**Title: Asthma Prediction Using Machine Learning**

Team Members: PREETHI MATHARI, SOWMYA ILA and SRILEKHA

Affiliation: Department of Applied Computing, Michigan Technological University

EET 4501 Applied Machine learning

**GitHub Link:** [**https://github.com/Mpreethi20/ML-Asthma-prediction-project/tree/main**](https://github.com/Mpreethi20/ML-Asthma-prediction-project/tree/main)

**Abstract**

This project explores the application of machine learning techniques to predict the severity of asthma attacks using a dataset sourced from Kaggle. Motivated by the need for early and accurate asthma severity prediction to improve patient outcomes, we employed several classifiers, including Random Forest, Gradient Boosting, and K-Nearest Neighbors. The results demonstrate a promising direction for using predictive analytics in healthcare, with our models achieving an accuracy of around 75%, highlighting the potential to enhance proactive healthcare interventions.

**Introduction**

Importance of Asthma Prediction

The capacity to foresee the severity of asthma attacks holds profound implications for healthcare, potentially enhancing patient care and optimizing resource allocation. Accurate predictions could enable preemptive medical interventions, significantly improving patient management and reducing emergency healthcare situations.

Motivation for the Study

The driving force behind this project is the transformative potential of machine learning in chronic disease management, particularly asthma. The integration of predictive analytics into healthcare practices for asthma can facilitate timely and targeted interventions, thus mitigating severe asthma episodes and enhancing patient health outcomes.

Machine Learning's Role in Healthcare

Machine learning has reshaped many sectors, including healthcare, by extracting meaningful insights from complex datasets. For this project, the following machine learning techniques are pivotal:

Neural Networks: These have been widely used for complex pattern recognition tasks in medical imaging and diagnostics, helping identify subtle patterns in patient data that may indicate disease states.

Decision Trees and Ensemble Methods (such as Random Forests): These methods are crucial for their robust performance in clinical decision-making tools, where they classify patients based on clinical symptoms and histories.

Gradient Boosting Machines: Employed for their high level of accuracy in predictive tasks, these models are adept at handling varied datasets typical in medical research.

K-Nearest Neighbour’s (KNN): This method, known for its simplicity, is used to classify patients based on the similarity of their cases to previously known examples.

Survey of Related Work

Various studies have utilized machine learning to enhance asthma management, focusing on predicting exacerbation from clinical and environmental data. Such research underscores the capabilities and limitations of different models—ranging from the precision and resource-intensiveness of deep learning models to the simplicity and sometimes insufficient sensitivity of models like KNN.

**Dataset**

The dataset was sourced from Kaggle and is designed to predict the severity of asthma based on symptoms and demographics.

It comprises approximately 63,000 instances, each representing individual patient data.

Features include symptoms such as Tiredness, Dry-Cough, Difficulty-in-Breathing, and demographics like age and gender.

Training/Validation/Test Split:

The data was divided into training and testing sets with an 80/20 split. Specifically, 80% of the data was used for training the models, and the remaining 20% was reserved for testing the effectiveness of the predictions.

Preprocessing Steps:

Handling Missing Values: The dataset was checked for missing values, which were then imputed to maintain the integrity of the dataset. Numerical missing values were filled using the mean of the respective column, while categorical missing values could be filled using the mode if applicable.

Encoding Categorical Variables: Categorical features such as symptoms were encoded to transform them into a format that could be efficiently processed by machine learning algorithms.

Data Normalization:

Standardization techniques were applied to the data to ensure that the input features have mean zero and standard deviation of one. This normalization is crucial for models that are sensitive to the scale of input data, such as K-Nearest Neighbors and Gradient Boosting.

Size of Input Data:

The input data size for each instance comprises various features, primarily categorical, indicating the presence or absence of symptoms and demographic characteristics. There is no typical 'image size' as the data is structured.

Data Features and Visualization:

Features Used: The dataset features include various symptoms related to asthma, age groups, and gender. Symptoms like 'Tiredness', 'Dry-Cough', and 'Difficulty-in-Breathing' are binary (1 for presence, 0 for absence).

Visualizations: Visual aids like histograms and pie charts were used to understand the distribution of features like age and gender. For example, pie charts were used to display the proportion of data belonging to different age groups and genders.

Correlation Analysis: A heatmap was used to analyze the correlation between different features, helping in understanding how different symptoms and demographics are related to the severity of asthma.

**Methods**

This project utilized three main machine learning algorithms:

Random Forest Classifier

Random Forest is an ensemble learning method renowned for its accuracy and robustness. It consists of a multitude of decision trees that operate as a collective. Each tree in the ensemble is built from a sample drawn with replacement (bootstrap sample) from the training data. When constructing the trees, the best split for each node is selected from a random subset of the features, enhancing diversity among the trees and consequently reducing the risk of overfitting. The final prediction is derived by aggregating (majority voting) the predictions from all trees, ensuring that the most common outcome among the individual trees is chosen as the final class for each input sample.

Gradient Boosting Classifier:

Gradient Boosting is a sophisticated boosting technique that builds an ensemble of models incrementally. It constructs a series of weak prediction models, typically decision trees, each correcting its predecessor, thereby improving the model's accuracy with each step. New models are added to compensate for the errors made by earlier ones, and the predictions are made based on a weighted sum of these models. This process involves optimizing a differentiable loss function, which quantitatively assesses how far the ensemble's predictions deviate from the actual data, with each new tree aiming to reduce this residual as much as possible.

K-Nearest Neighbors (KNN) is a simple yet effective classification method that predicts the class of a given point based on the classes of the nearest points in the feature space. It operates under the assumption that similar instances tend to cluster together. Thus, a new instance is classified by a plurality vote of its neighbors, with the new instance being assigned the class most common among its nearest.

**Experimental Setup and Hyperparameter Tuning**

In this project, the dataset was partitioned into training (80%) and testing (20%) sets to evaluate the performance of various predictive models. We utilized GridSearchCV for rigorous hyperparameter tuning, which systematically explores a range of configurations and applies cross-validation to prevent overfitting. This process was crucial for determining the optimal settings for each model:

Random Forest: Tuning focused on the number of decision trees (n\_estimators) and the depth of each tree (max\_depth), testing ranges from 100 to 200 trees and allowing for various tree depths from shallow to deep (unlimited).

Gradient Boosting: Adjustments included the number of stages (n\_estimators), learning rate adjustments (learning\_rate), and tree depths (max\_depth), with specific attention to how these parameters influence model complexity and learning speed.

Nearest Neighbors (KNN): We explored different values for the number of neighbors (k) and weighting strategies (weights), determining how each configuration affects the prediction accuracy and generalization.

These models underwent 5-fold cross-validation, which splits the training dataset into five subsets, using each in turn for validation while training on the remaining four.

Results and Analysis

Model evaluations were based on accuracy, precision, recall, and F1-score, providing a comprehensive view of performance:

Random Forest showed a tendency to classify instances as positive more frequently than warranted, evidenced by high recall but lower precision.

Gradient Boosting mirrored Random Forest in high recall, indicating effective identification of positive cases but at the expense of precision, leading to more false positives.

K-Nearest Neighbors demonstrated slightly lower accuracy but improved balance between precision and recall, suggesting a more cautious approach to classification.

Challenges Encountered

Key challenges included:

Overfitting: Particularly prevalent in complex models like Random Forest and Gradient Boosting. Strategies such as adjusting the maximum tree depth and increasing the ensemble size were employed to mitigate this.

Imbalanced Classes: Addressed using SMOTE to artificially enhance the minority class presence during training, aiming to improve the models' ability to recognize less frequent outcomes.

Discussion of Visual Results

Visual aids like confusion matrices and heatmaps were instrumental in illustrating model performance, showing the true and false positives and negatives for each class. These visuals helped in understanding the practical implications of model metrics and informed adjustments to model configurations.

Count plots and correlation heatmaps provided insights into the distribution of data and inter-feature relationships, essential for feature selection and engineering.

**Conclusion/Future Work**

The project demonstrated that machine learning techniques could effectively predict the severity of asthma, achieving an accuracy of approximately 75%. The use of algorithms like Random Forest, Gradient Boosting, and K-Nearest Neighbors provided insights into the potential of predictive analytics in healthcare. These findings underscore the relevance and potential impact of machine learning in enhancing asthma management practices by offering more timely and tailored interventions.

Contributions to the Field

This study contributes to the ongoing efforts to integrate machine learning into healthcare, specifically in the management of chronic diseases such as asthma. By successfully predicting asthma severity, healthcare providers can better allocate resources and plan treatments, potentially leading to improved patient outcomes and reduced healthcare costs.

Implications of the Results

The results suggest that machine learning models can handle complex health data and extract meaningful predictions that can aid in clinical decision-making. The high recall rates observed in the models indicate a strong ability to identify potential severe asthma cases, which is crucial for preventing severe asthma attacks.

Directions for Future Research

Integrating Larger Datasets: Future studies could incorporate more comprehensive datasets, possibly including additional patient data such as lifestyle factors, environmental conditions, and genetic information, to enhance the models' accuracy and applicability.

Exploring More Complex Models: Investigating more sophisticated machine learning models, such as deep learning techniques, could potentially improve the predictive accuracy and uncover more complex patterns in asthma severity.

Clinical Application and Validation: Implementing these models in a clinical setting would be a critical step forward. Future work should focus on real-world testing and validation of the models to ensure they perform well in practical healthcare environments.

Cross-Disciplinary Approaches: Collaboration with biomedical experts could refine the feature selection and model tuning processes, tailoring the models to be more sensitive to clinically relevant predictors of asthma attacks.

Personalized Medicine: There is an opportunity to develop personalized asthma management plans based on individual risk assessments generated by machine learning models. This approach could lead to more personalized, proactive healthcare interventions.

**Contributions**

Our project benefited greatly from the collaborative efforts of each team member. Srilekha's , focused on ensuring data quality through tasks like cleaning and feature transformation, while also meticulously documenting these processes.

Preethi Mathari expertise in machine learning algorithms was essential for refining our predictive models, utilizing techniques like GridSearchCV to optimize performance.

Sowmya Ila played a crucial role in analyzing model outputs and crafting insightful visual representations, contributing significantly to the discussion section of our report. Together, our contributions advanced our understanding of the data and generated robust predictive models with meaningful implications for our field.

**References**

<https://www.kaggle.com/datasets/deepayanthakur/asthma-disease-prediction>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10455492/>