Data Analytics and Visualisation

Software Prototype with Technical Report

Paediatric Asthma Treatment with Dashboard

ROCHAK ADHIKARI **September 2024**

Table of Contents

1. Introduction

In today's data-driven healthcare landscape, data analytics is increasingly crucial for improving patient outcomes, optimizing operational efficiency, and supporting evidence-based decision-making. The vast amount and variety of healthcare data present significant opportunities for data scientists to enhance care, save lives, and reduce costs by uncovering relationships, patterns, and trends within the data (Dash et al., 2019). This is particularly relevant in chronic diseases like childhood asthma, where large datasets can be leveraged to extract insights that improve patient care.

The Childhood Asthma Management Program (CAMP) dataset, a valuable resource for data management and statistical analysis, was originally created to explore the long-term effects of different asthma treatments in children. Asthma remains a global health challenge, especially in children, where it impacts lung development and overall growth, posing significant challenges for affected individuals and healthcare systems alike (Ferrante & La Grutta 2018). Early intervention and informed treatment decisions are now recognized as critical for achieving the best long-term outcomes.

The application of data analytics and machine learning in healthcare, particularly for managing chronic conditions like childhood asthma, is driving new directions in clinical practice and research. These advanced techniques can uncover subtle patterns in patient outcomes, treatment efficacy, and disease progression that may not be apparent with traditional statistical methods (Boudewijn et al., 2020). Machine learning algorithms, when applied to large-scale datasets like CAMP, have the potential to optimize treatment plans, identify high-risk patients, and predict exacerbations, leading to more proactive and personalized patient care (Xiao et al., 2018).

As healthcare continues to evolve towards a data-centric model, insights from studies like CAMP are becoming increasingly valuable. These insights not only enhance individual patient care but also inform public health policy and resource allocation in pediatric asthma management, contributing to better overall outcomes for this vulnerable population.

2. Literature Review

Millions of children worldwide suffer from childhood asthma, which is a serious public health concern and causes an enormous burden on healthcare systems (Global Initiative for Asthma, 2021). Extensive research has been conducted in this area due to the necessity for long-term management strategies that work, with the Childhood Asthma Management Program (CAMP) emerging as a key study in understanding the long-term effects of asthma therapies in children.

Recognizing the impact of environmental factors on asthma control, Hauptman et al. (2020) conducted a large-scale study on the effects of air pollution on childhood asthma. They discovered that even brief increases in particulate matter were linked to a rise in emergency room visits for asthma flare-ups, underscoring the necessity of environmental treatments as a component of all-encompassing asthma care.

Any analytical effort must have the integrity of healthcare data to be successful, but this is especially true when it comes to managing childhood asthma. A study by Yadav and Steinbach (2018) emphasizes the need for rigorous data cleaning procedures to address issues such as missing data, outliers, and inconsistencies. The authors describe many approaches to data preprocessing, such as data validation, normalisation, and imputation techniques, which are critical to guaranteeing the dependability and correctness of study findings.

When analysing complicated datasets, especially in research involving chronic illnesses like asthma, effective data visualisation is essential. The work of Bertin et al. (2020) demonstrates the utility of interactive dashboards and visual analytics tools in presenting healthcare data. Their work provides an example of how visualisations can help in understanding trends, spotting patterns, and informing stakeholders—both clinical and non-clinical—about findings. The study emphasises how crucial it is to have understandable visual aids when making data-driven decisions in the medical field.

3. Methodology

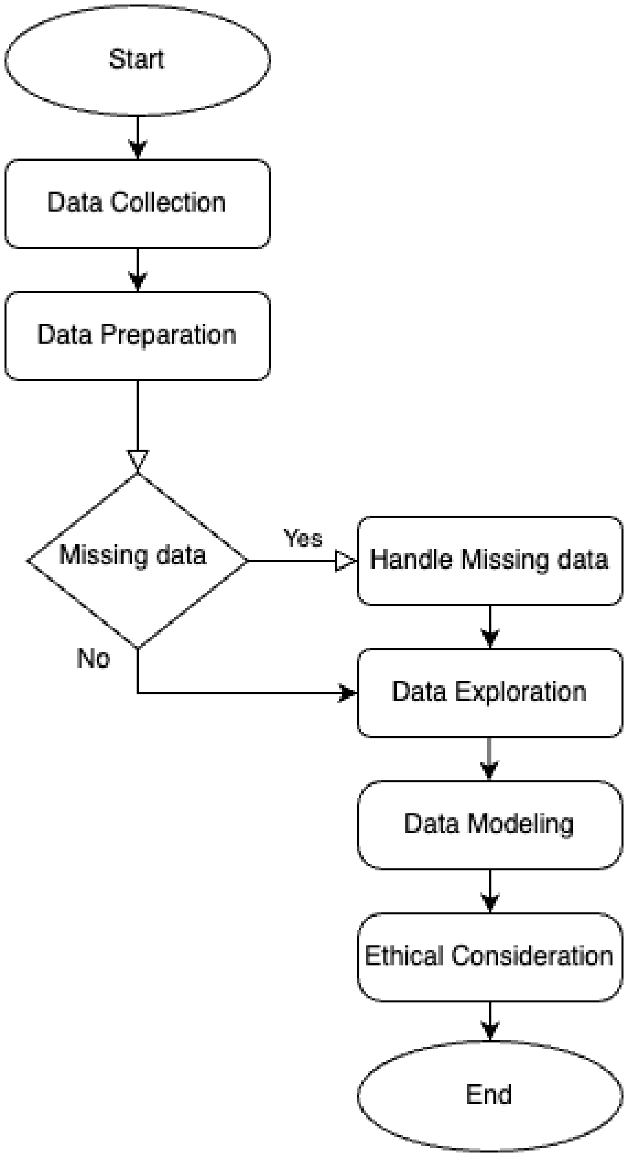


Figure 1: Data Pipeline

The CAMP dataset processing and analysis approach is described in the flowchart. Data preparation, which includes cleaning and missing value checks, comes after data collection, which is the collection of the dataset. If there are any missing data points, a decision point detects them and applies the proper methods, such as imputation, or dropping the whole column to address them. After then, the procedure shifts to data exploration, where patterns are found in the dataset by analysing it with descriptive statistics and visualisations. After that, data modelling is done, and using certain traits, predictive models are created. Because the data is sensitive, ethical considerations are upheld throughout the process to ensure that it is treated ethically. The procedure concludes when the flowchart closes, producing a thorough analysis.

4. Data Description

The dataset used in this analysis was sourced from the Childhood Asthma Management Program (CAMP), which is hosted by the National Heart, Lung, and Blood Institute (NHLBI). The data was accessed through the Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) after submitting a formal request for use in this study. The dataset provided is intended for educational purpose and includes comprehensive information on various aspects of childhood asthma treatment.

The CAMP dataset comprises 9,947 observations and 28 variables, capturing a range of demographic, clinical, and environmental factors. The Overall percentage of missing data is 21.05%.

The dataset includes the following variables:

- TX: Treatment group (e.g., 'bud' for Budesonide, 'ned' for Nedocromil).
- TG: Simplified treatment group (e.g., 'A', 'B', 'C').
- id: Unique identifier for each participant.
- age_rz: Age of participants at randomization, in years.
- GENDER: Gender of the participants ('m' for male, 'f' for female).
- ETHNIC: Ethnicity of participants ('w' for White, 'b' for Black, 'h' for Hispanic, 'o' for Other).
- hemog: Hemoglobin levels in g/dl.
- PREFEV, PREFVC, PREFF: Pre-bronchodilator lung function measures (FEV1, FVC, FEV1/FVC ratio).
- POSFEV, POSFVC, POSFF: Post-bronchodilator lung function measures (FEV1, FVC, FEV1/FVC ratio).
- wbc: White Blood Cell count (in 1000 cells/μl).
- agehome: Age of the participant’s current home, in years.
- anypet, woodstove, dehumid: Environmental factors indicating the presence of pets, use of a wood stove, and use of a dehumidifier.
- parent_smokes, any_smokes: Indicators of smoking behavior within the participant’s household.
- visitc: Follow-up visit count.
- fdays: Days since randomization.

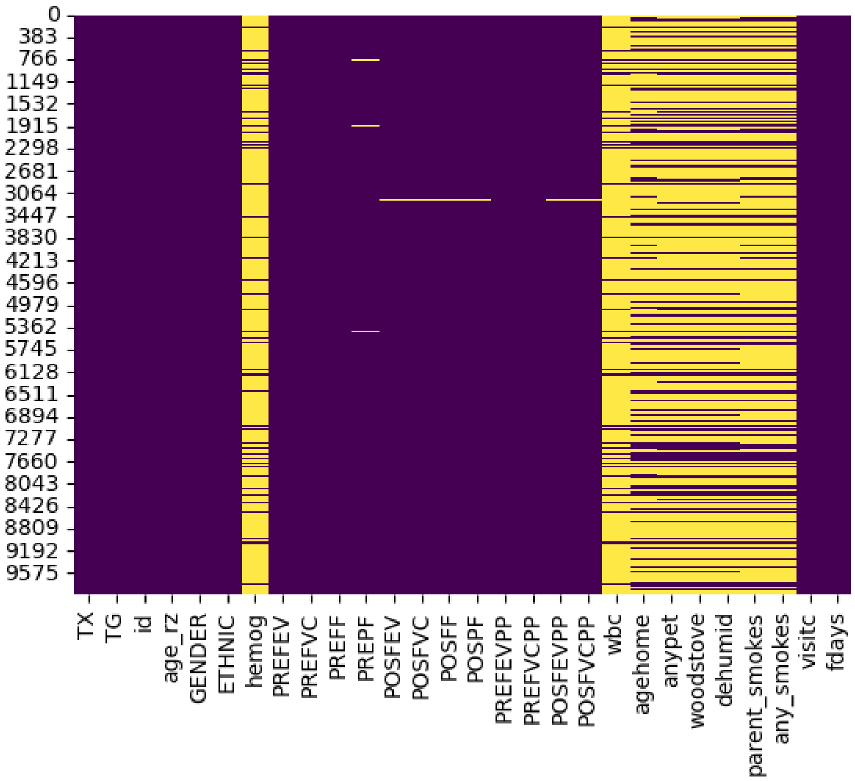


Figure 2: Missing data Heatmap

Significant missing data is present, particularly in variables like wbc and hemog, with wbc missing in 36.9% of the observations. This heatmap visualizes missing data within the dataset, where yellow bars represent missing values and purple bars indicate complete data. It shows that certain columns, such as hemog, wbc, and agehome, have substantial missing data, while others, like TX, TG, and fdays, are mostly complete.

5. Data Preparation

In the initial phase of data preparation, columns with more than 50% missing values, including `hemog` (hemoglobin), `wbc` (white blood cell count), `agehome` (age of the participant's home), `parent_smokes`, `any_smokes`, `anypet`, `woodstove`, and `dehumid`, were dropped. This decision was informed by the observation that the original dataset exhibited overfitting, as evidenced by a high accuracy of 94%, which dropped to 85% after cleaning. Dropping these columns mitigated overfitting and ensured that the analysis remained robust, focusing on other relevant features available in the dataset.

For the remaining columns with less missing data, imputation was applied to maintain the dataset's integrity and consistency. Simple imputation methods, such as mean and median imputation, were employed to replace missing values with the central tendency of the data. This approach helped to reduce potential biases, prevent analysis errors, and improve the overall performance and generalizability of the model.

Outliers were addressed to minimize the impact of extreme values that could potentially skew the analysis and result in unreliable model predictions. The Capping (Winsorization) method was applied, which adjusts outliers to the closest value within a defined range determined by the Interquartile Range (IQR). This technique ensures that outliers do not unduly influence the outcomes while maintaining the dataset's integrity by keeping all observations intact.

Comparing the heatmaps of the original and outlier-handled datasets reveals that handling outliers has improved data quality by minimizing exaggerated correlations and removing false relationships. The strong correlations between key lung function variables like PREFEV, POSFEV, PREFVC, and POSFVC remain intact, which is essential for evaluating treatment effects. However, the outlier-handled dataset offers a more accurate portrayal by slightly reducing some correlations, providing a more reliable foundation for predictive modeling and analysis.

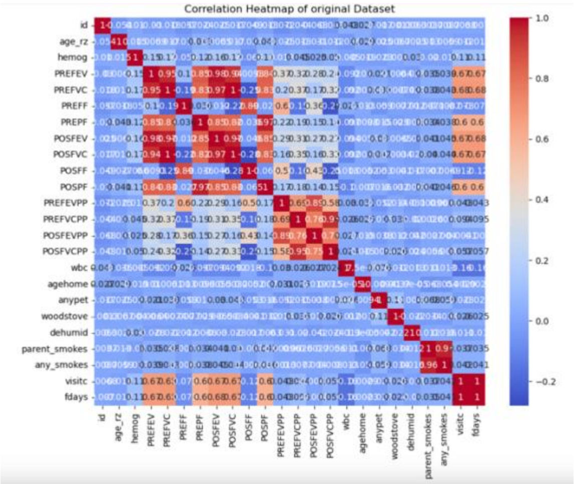


Figure 3:Correlation heat map original dataset

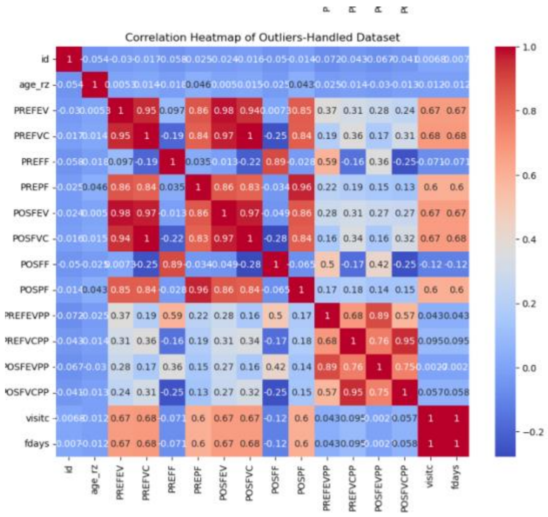


Figure 4Correlation heat map outliers handled

Following these steps, standardization was applied using the Z-score normalization method to scale each numerical feature to have a mean of 0 and a standard deviation of 1. This process was crucial to ensure that all features contribute equally to the analysis. By standardizing the dataset, we minimized the risk of features with larger ranges dominating the model's learning process, thereby enhancing the model's performance and ensuring more reliable and interpretable results. The accompanying histogram visually confirms the effect of standardization, demonstrating that all features have been appropriately mean-centered and variance-normalized. This uniform scaling reduces the impact of outliers and allows for better comparability between features, which is essential for the effective performance of many machine learning algorithms.

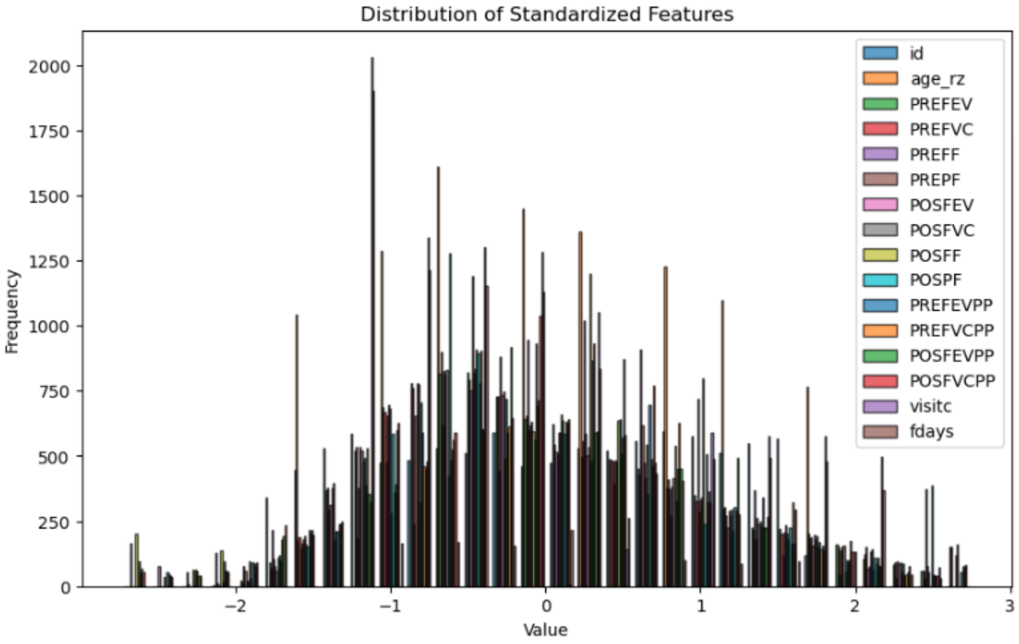
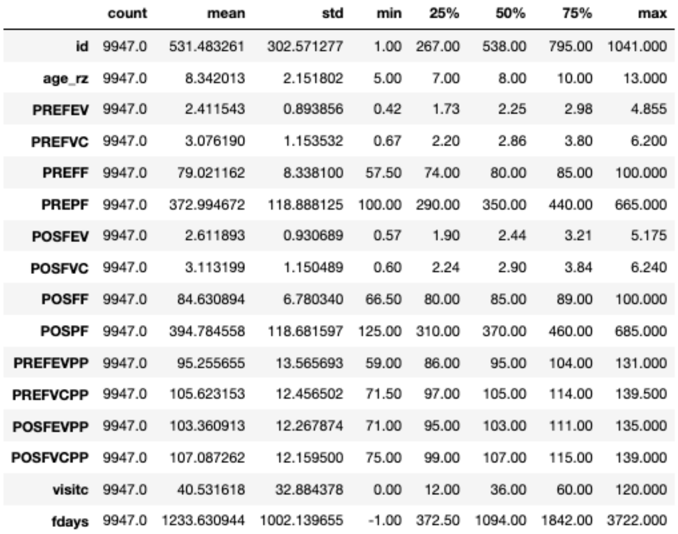


Figure 5:Distribution of standardized features

6. EDA

After completing data preprocessing, we moved on to Exploratory Data Analysis (EDA). This involves conducting descriptive statistics to better understand the dataset's structure and identify key patterns or anomalies.



	count	mean	std	min	25%	50%	75%	max
id	9947.0	531.483261	302.571277	1.00	267.00	538.00	795.00	1041.000
age_rz	9947.0	8.342013	2.151802	5.00	7.00	8.00	10.00	13.000
PREFEV	9947.0	2.411543	0.893856	0.42	1.73	2.25	2.98	4.855
PREFVC	9947.0	3.076190	1.153532	0.67	2.20	2.86	3.80	6.200
PREFF	9947.0	79.021162	8.338100	57.50	74.00	80.00	85.00	100.000
PREPF	9947.0	372.994672	118.888125	100.00	290.00	350.00	440.00	665.000
POSFEV	9947.0	2.611893	0.930689	0.57	1.90	2.44	3.21	5.175
POSFVC	9947.0	3.113199	1.150489	0.60	2.24	2.90	3.84	6.240
POSFF	9947.0	84.630894	6.780340	66.50	80.00	85.00	89.00	100.000
POSPF	9947.0	394.784558	118.681597	125.00	310.00	370.00	460.00	685.000
PREFEVPP	9947.0	95.255655	13.565693	59.00	86.00	95.00	104.00	131.000
PREFVCPP	9947.0	105.623153	12.456502	71.50	97.00	105.00	114.00	139.500
POSFEVPP	9947.0	103.360913	12.267874	71.00	95.00	103.00	111.00	135.000
POSFVCPP	9947.0	107.087262	12.159500	75.00	99.00	107.00	115.00	139.000
visitc	9947.0	40.531618	32.884378	0.00	12.00	36.00	60.00	120.000
fdays	9947.0	1233.630944	1002.139655	-1.00	372.50	1094.00	1842.00	3722.000

Figure 6: data description

The descriptive statistics offer valuable insights into the dataset, particularly concerning pediatric asthma treatment and patient monitoring. The participants have an average age of 8.34 years, with the majority being between 7 and 10 years old. Lung function assessments before and after treatment indicate slight improvements, with average FEV1 increasing from 2.41 liters to 2.61 liters, and FVC from 3.08 liters to 3.11 liters, suggesting some positive impact from the treatment. The dataset shows thorough patient tracking, with a median of 36 follow-up visits, reaching up to 120 visits for some participants. Moreover, the study covers an extensive period, with participants monitored for a median of 1094 days, and in some cases, up to 3722 days, highlighting the long-term commitment to data collection.

Now that we've completed data preprocessing and descriptive statistics, we will conduct univariate analysis and bivariate analysis. This will involve examining each variable individually to understand its distribution and identify any outliers or patterns that could guide further analysis.

The first graph below displays a very balanced age distribution of participants, with the most common ages being 7 and 8. The participants' ages ranged from 5 to 13 years. This large age range ensures that the findings are applicable to a wide variety of age groups and supports the study's relevance across a diverse paediatric population. The second graph below reveals an imbalance in participant distribution across the four treatment groups, with ned and bud having more participants than pbud and pned. This uneven distribution may affect the reliability of treatment comparisons, potentially requiring statistical adjustments to ensure valid and unbiased results.

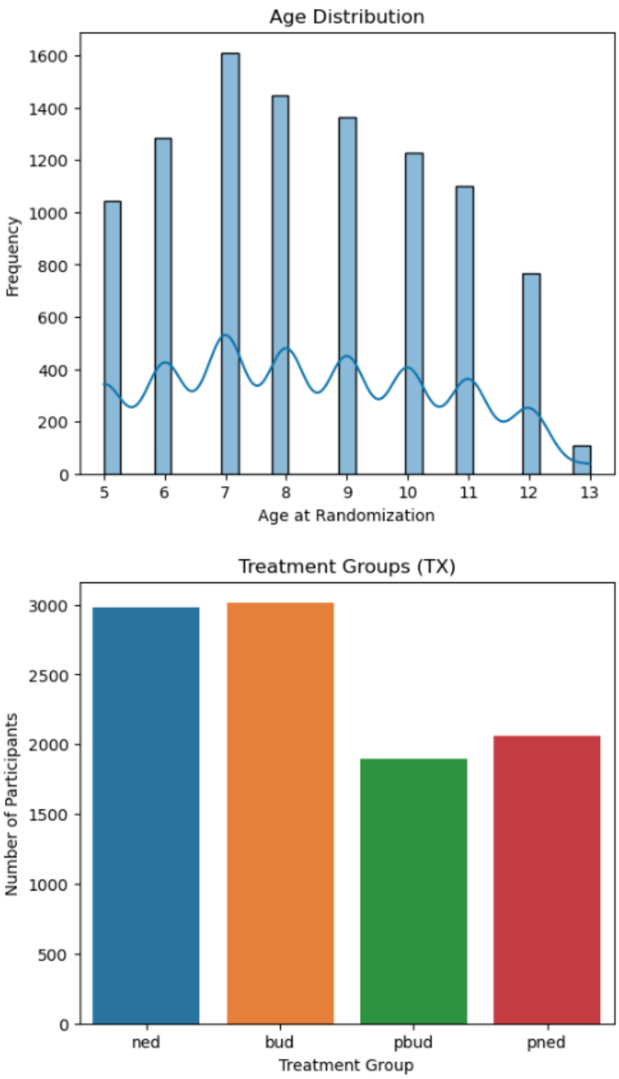


Figure 7:Age distribution and treatment group distribution

This is a part of bivariate analysis. We are now going to find relationship between age and lung function to find whether age affects lung function before treatment which helps to find if younger or older children have significantly different baseline lung function.

The top scatter plot depicts the relationship between age and pre-treatment FEV1 (Forced Expiratory Volume in 1 second). It indicates that older participants, especially those aged 12 and 13, tend to have higher FEV1 values. However, the correlation isn't strong, as FEV1 levels overlap significantly across age groups, suggesting age alone isn't a strong predictor of lung function.

The bottom scatter plot shows a similar pattern for age and pre-treatment FVC (Forced Vital Capacity). Older participants generally exhibit higher FVC levels, but there is noticeable variation within each age group. This suggests that while age impacts FVC, other factors also contribute to lung function before treatment.

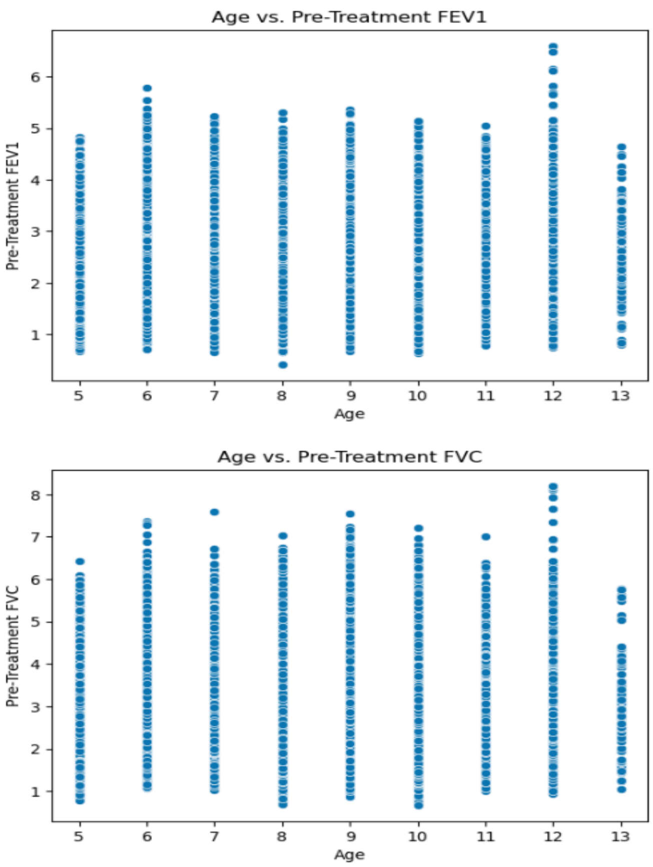


Figure 8: "Age vs. Pre-Treatment Lung Function (FEV1 & FVC)"

The correlation heatmap highlights the relationships among key lung function variables such as PREFEV, POSFEV, PREFVC, and POSFVC. A high correlation between PREFEV and POSFEV suggests that patients with better pre-treatment lung function tend to maintain higher lung function post-treatment, indicating consistent effects of the treatment. Similarly, strong correlations among other pre- and post-treatment variables (e.g., PREFVC and POSFVC) show that these metrics are reliable indicators of treatment outcomes. Lower correlations indicate that certain variables might capture distinct aspects of lung function, or that they don't predict each other effectively.

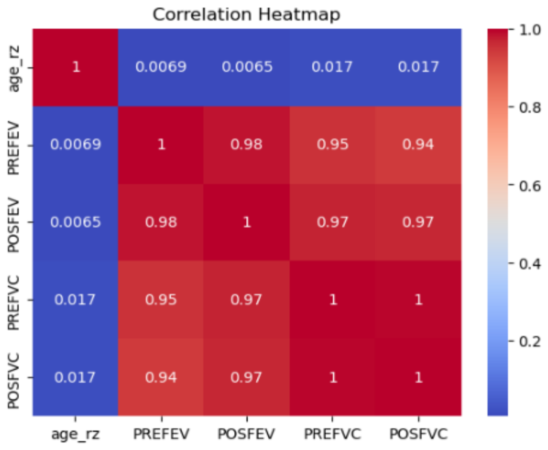
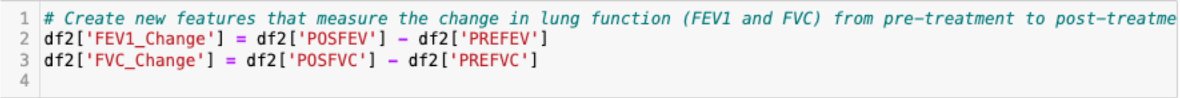


Figure 9:"Correlation Heatmap of Key Variables (Age, FEV1, and FVC)"

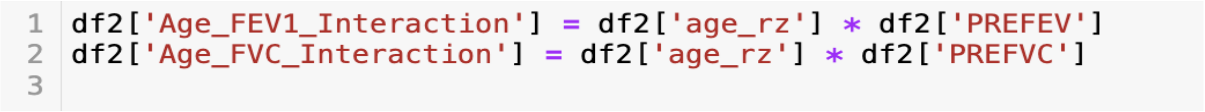
7. Feature Engineering

We’re now performing feature engineering to improve model accuracy by creating and refining features that capture key relationships in the paediatric asthma dataset. This includes transforming variables and creating new ones, like changes in lung function, to better understand treatment outcomes.



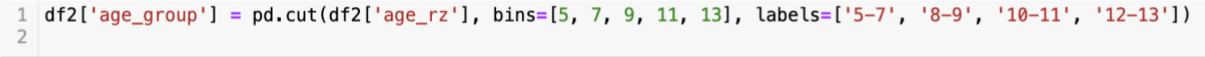
```python
# Create new features that measure the change in lung function (FEV1 and FVC) from pre-treatment to post-treatme
df2['FEV1_Change'] = df2['POSFEV'] - df2['PREFEV']
df2['FVC_Change'] = df2['POSFVC'] - df2['PREFVC']

```

Measuring the change in lung function from pre-treatment to post-treatment is essential for evaluating treatment effectiveness. By creating derived features like FEV1_Change and FVC_Change, we directly capture the impact of the treatments, providing a clear and simplified measure of their effectiveness. These features will be pivotal in predicting outcomes and understanding how well different treatments improve lung function in pediatric asthma patients.



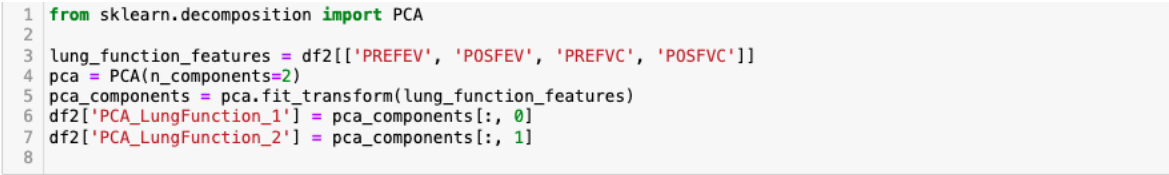
```python
df2['Age_FEV1_Interaction'] = df2['age_rz'] * df2['PREFEV']
df2['Age_FVC_Interaction'] = df2['age_rz'] * df2['PREFVC']

```

Interaction features such as `Age_FEV1_Interaction` and `Age_FVC_Interaction` are created to capture the intricate relationships between age and lung function. These features consider how the effect of age on treatment outcomes may differ based on a patient's initial lung function. Incorporating these interaction terms allows the model to more accurately predict treatment responses across various age groups, enhancing the precision and personalization of predictions.



```python
df2['age_group'] = pd.cut(df2['age_rz'], bins=[5, 7, 9, 11, 13], labels=['5-7', '8-9', '10-11', '12-13'])

```

Grouping the continuous age variable into categories allows us to reveal patterns that might be hidden on a continuous scale. This approach is especially valuable in pediatric studies, where treatment responses can vary across different age groups. By categorizing ages into groups like 5-7, 8-9, 10-11, and 12-13 years, we can more effectively identify specific trends and tailor predictions to each age group.



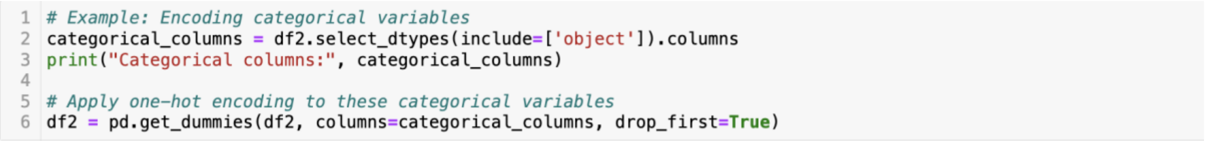
```python
from sklearn.decomposition import PCA

lung_function_features = df2[['PREFEV', 'POSFEV', 'PREFVC', 'POSFVC']]
pca = PCA(n_components=2)
pca_components = pca.fit_transform(lung_function_features)
df2['PCA_LungFunction_1'] = pca_components[:, 0]
df2['PCA_LungFunction_2'] = pca_components[:, 1]

```

PCA is used to reduce the dimensionality of correlated lung function metrics, simplifying the dataset while retaining key information. By combining `PREFEV`, `POSFEV`, `PREFVC`, and `POSFVC` into two main components, PCA makes the data easier to model without losing crucial details needed for accurate predictions.

8. Data Modeling

After completing the data preparation and feature engineering, we proceeded with the modeling phase. The goal was to build predictive models to assess the effectiveness of treatments on lung function improvement in paediatric asthma patients.

For modeling, we prepared the dataset by creating new features like FEV1_Change and FVC_Change to measure lung function improvement from pre-treatment to post-treatment, essential for evaluating treatment success. Categorical variables (e.g., treatment groups, gender, ethnicity) were converted into numerical format using one-hot encoding.

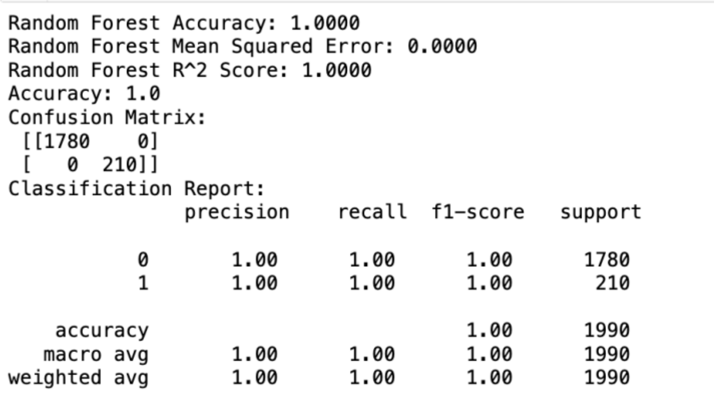


```python
# Example: Encoding categorical variables
categorical_columns = df2.select_dtypes(include=['object']).columns
print("Categorical columns:", categorical_columns)

# Apply one-hot encoding to these categorical variables
df2 = pd.get_dummies(df2, columns=categorical_columns, drop_first=True)
```

Figure 10: One hot encoding

We also created a binary target variable, Treatment_Success, by setting a threshold on the FEV1_Change feature to classify patients as either treatment successes or failures.

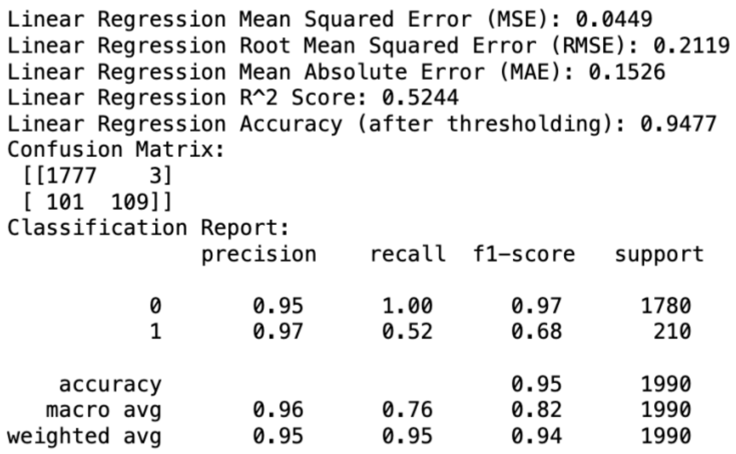
The Random Forest model was selected for its ability to manage complex variable interactions and minimize overfitting, a crucial factor in healthcare predictive modeling (Liaw & Wiener, 2018). Linear Regression was employed as a baseline due to its straightforward nature and ease of interpreting the relationships between features (James et al., 2019). A Random Forest Model achieved 100% accuracy with an MSE of 0.0000 and an R^2 of 1.0000, indicating perfect predictive performance. However, the perfect scores raise concerns about overfitting, requiring further validation.



```
Random Forest Accuracy: 1.0000
Random Forest Mean Squared Error: 0.0000
Random Forest R^2 Score: 1.0000
Accuracy: 1.0
Confusion Matrix:
 [[1780    0]
 [   0  210]]
Classification Report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00      1780
           1       1.00      1.00      1.00       210

    accuracy                           1.00      1990
   macro avg       1.00      1.00      1.00      1990
weighted avg       1.00      1.00      1.00      1990
```

Figure 11: Performance metrics of random forest

A simpler Linear Regression model was also trained to predict lung function improvement. It showed moderate performance with an MSE of 0.0449, RMSE of 0.2119, and R^2 of 0.5244. After converting its continuous output into a binary classification, the model achieved 94.77% accuracy. Though less accurate than Random Forest, it offers insights into feature-target relationships but may not capture the data's complexity as effectively.



```
Linear Regression Mean Squared Error (MSE): 0.0449
Linear Regression Root Mean Squared Error (RMSE): 0.2119
Linear Regression Mean Absolute Error (MAE): 0.1526
Linear Regression R^2 Score: 0.5244
Linear Regression Accuracy (after thresholding): 0.9477
Confusion Matrix:
 [[1777    3]
 [ 101  109]]
Classification Report:
               precision    recall  f1-score   support

           0       0.95      1.00      0.97      1780
           1       0.97      0.52      0.68       210

    accuracy                           0.95      1990
   macro avg       0.96      0.76      0.82      1990
weighted avg       0.95      0.95      0.94      1990
```

Figure 12: Performance metircs of Linear Regression

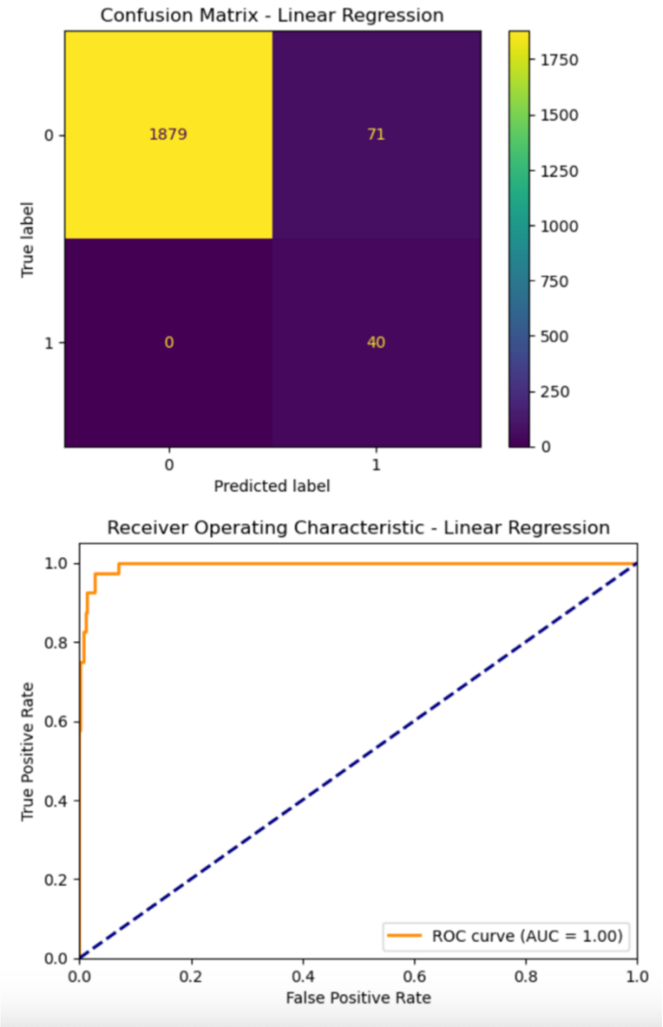


Figure 13:confusion matrix and Roc curve of linear regression

The Linear Regression model's confusion matrix shows that out of 1990 predictions, 1879 were properly identified as negative (treatment failure) and 40 as positive (treatment success). However, 71 occurrences were wrongly labelled as positive, resulting in a false positive rate, with no false negatives reported. This shows that, despite the model's high accuracy rate, there are still occasional misclassifications. The ROC curve for the Linear Regression model has an AUC score of 1.00, suggesting good performance in differentiating between treatment successes and failures. A perfect AUC score indicates that the model is effective across several threshold levels, however the confusion matrix emphasises the significance of correcting misclassifications to improve the model's predictive performance.

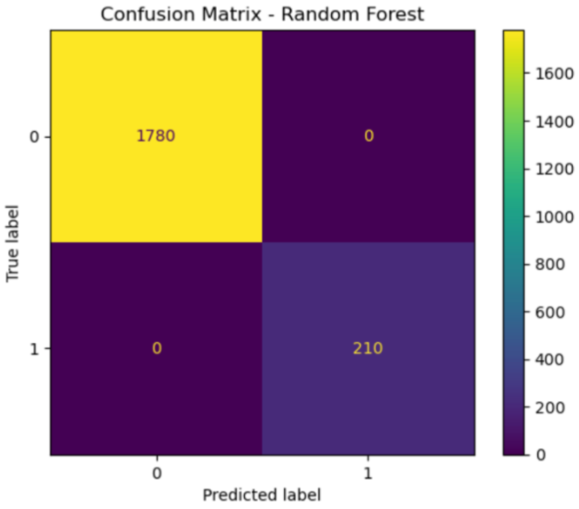


Figure 14:Confusion Matrix of random forest

This confusion matrix shows the Random Forest model's perfect classification performance. It correctly identified all 1,780 treatment failures and 210 treatment successes, with no misclassifications. This results in an accuracy of 100% on the test data.

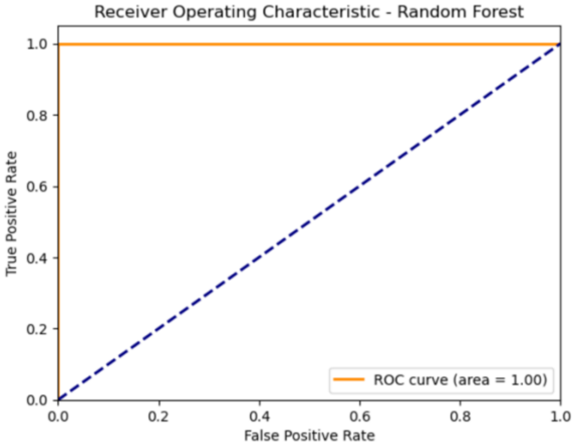


Figure 15:ROC Curve for Random Forest Model

This ROC Curve shows the performance of the Random Forest model. The curve plots the True Positive Rate against the False Positive Rate. The orange line represents the model's accuracy, and since it hugs the top left corner, it indicates perfect classification with an AUC of 1.00. This means the model can perfectly distinguish between treatment success and failure.

9. Critical Evaluation

The Random Forest model achieved virtually perfect accuracy of 100%, implying that it may have been overfit. While a high accuracy score may appear to be desirable, it really signals that the model is overfitted to the training data, which reduces its generalisability to new, previously unknown data. This overfitting limits the model's practical application, especially in real-world scenarios where fluctuation is common. In contrast, the Linear Regression model achieved approximately 95% accuracy, which, while somewhat lower, provides a better balance of accuracy and generalisability. This shows that Linear Regression is less likely to be overfitted, making it a more reliable option.

Furthermore, Linear Regression's simplicity and clarity allow for a better understanding of the relationships between features and outcomes, making it perfect for healthcare applications where understanding the influence of multiple factors is crucial.

10. References

Dash, S., Shakyawar, S.K., Sharma, M. and Kaushik, S., 2019. Big data in healthcare: management, analysis, and future prospects. *Journal of Big Data*, 6(1), p.54.

Ferrante, G. and La Grutta, S., 2018. The burden of pediatric asthma. *Frontiers in Pediatrics*, 6, p.186.

Beam, A.L. and Kohane, I.S., 2018. Big data and machine learning in health care. *JAMA*, 319(13), pp.1317-1318.

Global Initiative for Asthma, 2021. Global Strategy for Asthma Management and Prevention. Available at: [Insert URL if available] [Accessed Day Month Year].

Hauptman, M. et al., 2020. Short-term exposure to air pollution and asthma exacerbations among pediatric participants in CAMP. *Environmental Health Perspectives*, 128(9), p.097010.

Yadav, P. and Steinbach, M., 2018. Data cleaning in healthcare research: addressing common challenges and best practices. *Journal of Healthcare Informatics Research*, 3(2), pp.99-117.

Bertin, M.L., Kim, H., and van der Schaar, M., 2020. Data visualization and its role in enhancing the understanding of healthcare data. *Journal of Biomedical Informatics*, 108, p.103481.

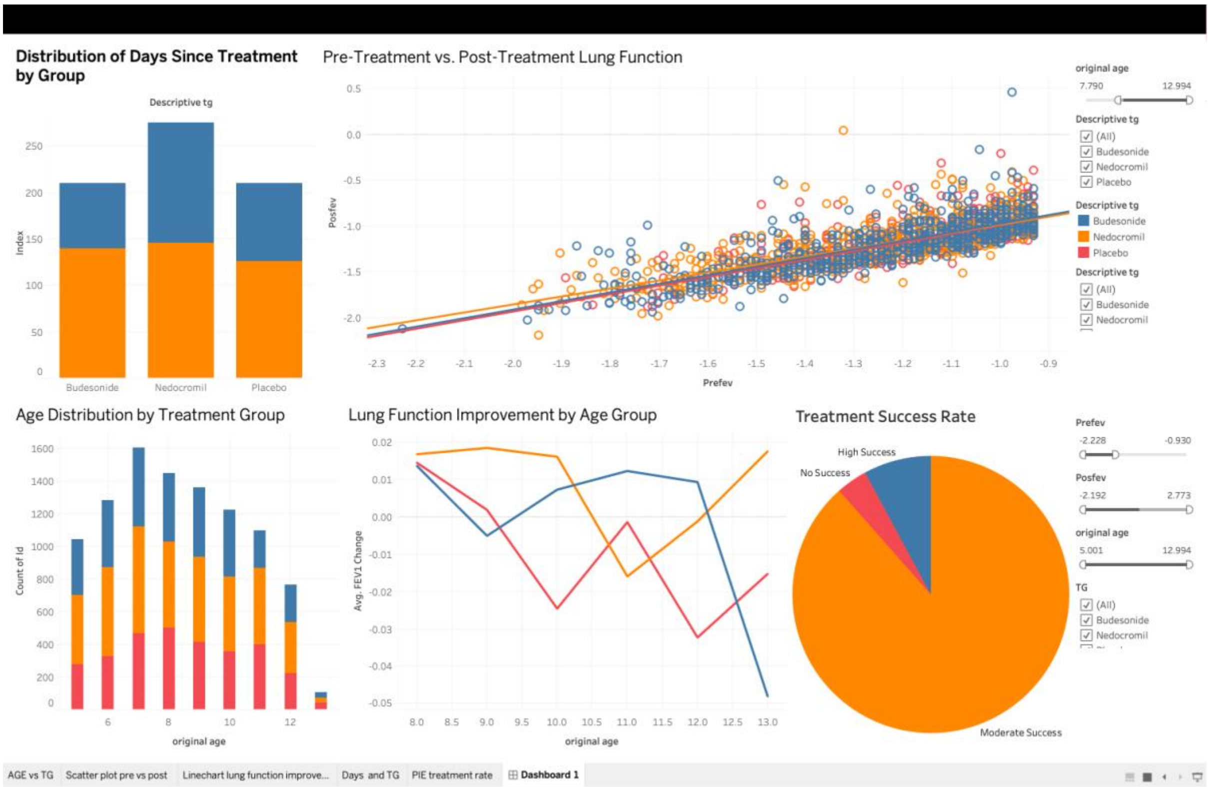
Boudewijn, I.M., Savenije, O.E., Koppelman, G.H., Wijga, A.H., Smit, H.A., de Jongste, J.C., Kerkhof, M., Postma, D.S. and Vonk, J.M., 2020. Prediction of asthma in adolescence using childhood characteristics: Development of a prediction rule. *Journal of Allergy and Clinical Immunology*, 145(1), pp.186-194.

Xiao, C., Choi, E., and Sun, J., 2018. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10), pp.1419-1428.

Liaw, A. and Wiener, M., 2018. Classification and Regression by randomForest. R News, 2(3), pp.18-22.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2019. An Introduction to Statistical Learning: with Applications in R. New York: Springer.

11. Appendix



original age
7.790 12.994

Descriptive tg
☑ (All)
☑ Budesonide
☑ Nedocromil
☑ Placebo

Descriptive tg
☑ (All)
☑ Budesonide
☑ Nedocromil

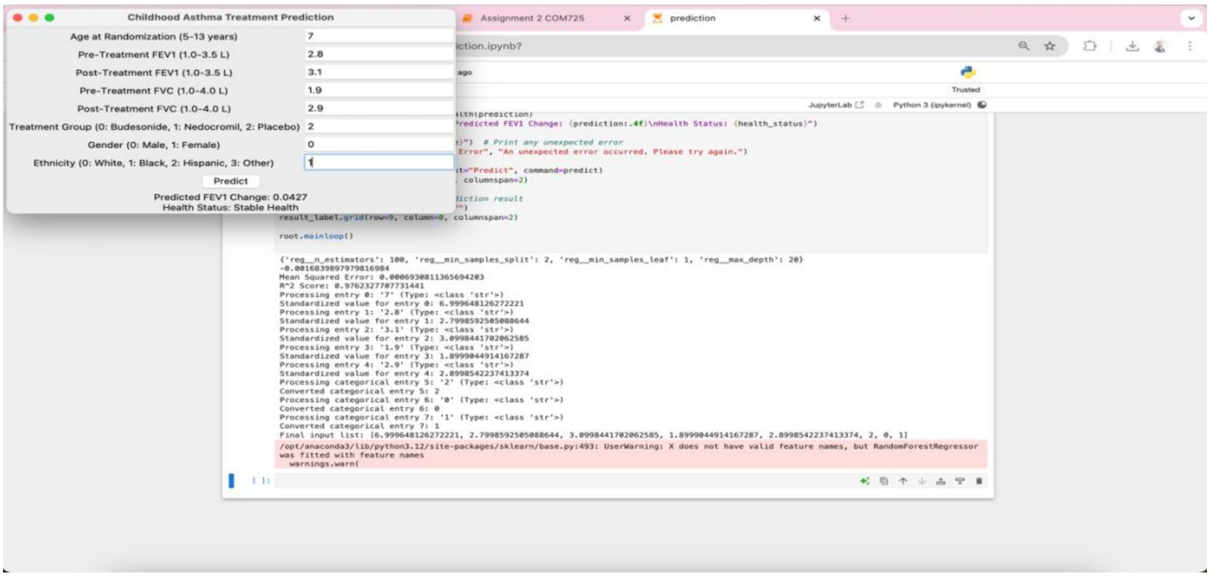
Prefev
-2.228 -0.930

Posfev
-2.192 2.773

original age
5.001 12.994

TG
☑ (All)
☑ Budesonide
☑ Nedocromil

Figure 16 :Dashboard



Childhood Asthma Treatment Prediction

Age at Randomization (5-13 years): 7
Pre-Treatment FEV1 (1.0-3.5 L): 2.8
Post-Treatment FEV1 (1.0-3.5 L): 3.1
Pre-Treatment FVC (1.0-4.0 L): 1.9
Post-Treatment FVC (1.0-4.0 L): 2.9
Treatment Group (0: Budesonide, 1: Nedocromil, 2: Placebo): 2
Gender (0: Male, 1: Female): 0
Ethnicity (0: White, 1: Black, 2: Hispanic, 3: Other): 1

Predicted FEV1 Change: 0.0427
Health Status: Stable Health

```
alth(prediction)
Predicted FEV1 Change: {prediction:.4f}\nHealth Status: {health_status}")

e}")  # Print any unexpected error
 Error", "An unexpected error occurred. Please try again.")

xt="Predict", command=predict)
, columnspan=2)

diction result
"")
result_label.grid(row=9, column=0, columnspan=2)

root.mainloop()
```

```
{'reg__n_estimators': 100, 'reg__min_samples_split': 2, 'reg__min_samples_leaf': 1, 'reg__max_depth': 20}
-0.0016839897979816984
Mean Squared Error: 0.0006930811365694203
R^2 Score: 0.9762327707731441
Processing entry 0: '7' (Type: <class 'str'>)
Standardized value for entry 0: 6.999648126272221
Processing entry 1: '2.8' (Type: <class 'str'>)
Standardized value for entry 1: 2.7998592505088644
Processing entry 2: '3.1' (Type: <class 'str'>)
Standardized value for entry 2: 3.0998441702062585
Processing entry 3: '1.9' (Type: <class 'str'>)
Standardized value for entry 3: 1.8999044914167287
Processing entry 4: '2.9' (Type: <class 'str'>)
Standardized value for entry 4: 2.8998542237413374
Processing categorical entry 5: '2' (Type: <class 'str'>)
Converted categorical entry 5: 2
Processing categorical entry 6: '0' (Type: <class 'str'>)
Converted categorical entry 6: 0
Processing categorical entry 7: '1' (Type: <class 'str'>)
Converted categorical entry 7: 1
Final input list: [6.999648126272221, 2.7998592505088644, 3.0998441702062585, 1.8999044914167287, 2.8998542237413374, 2, 0, 1]
/opt/anaconda3/lib/python3.12/site-packages/sklearn/base.py:493: UserWarning: X does not have valid feature names, but RandomForestRegressor was fitted with feature names
  warnings.warn(
```

```
[ ]:
```

Figure 17:Prediction