Asthma Disease Diagnosis and Visualization

Priyadarsini Panda

EC Utbildning

Kunskapskontroll2_DLP

2025-09

Abstract

This Project aims to develop and evaluate a machine learning model to predict asthma diagnosis using Python. It includes data preprocessing, class balancing with SMOTEEN and model training using Logistic Regression,Random Forest, XGBoost. Hyperparameter tuning with GridSearch and threshold tuning was applied to prioritize recall to minimize missed asthma cases. The final models were integrated into Python and exported for visualization in PowerBI, enabling interactive dashboards that display predicted probabilities, asthma risk scores and feature importances. Diagnosis labels are defined as 0 for non-asthmatic and 1 for asthmatic.

In additon to the implementation was carried out in Jupyter Notebook, where exploratory data analysis, preprocessing and model testing were performed using both SMOTE and SMOTEEN to compare the result.

The dataset was obtained from Kaggle by Rabie El Kharoua, (2024). https://www.kaggle.com/datasets/rabieelkharoua/asthma-disease-dataset/code . It is synthetic and designed for educational purposes. In addition to more real-world data would improve precision.

Keywords: Asthma Dataset(From kaggle), Asthma diagnosis, Preprocessing, Balancing data, Hyperparameter Tuning, Model Training, Model Evaluation, Confusion Matrix, classification report, Feature extraction, Power BI, Streamlit.

Acknowledgement

I am thankful to my Mentor Antonio for their invaluable guidance and expertise , which have been a constant source of inspiration and learning and heartfelt thanks to my classmates for their kindness and generosity.

Innehållsförteckning

Abstract 2

1 Introduction 1

2 Methodology 1

2.1 Data Collection and Cleaning 1

2.2 Preprocessing 1

2.3 Models Tested with Threshold Adjustment 1

3 Evaluation 2

4 Results and Discussion 3

4.1 Confusion Matrix 3

4.2 Feature Importance 4

5 Power BI Integration 4

6 Streamlit App Deploy 5

7 Conclusion 6

Appendix A 7

References 7

1 Introduction

Asthma is chronic respiratory disease that causes the airways to become inflamed and overreactive, leading to symptoms like wheezing, shortness of breath, coughing, chest tightness etc. Therefore, a diagnosis is what a doctor makes after assessing a patient’s symptoms. It is the process for figuring out if a person is having asthma or not.

The purpose of this project to apply Python programming and data science techniques to build a predictive model for asthma diagnosis. In healthcare, early detection of asthma is crucial because it allows timely treatment and prevention of complications. Machine learning provides a way to combine multiple patient related factors such as Chest-tightness, Shortness of breath, wheezing, family history asthma, pollution etc..symptoms to predict whether a person is likely to have asthma or not.

2 Methodology

The project was implemented in python but it was first written in Jupyter Notebook to see the pattern of data set where exploratory data analysis were performed and how the dataset shape changes before and after applying SMOTEEN.

2.1 Data Collection and Cleaning

The asthma data set was downloaded from Kaggle. Missing values were checked through EDA and irrelevant columns PatientID and DoctorInCharge were removed to avoid bias in prediction.

2.2 Preprocessing

The dataset was split into training(80%) and test set(20%) to preserve class distribution. Class distribution was imbalanced (0= 2268, 1=124). To address imbalance class, SMOTEEN was applied that combines SMOTE oversampling and Edited Nearest Neighbors(ENN). Features were standardized using StandardScaler.

2.3 Models Tested with Threshold Adjustment

Three models (Logistic Regression, Random Forest, XGBoost) were tested with GridSearch for hyperparameter tuning in Jupyter Notebook. Instead of using the default threshold value 0.5, a lower threshold of 0.3 was chosen to increased recall for detecting asthma patients.

Based on the results, Logistic Regression and Random Forest were implemented in the final Python pipeline for further optimization and integration.

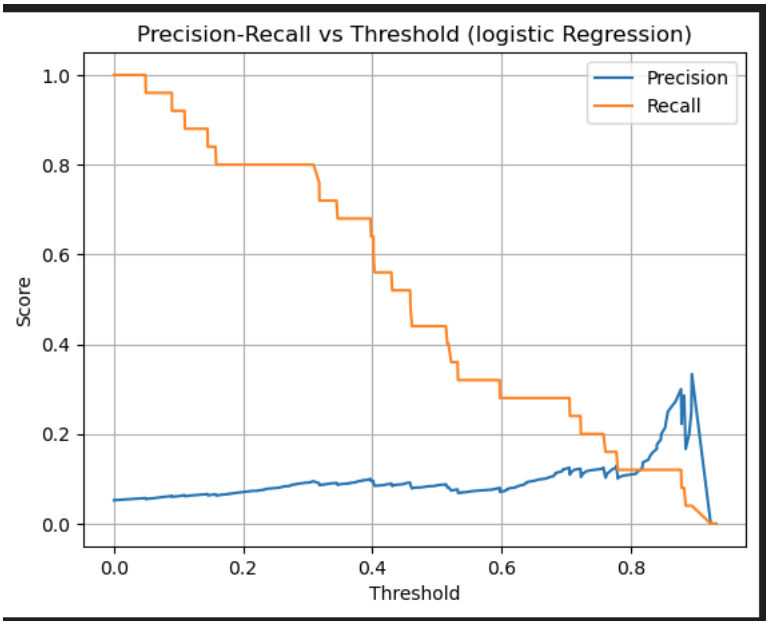


Image 1: precision and recall plot from Jupyter Notebook across different threshold

3 Evaluation

The model is evaluated using Confusion Matrix and classification report to calculate precision, recall, F1 score , accuracy.

Feature Importance analysis to visualize which factors influenced more in asthma diagnosis.

Confusion Matrix to track true positives, false positives , true negetavies and false negatives.

Precision: Out of all the patients the model said have asthma, how many actually do. High precision means fewer false alarms.

Recall: Out of all patients who actually have asthma, how many did the model catch. High recall means fewer missed cases.

F1 Score: A balance between precision and recall. Useful when you want both fewer false +ve and fewer false -ve.

Accuracy: Overall, how many prediction were correct. It can be misleading if one class dominates.

ROC-AUC : Measures how well the model seperates asthma v/s non-asthma across all threshold. Higher is better.

Diagnosis: 0 = non-asthmatic and 1 = asthmatic.

Class: 1(Asthma):

Model Name	Recall	Precision	F1-Score	Accuracy	ROC-AUC
Logistic Regression	0.800	0.092	0.165	0.576	0.670
Random Forest	0.440	0.088	0.147	0.733	0.622

4 Results and Discussion

Logistic Regression achieved the highest recall(0.800), correctly identifying 80% of asthma casses. This makes it the most suitable model for initial screening, where missing a diagnosis could lead serious health consequences. Even though its low precision(0.092), is acceptable in a medical context, as false positives can be followed up with further clinical testing.

Random Forest showed better overall accuracy with 0.733 higher than Logistic Regression but its recall 0.440 is lower Logistic Regression.

4.1 Confusion Matrix

Logistic Regression

Confusion Matrix:

```
[[256   198]
 [ 5      20]]
```

Random Forest

Confusion Matrix:

```
[[340   114]
  [14     11]]
```

Logistic Regression:

True Positives (TP): 20 patients correctly identified as asthmatic

False Positives (FP): 198 patients incorrectly flagged

False Negatives (FN): 5 missed asthma cases

True Negatives (TN): 256 correctly identified as non asthmatic

Random Forest:

True Positives (TP): 11 patients correctly identified as asthmatic

False Positives (FP): 114 patients incorrectly flagged

False Negatives (FN): 14 missed asthma cases

True Negatives (TN): 340 correctly identified as non asthmatic

Logistic regression had only 5 false negatives which strength in identifying true asthma cases but has many false positives. This is acceptable in medical screening, where missing a diagnosis is riskier than over predicting. Where Random Forest had 14 missed asthma cases which could be critical for asthmatic patients.

4.2 Feature Importance

Feature Importance provides insight into the top 15 features(such as HistoryOfAllergies, Eczema, ChestTightness, ShortnessOfBreath, Coughing, DustExposure, PetAllergy etc..) contributing to asthma diagnosis.

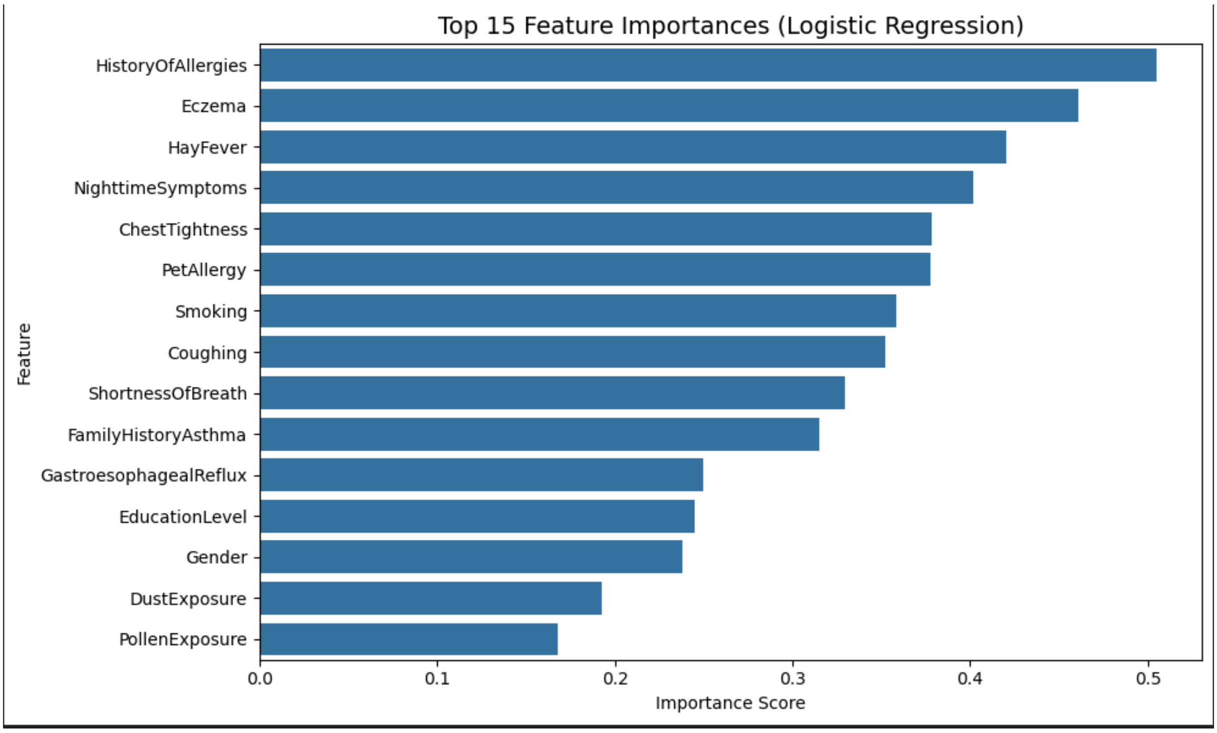
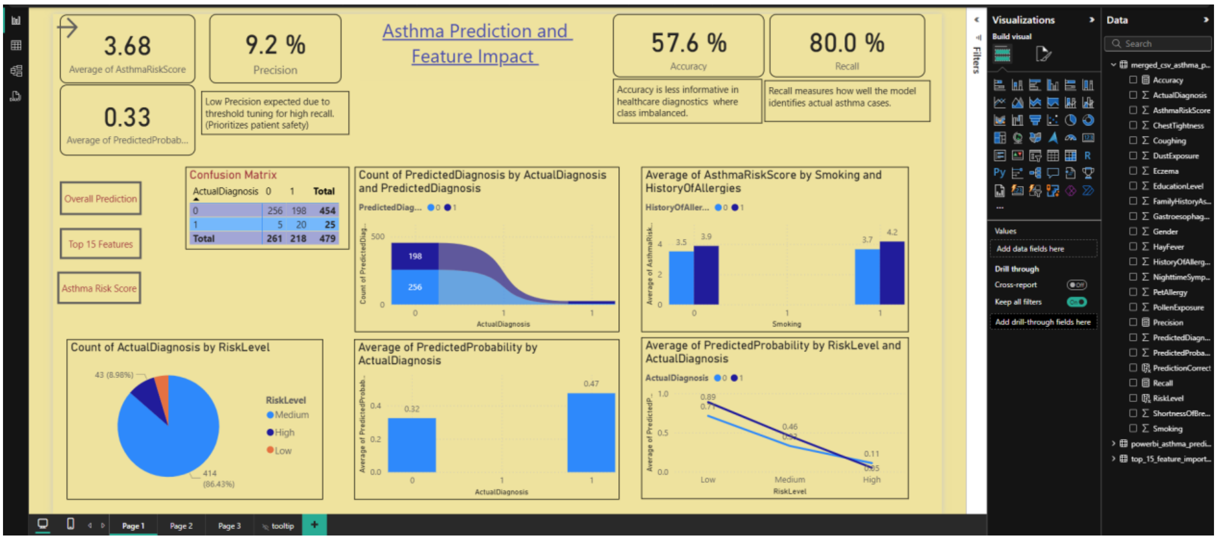


Image 2. From Jupyter Notebook top 15 features

5 Power BI Integration

Power Bi dashboard were created to visualize asthma diagnosis result and to support clinical decision making. This integration enables interactive visualzation of model prediction and feature contribution, making the results more accessible to healthcare professionals.



Low Precision expected due to threshold tuning for high recall. (Prioritizes patient safety)

Accuracy is less informative in healthcare diagnostics where class imbalanced.

Recall measures how well the model identifies actual asthma cases.

Confusion Matrix

ActualDiagnosis	0	1	Total
0	256	198	454
1	5	20	25
Total	261	218	479

Image 3. ScreenShot taken From Power BI Dashboard overall view

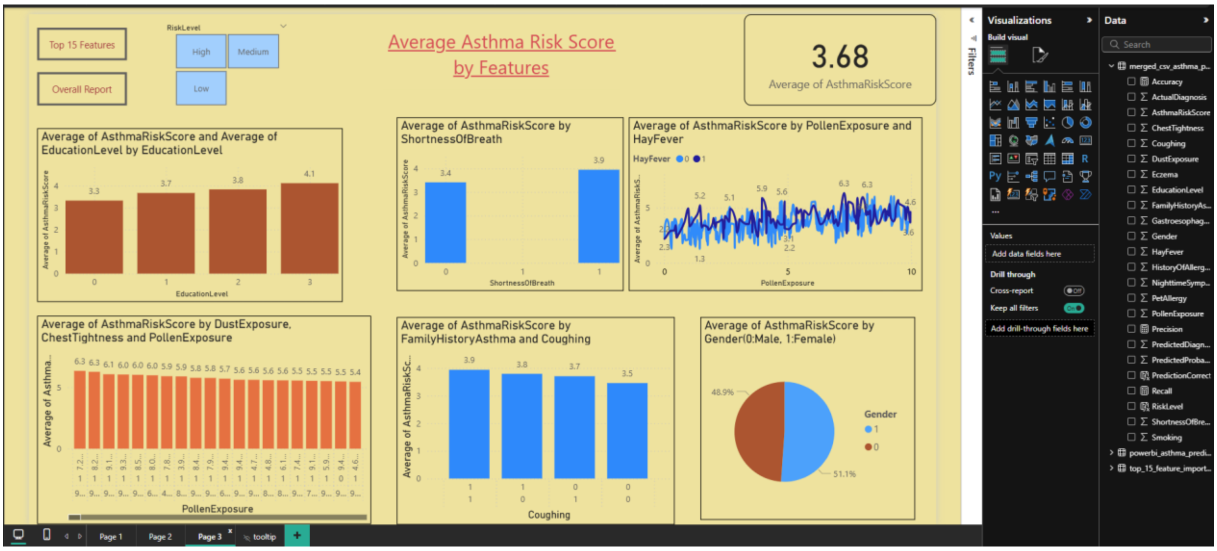
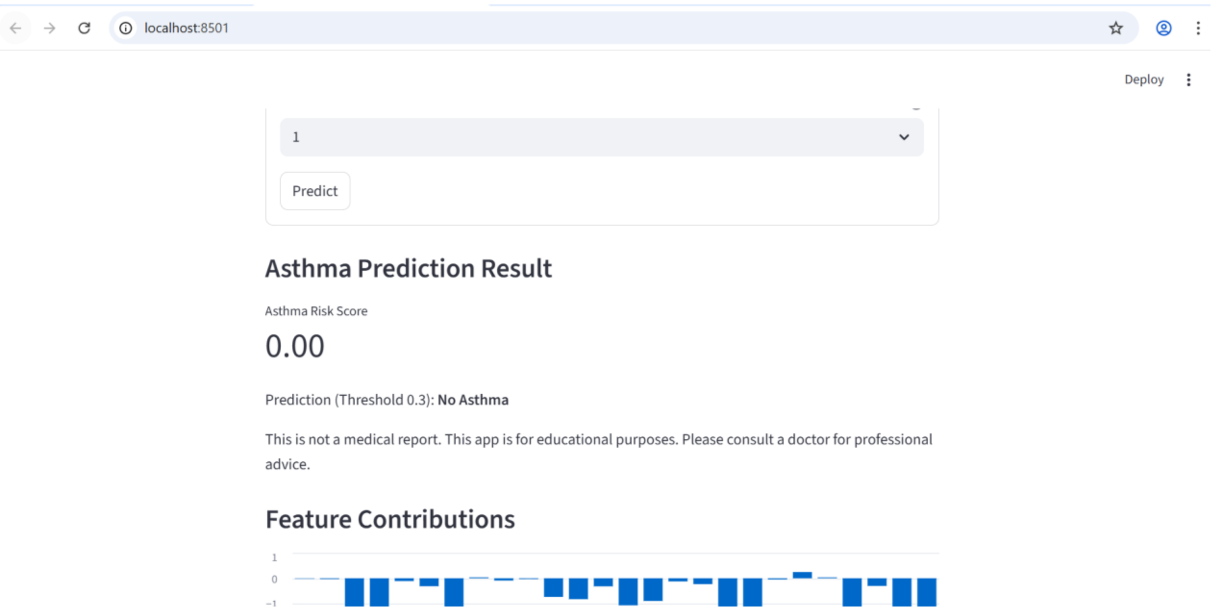


Image 4. ScreenShot taken from Power BI Dashboard average asthma RiskScore by features

6 Streamlit App Deploy

A streamlit web application was developed to make asthma prediction. This app allows to input patient details such as habits, environmental exposures, allergy history, clinical symptoms etc to get a real time asthma risk prediction.

It uses the selected model (Logistic Regression) and Standard Scaler saved from the pipeline and all the features (instead of top 15 features) that matching with original training data set. As a result it calculates and display the Asthma Risk Score based on model probability and give a prediction of asthma or no asthma. It also allows users to download their prediction result as CSV file.



Asthma Prediction Result

Asthma Risk Score

0.00

Prediction (Threshold 0.3): **No Asthma**

This is not a medical report. This app is for educational purposes. Please consult a doctor for professional advice.

Feature Contributions

Image:5 Screenshot taken from Streamlit app

7 Conclusion

This project demonstrated the application of machine learning technique to predict asthma diagnosis using a synthetic dataset. Through Preprocessing, class balancing with SMOTEEN, and threshold tuning, the models were optimized to prioritize recall – ensuring that true asthma cases were identified.

Among the two models tested, Logistic Regression was selected for its strong performance, achieving a recall of 80% and the highest ROC-AUC(0.670). Although its precision was low, resulting a high number of false positives which is acceptable in medical context. In healthcare, it is more critical to avoid missing true cases than to eliminate false alarms. Therefore, accuracy is less important than sensitivity.

The model outputs were integrated into a Power BI dashboard, which visualizes prediction, risk scores and feature importances. This dashboard makes the results accessible and interpretable for clinical decision support, allowing users to explore insights interactively and visually.

Although the data set was synthetic, the project lays a strong foundation for future work with real world clinical data.

In conclusion, this project highlights the potential of data science to support early asthma diagnosis and shows how technology can enhance medical decision making. In addition, the dashboard strengthens users to explore prediction visually and making complex results easier to understand for everyone.

Appendix A

Image 1: precision and recall plot from Jupyter Notebook across different threshold

Image 2: From Jupyter Notebook top 15 features

Image 3: Screenshot taken From Power BI Dashboard overall overview

Image 4: Screenshot taken from Power BI Dashboard average asthma RiskScore by features

Image:5 Screenshot taken from Streamlit app

References

https://www.geeksforgeeks.org/machine-learning