A7: LUNGVISION: Seeing the Future of Asthma Risk

Deepta Susan Beji | Ericsson A. David | Fabricio Montero Zuñiga
Group 1: Bioengineering
Prof. Hasan Tanvir

First steps of the project

Deadline: Monday, Dec 1, at noon (12:00)

The goal of this homework is to get you started with your project. Make sure that you know all the requirements for projects. Task 1 asks you to set up a repository. Tasks 2 and 3 are about the first two steps in CRISP-DM: business and data understanding. Task 4 asks you to make a work plan.

Task 1. Setting up (0.25 points)

- Create a project repository, either in GitHub or Bitbucket, if You have not done this already.
- Make sure that instructors have access to the repository. Invite us by using our usernames. If that does not work, invite us by our e-mail.
  - Markus Haug - haugmarkus (GitHub) and markushaug123@gmail.com (Bitbucket)
  - Kati Laidus - katilaidus (Github)
  - Heili Aavola - heiliaavola (GitHub)
  - Victor Pinheiro - victorhenriquecp (GitHub) and victorhenriquecp (Bitbucket)
  - Hasan Tanvir - hasantanvir79 (GitHub) and hasan.tanvir@ut.ee (Bitbucket)
- Add reports from the following subtasks (business understanding, data understanding and planning) to the repository as a single separate PDF file named GROUP_NR_report.pdf (e.g. "A0_report.pdf").
- Add the link of the repository also to the report.

Github Link: https://github.com/fabriciomz24IMIM/IDS-LUNGVISION.git

TASK 2
BUSINESS UNDERSTANDING

BACKGROUND

Currently there is no direct way of concluding if a person has asthma, and diagnosis for the condition is a combination of patient interviews, physical examination, or if there was a recorded attack prior (World Health Organization, 2013). Certain specialized tests can be done to confirm the subjective opinion of medical professionals, however these may be costly and are not a viable first-layer form of diagnosis for patients with asthma-like symptoms (National Institute for Care and Excellence, 2024).

Certain Biomarkers are available that have correlation to asthma but it is not clear whether these can predict if it's for childhood-onset asthma or adulthood-onset asthma. Both varieties of the pulmonary condition have very different approaches for treatment or care (Buhl et al., 2020). Additionally, there is the concept of “co-biomarkers”. This noted interaction from these biomarkers could be a good indication of the type and severity of asthma that an individual has (Lavere et al., 2025).

BUSINESS GOALS AND SUCCESS CRITERIA

Over the past years, numerous biomarkers have shown potential to be good diagnosis tools for asthma, prompting the need for a good analytical model to analyze them in a multi-variable study. This project aims to create the aforementioned model with hopes of integrating it as a potential initial front-line diagnostic tool in many healthcare systems. Specifically, the following goals could be applied to a future company or start-up with this advocacy in mind:

- To be able to create a model that can be used as a precursor for the early detection of asthma, preventing life-threatening attacks and avoiding unnecessary costly tests for unclear initial diagnosis.
- To be able to integrate the model into the healthcare system as one of the basic periodic tests done during annual physical examinations (APEs) and make it a general screening method both in school children and adults.
- To raise awareness about the complexity of asthma and how certain factors both physical and environmental may alter the prognosis of asthma into the later years of an individual
- To be able to identify specific biomarkers and create an easy rapid-test to analyze if a person has asthma and may categorize it as either childhood or adulthood-onset asthma.

The following success criteria can also be applied to determine whether the goals are achieved or are incipient to meeting them:

- **Accuracy** - *Is the model correctly predicting whether the patient has asthma and is able to put a category on its type or severity?*
- **Reliability** - *Can the model consistently predict accurate results and is it a good first-line diagnostic tool that eliminates ambiguity?*
- **Ease of Use** - *Can the model be widely used by medical or healthcare professionals and can easily be interpreted by the common individual?*
- **Acceptability** - *Is it Robust and Empirical enough to be widely accepted both by the general public and the medical field?*

- **Room for Improvement** - *Is the model easily trainable to handle improvements or new research regarding new biomarker data or the model can be extended into something more tangible like a product in the future?*

CURRENT SITUATION

Currently, the team has the following resources that they can use to fulfil the goals of the project.

- **NHANES Data Sets** - The National Health and Nutrition Examination Survey by the US Centers for Disease Control and Prevention is a survey updated every year and published quadrennially has numerous data sets which include the prevalence of asthma and as well other laboratory and bodily measurements that can be used to test for potential biomarkers that correlate to the prevalence of the condition.

- **GitHub Folder** - The team uses this as the master folder to keep all files related to the research and creation of the model. This also serves as the online digital collaboration platform between each member.

- **Literature for Verification of Biomarkers** - The Biomarkers selected to be screened were initially screened to minimize total run-times for the model. Research and papers were compiled by the team to serve as the basis for biomarker banking for testing.

With the current resources, the team now employs the following requirements, assumptions, and constraints within the project:

Requirements

- Use data from the NHANES dataset to create a training and validation model for the project to test whether the biomarkers described actually have significant statistical correlation with the onset of asthma.
- Clean and process the data to let the machine have a better “learning” experience in predicting the outcome.
- Analyze the biomarkers that are seen to have significant correlation and determine their interaction effects with other biomarkers analyzed.
- The output of the model should be easily understood and definitive so as to tell the next course of treatment for the patient.

Assumptions

- The data provided by NHANES is factual, timely, and correct and are from real individuals that have been interviewed.
- The individuals with a positive result for asthma in the data have been previously seen by a medical professional and have been confirmed for the diagnosis of asthma.
- The certain markers for analysis have been verified and done by a legitimate medical collection point.
- For lifestyle markers, it is assumed that these are average values given by the respondents and actual amounts vary, but not deviate largely from the reported value.

Constraints

- The NHANES data only account for respondents in the United States, hence this model will solely be learning from data from an isolated group, thus not take into account Race, Ethnicity, Religious Beliefs (Diets), and other external factors which may be considered for global applications

- Since the data set involves both children and private mental health data, we do not have permission to access some of those sensitive values thus the model would not take them into consideration, but nevertheless they should be analyzed as well.

Even though the development of the team's current model is a big leap for the medical field with regards to the integration of machine learning and diagnostics, it should be noted that there are certain factors which must be considered before making the model public for use.

Risks and contingencies

- Although the model would be standing as a frontline diagnostic tool, it must be noted that it should not be treated as the main confirmatory diagnostic. This tool shall only be used for initial screening and the resulting outcomes should be analyzed by a licensed medical professional for the next course of action to be taken.
- Since the model is being trained in a dataset within a specific country, there may be variability in outcomes when used for patients outside the country and may or may not give accurate results

Terminology

- Asthma - *Asthma is a chronic inflammatory disorder of the airways resulting in, variable airflow bronchial obstruction which is potentially reversible with appropriate therapy or spontaneously. It is typically characterized by episodic attacks of breathlessness, cough, and wheezing ("asthma triad") (World Health Organization, 2013).*

- Childhood-onset Asthma **-** *Asthma that occurs in children and includes different treatments catered towards them. Non-mitigation at this age may allow them to progress with the condition to adulthood (Trivedi & Denton, 2019).*

- Adulthood-onset Asthma - *Asthma that is prevalent in ages of 20 and above and usually more associated with lifestyle effects or are from remission as someone who had asthma previously (Trivedi & Denton, 2019).*

- Biomarkers - Characteristic indicators of biological or pathogenic processes, responsive to exposures or interventions *(FDA-NIH Biomarker Working Group, 2016).*

Costs and benefits

- Costs
  - Data Acquisition and Gathering
  - Global Conditions and Empricity
  - Continuous updating of data

- Benefits
  - More effective-evidence based first-line diagnostic tool
  - Elimination of costs in unnecessary tests
  - Widespread and easy wide-scale diagnosis of population
  - Early detection of Asthma in patients prior initial attack

RELATIONSHIP TO DATA-MINING

With the rise of the use of machine learning and AI in the medical field, the team hopes to be a bridge between the usage of modern technology, data analysis, and modern medicine to create a solution to progress the healthcare scenario. In general, the team has the following goals and criteria on how the data would be used, processed, and interpreted:

Data-Mining Goals

- To be able to gather required data and perform pre-analysis treatment prior to utilizing the data to create the model
- To be able to observe the correlation between potential biomarkers with the prevalence of Asthma.
- Discover certain cross-variable interactions between the biomarkers to determine possibility of co-biomarker/morbidity effects that influence asthma prevalence or severity
- Train the model to determine whether an individual has asthma based from certain biomarker data
- Configure the model to determine whether positive asthma patients are childhood-onset or adulthood-onset variants

Data-Mining Success Criteria

- **Accuracy** - The data was properly cleaned and the model was trained to give out good accurate results once verified in the testing data
- **Reliability** - The model is able to adapt to all types of possible inputs and data for analysis and adjusts the results accordingly with regards to proper model training
- **Ease of Use** - Output data is easy easy to understand by all individuals and is able to be used in first-line diagnosis
- **Specificity and Decreased Ambiguity** - Output prediction is empirical and leaves no to little room for ambiguity as it would be used for the direction of treatment moving forward.

REFERENCES

Buhl, R., Korn, S., Menzies-Gow, A., Aubier, M., Chapman, K. R., Canonica, G. W., Picado, C., Donica, M., Kuhlbusch, K., Korom, S., & Hanania, N. A. (2020). Prospective, single-arm, Longitudinal Study of biomarkers in real-world patients with severe asthma. *The Journal of Allergy and Clinical Immunology: In Practice*, *8*(8). https://doi.org/10.1016/j.jaip.2020.03.038

FDA-NIH Biomarker Working Group. (2016). BEST (biomarkers, EndpointS, and other tools) resource [internet].

Lavere, P. F., Phillips, K. M., Hanania, N. A., & Adrish, M. (2025). Established and emerging asthma biomarkers with a focus on biologic trials: A narrative review. *Journal of Personalized Medicine*, *15*(8), 370. https://doi.org/10.3390/jpm15080370

National Institute for Care and Excellence. (2024). *Asthma: diagnosis, monitoring and chronic asthma management (BTS, NICE, SIGN)* . National Institute for Care and Excellence.

Trivedi, M., & Denton, E. (2019). Asthma in children and adults—what are the differences and what can they tell us about asthma? *Frontiers in Pediatrics*, *7*. https://doi.org/10.3389/fped.2019.00256

World Health Organization. (2013). *Diagnosis and Management of Asthma: Primary health care service Delivery Iraq 2013 guidelines*. Iraq; World Health Organization.

TASK 3
DATA UNDERSTANDING

GATHERING DATA

Outline data requirements

The goal of this project is to build a machine-learning model that predicts whether an individual's asthma is childhood-onset or adult-onset based on demographic, behavioral, clinical, and laboratory characteristics. To support this goal, the required data must include:

1) Asthma variables indicating diagnosis, age of onset, current asthma status, recent symptoms, and relevant family history.
2) Sociodemographic data such as age, sex, race/ethnicity, insurance coverage, and general health.
3) Behavioral and environmental factors, including physical activity, tobacco use, alcohol consumption, and household smoking exposure.
4) Clinical and biomarker measures, including BMI, blood pressure, metabolic markers, pulmonary-related labs, and systemic inflammation indicators.
5) Identifiers enabling linkage across NHANES components.

Verify data availability

NHANES provides publicly accessible datasets covering demographics, medical conditions, physical activity, tobacco use, environmental exposures, lab tests, and health behaviors. All variables required for this study were available across multiple XPT modules. Data dictionaries confirmed that all selected variables exist and are well-documented. The raw SAS transport files were successfully downloaded for all target cycles and converted into CSV format to ensure compatibility with Python-based data-mining tools. All files included a unique respondent identifier (SEQN), enabling reliable merging.

Define selection criteria

Selection focused on NHANES participants who completed interviews and examinations. Only variables relevant to asthma onset prediction were retained. This included:

- Demographic fields (age, gender, race/ethnicity, etc.).
- Health insurance and self-reported health status.
- Tobacco use frequency, exposure, and family smoking history.
- Physical activity measures.
- Asthma-related clinical history.
- Chronic disease indicators.
- Anthropometric variables (BMI, weight, height).
- Laboratory measures associated with inflammation, metabolism, or respiratory function.
- Variables not related to the prediction goal (dental, dietary recall, agricultural exposures, etc.) were excluded.

DESCRIBING DATA

After selection and merging, the dataset contained all respondents from the 2017–2020 NHANES cycles. Each row corresponds to one individual, and each column represents a demographic, behavioral, clinical, or laboratory characteristic. All variables were renamed with descriptive human-readable labels to simplify interpretation. The same procedure was performed for the 2015-2016 dataset.

The dataset includes:

- Demographics.
- Health behaviors including alcohol intake, smoking history, tobacco product use, and physical activity.
- Asthma history including diagnosis, age of onset, attacks, emergency visits, and family asthma history
- Chronic conditions such as heart disease, thyroid disorders, liver conditions, and cancer.
- Anthropometrics including weight, height, BMI and birth-related measures.
- Laboratory biomarkers including metabolic markers, liver enzymes, electrolytes, inflammatory markers, lipids, and blood counts.

Descriptive statistics show that most demographic variables were complete and follow expected NHANES distributions. Laboratory measures exhibit natural biological ranges, though some contain missing values due to NHANES subsampling.

EXPLORING DATA

Exploration involved summary statistics, frequency distributions, and checks for outliers. Age and gender distributions aligned with typical NHANES population-level sampling. Tobacco use and physical activity variables displayed expected variation across age groups. Some laboratory values showed right-skewness, which is consistent with biological measures.

Asthma variables revealed a mix of childhood-onset and adult-onset cases, enabling supervised learning.

During exploration, extremely small floating-point artifacts ($5.3976e^{-79}$) appeared in certain fields and were determined to be errors caused during XPT-to-CSV conversion. These values were treated as 0.

VERIFYING DATA QUALITY

Data quality checks revealed:

- Missing values in several lab variables due to NHANES subsampling designs; these will require imputation or exclusion during modeling.
- Consistent respondent IDs across all tables, ensuring valid merges.
- No duplicated respondents after merging.
- Outlier checks indicated biologically implausible values only where the floating-point artifact occurred; these were corrected.
- Variable meaning validated using NHANES data dictionaries to ensure proper interpretation.

Overall, the dataset is of high quality and fully suitable for modeling after routine cleaning steps (handling missing values, normalization, and type conversion).

TASK 4

PLANNING THE PROJECT

Task 1: Data Acquisition and Preprocessing (18 hours)

- Fabricio (15h): Download NHANES datasets (2017-2023), merge tables, label asthma patients by onset age, handle missing values, remove outliers
- Deepta (2h): Review data structure and quality issues
- Eric (1h): Define evaluation metrics (accuracy >75%, precision, recall, F1, AUC-ROC)

Task 2: Feature Selection and Statistical Testing (12 hours)

- Fabricio (10h): Extract features, perform t-tests and chi-square tests, calculate correlations, keep features with $p<0.05$, document significance
- Eric (2h): Review statistical methodology, suggest demographic stratifications

Task 3: Feature Engineering and Pipeline (10 hours)

- Fabricio (8h): Encode categorical variables, normalize continuous features, build preprocessing pipeline, create 80/20 train-validation split
- Deepta (2h): Validate pipeline, verify data distributions, check for leakage

Task 4: Model Development and Training (20 hours)

- Deepta (18h): Build and tune four models: Logistic Regression (tune regularization), K-Nearest Neighbors (optimize K value and distance metrics), Random Forest (tune trees, depth, features via grid search), Neural Network (design architecture, tune learning rate and dropout). Compare models, ensure >75% accuracy, document hyperparameters
- Eric (1h): Track model performance
- Fabricio (1h): Monitor data quality during training

Task 5: Temporal Validation on Test Set (14 hours)

- Eric (12h): Test models on 2015-2016 data, calculate all metrics, create confusion matrices, perform significance testing, compare test vs validation performance
- Deepta (1h): Adjust models if needed
- Fabricio (1h): Verify preprocessing consistency

Task 6: SHAP Explainability Analysis (16 hours)

- Eric (14h): Generate SHAP values, create feature importance visualizations, stratify by demographics (age, sex, race, income), identify modifiable risk factors
- Deepta (2h): Support SHAP integration

Task 7: Documentation and Reporting (14 hours)

- Eric (6h): Discussion, implications, limitations, presentation slides, executive summary

- Deepta (5h): Model methodology, results, technical appendix
- Fabricio (3h): Data sources, methods, feature selection documentation

Methods and Tools

Python with Pandas, NumPy, scikit-learn, TensorFlow, SHAP, Matplotlib, Seaborn. Data from CDC NHANES (2017-2020 training, 2015-2016 testing).

Key Clarifications

Temporal validation proves models work on future patients. Feature testing ensures only significant predictors are included. SHAP analysis identifies demographic-specific and modifiable risk factors for clinical application.

Total Hours: Fabricio: 38h | Deepta: 31h | Eric: 33h