to provide insights into how age, breathing difficulty, and possibly sleep time are related to asthma cases or symptoms within a dataset.

A screenshot of a computer

Description automatically generated

**Impact Across Age, Sex and Race:**

The treemap displayed here offers a stratified perspective on the distribution of asthma based on demographic criteria like age, gender, and ethnicity. Each section within the visual represents a specific subgroup, with its size indicating the subgroup's relative frequency or impact in terms of asthma.

Key observations from the visualization include:

Predominantly, the most prominent sections are those representing older White females, particularly in the 60-74 age range, indicating a higher frequency of asthma or a greater significance in terms of healthcare engagement within this demographic.

In contrast, the smaller sections for the younger demographics, notably the 18-24 and 25-29 age brackets, suggest a lesser frequency or impact.

The data points to a potential trend where asthma's impact increases with age, with a marked prominence among females. Nonetheless, there's also a noticeable presence of asthma impact among older males, especially in the 65-74 age category.

If color variations are present, they may denote different layers of data, such as intensity of asthma conditions or levels of healthcare usage. Darker hues could indicate more intense asthma episodes or a higher use of healthcare services.

The treemap thus serves as an efficient means to encapsulate and communicate multifaceted data, highlighting demographic patterns and potential areas of focus for healthcare planning and asthma intervention measures.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Model Evaluation:**

The Random Forest model, as depicted in the screenshots, illustrates ‘Runny Nose’ as the most influential variable in the early detection of asthma, followed by ‘Physical Activity’ and ‘heart disease’. The model’s Average Squared Error (ASE) and the lift chart indicate that the model’s predictions align closely with the observed data at the 70th percentile, signifying a reliable predictive capability.

The Logistic Regression Model focused on the presence of ‘Pains’ as an event. Showcasing a Youden’s index of 0.2008, which suggests moderate discriminative power. The confusion matrix within this model emphasizes the low frequency of true positives (events correctly predicted as ‘Yes’), which indicates room for improvement in model sensitivity.

**Future Work:**

Data Collection Expansion: There is a call for broader data collection, potentially incorporating genetic and environmental variables, which could lead to more comprehensive predictive models.

Enhanced Data Quality and Monitoring: The project suggests that improving data quality and implementing real-time monitoring systems could significantly improve asthma management and patient care, indicating a direction towards more personalized and responsive healthcare solutions.

These findings underscore the potential of machine learning in transforming asthma diagnosis and management, paving the way for more individualized treatment plans and proactive healthcare strategies.

**Lessons Learned and Conclusion**

The analysis conducted through the Random Forest model revealed a high degree of predictive accuracy, with symptoms such as 'Runny Nose' identified as key predictors in the early detection of asthma. This finding underscores the potential of using symptomatology in predictive health analytics. However, while the Logistic Regression model provided valuable insights, the Youden index and the confusion matrix pointed towards moderate predictive capabilities, suggesting a need to refine model sensitivity further. The project’s findings advocate for the expansion of data collection to encompass genetic and environmental factors. Incorporating such multifaceted data promises to enhance the precision of predictive models and, by extension, the personalization of asthma management strategies. The envisaged integration of real-time monitoring systems aligns with the evolving paradigm of healthcare, which increasingly favors individualized patient care and responsive treatment plans. Our investigation has also highlighted the significant influence of age on asthma prevalence. The data indicates a higher likelihood of asthma with advancing age, signaling the importance of age as a factor in predictive health models. Furthermore, the direct association of symptoms such as sore throat and difficulty in breathing with the early prediction of asthma presents healthcare practitioners with actionable markers for early screening and intervention. As we look towards the future of healthcare, the insights gained from this project emphasize the transformative impact of machine learning in the early detection and management of asthma. By leveraging these technologies, we move closer to a healthcare system characterized by foresight, efficiency, and a deepened capacity for enhancing patient well-being.