Predicting Respiratory Diseases by the Patient’s Breathing Sounds

Team Members: Kaarel Tamuri, Stina-Marie Maripuu
Github: https://github.com/KaarelTamuri/RespiratoryDiseases/

Task 2. Business understanding (0.5 point)

1. Business Understanding

Background

Set within the healthcare industry, this project focuses on supporting the diagnostic process for respiratory diseases. It aims to harness the potential of machine learning to analyze breathing sounds, to help cure patients more effectively and reliably.

Business Goals

The key goal is to develop a machine learning model that assists doctors in identifying respiratory diseases from patients' breathing sounds. The objective is to supplement the traditional diagnostic methods, providing a data-driven tool that could potentially enhance the speed and accuracy of diagnoses.

Business Success Criteria

The project will be deemed successful if it proves to be a reliable and useful tool for doctors in their routine diagnostic work. The model's effectiveness will be assessed by its adoption rate in clinical settings and its impact on improving the diagnostic process.

2. Assessing the Situation

Inventory of Resources

Resources available for this project include a comprehensive dataset with recordings of patients' breathing, alongside their demographic information and medical diagnoses. The project team comprises two data scientists, equipped with Python and Jupyter for development.

Requirements, Assumptions, and Constraints

A one-week deadline imposes a significant time constraint on the project.
The team faces challenges due to limited knowledge in medical science and sound analysis.
Computational resources for model training are limited, which could restrict the complexity and scale of the model.

Risks and Contingencies

There's a risk that the project may not achieve a fully functional model within the stipulated time frame. In this event, the team plans to adapt the project scope if necessary, possibly focusing on a simpler model based on textual data if sound file analysis proves too challenging within the time constraints.

Terminology

Respiratory Disease: This encompasses conditions affecting the respiratory system, including chronic and acute illnesses.

Costs and Benefits

Costs:

The project will involve about 60 hours of work, with no direct financial costs but a considerable investment of time and effort.

Benefits:

The potential benefits include providing doctors with a tool that could make the diagnostic process for respiratory diseases more efficient and accurate. For the team, it offers valuable experience in applying machine learning techniques in a healthcare context.

3. Defining Data-Mining Goals

Data-Mining Goals

The project aims to identify distinctive patterns in breathing sounds that correlate with specific respiratory diseases. This involves analyzing audio data to find signatures that can be linked to known conditions.

Data-Mining Success Criteria

The success of the data-mining aspect will be measured by the model's accuracy and its recall rate. A high recall rate is particularly important to ensure no potential disease cases are overlooked, as doctors will rely on the model to flag possible conditions for further investigation.

Project Beneficiaries

The main beneficiaries are doctors and patients in healthcare settings. Doctors will benefit from having an additional tool to aid in the diagnosis of respiratory diseases, potentially leading to more accurate and timely diagnosis. Patients stand to gain from improved diagnostic accuracy, which can lead to better treatment outcomes. In a broader sense, the project also contributes to the field of medical technology, demonstrating the application of machine learning in enhancing healthcare diagnostics.

Task 3. Data understanding (1 points)

Gathering data:

Data requirements

The data needed for this project is health data regarding respiratory diseases. The main goal is to correctly diagnose a respiratory disease based on the respiratory sounds of the patient. For that to be doable, we need resiratory sounds and patient info regarding their health condition. The dataset has to include multiple diseases, but also healthy patients.

Data availablility

This data is available on Kaggle, named Respiratory Sound Database: https://www.kaggle.com/datasets/vbookshelf/respiratory-sound-database/data . From the database we have access to 920 .wav sound files and additionally 920 annotated .txt files that describe the .wav files. We have a table listing diagnosis for each patient. Finally, we have a table containing demographic information for each patient.

Selection criteria

From those files we will be using the .wav sound files to predict respiratory diseases. We are going to try to predict it directly using the sound files and not the annotated .txt files, but these are a backup. We also need to use the diagnosis data and also the gender and age of the patient to find any possible correlations.

Describing data:

.wav files

Starting with the .wav audio files. The database includes 920 .wav files, for almost every patient, there are 7 different sound files. The audio files are of varying length, 10s to 90s. The different sound files are all recordings from different parts of the chest, like Trachea (Tc), Anterior left (Al), Anterior right (Ar), Posterior left (Pl), Posterior right (Pr), Lateral left (Ll), Lateral right (Lr). These different chest location help us achieve our second goal of trying to find the best method to record breathing for the most accurate prediction. For every sound file there is also information about the recording device, for example if it was a sequential/single channel or simultaneous/multichannel recording device, meaning we can see if there is a difference between the general quality of the number of channels in recordings. Lastly, the recordings also include information regarding what specific equipment was used, like AKG C417L Microphone, 3M Littmann Classic II SE Stethoscope, 3M Litmmann 3200 Electronic Stethoscope, WelchAllyn Meditron Master Elite Electronic Stethoscope. Getting the recording device information also helps us analyse if there even is a best way to record breathing. In addition the sound files also have information regarding the patients ID.

.txt annotation files

In addition to the sound files, the database also includes annotated .txt files of all the files mentioned above. These files are helpful in making it easier for the computer to read the files since they describe the contents of the sound files. Annotation files include. Start and end time of a respiratory cycle in seconds, meaning inhale and exhale. Presence or absence of crackles, where presence = 1 and absence = 0. Crackles are intermittents sounds that are often audible during inhalation and may sound similar to bubbling and popping noises (https://www.medicalnewstoday.com/articles/lung-sounds#rhonchi). Presence or absence of wheezes, where presence = 1 and absence = 0. Wheezing is described as a high-pitched, musical, continuous sound (https://www.medicalnewstoday.com/articles/lung-sounds#wheezing).

Diagnosis information

Every patient has an ID and in this table, the ID relates to the patient's diagnosis. Diagnosis include: Healthy, Chronic Obstructive Pulmonary Disease - COPD, Upper Respiratory Tract Infection - URTI, Bronchiectasis, Pneumonia, Lower Respiratory Tract Infection - LRTI.

Demographic information

The whole dataset includes 126 patients. Regarding the demographic information, we have the patient ID, Age, Sex, Adult BMI (kg/mg2), Child Weight (kg), Child Height (cm). The patients span all groups - children, adults and elderly. The reason the demographic table includes child weight and height is because BMI is not an accurate enough of a condition to represent the children's healthiness.

Exploring data:

Demographic table:
Age:

The average age in the dataset is approximately 43 years, with a median (50th percentile) age of 60 years. The age range spans from a minimum of 0.25 years to a maximum of 93 years, highlighting a diverse age distribution.

BMI (Body Mass Index):

The mean BMI is around 27.19, with a median of 27.4. The BMI values vary between 16.5 and 53.5, indicating a broad spectrum of body mass index measurements in adults.

Weight:

The average weight is 21.36 kg, with a median of 15.1 kg. The weight range extends from a minimum of 7.14 kg to a maximum of 80.0 kg, showcasing the diversity in children's weights.

Height:

The mean height is 104.65 cm, and the median is 99.5 cm. Heights range from a minimum of 64.0 cm to a maximum of 183.0 cm, showcasing the diversity in children's heights.

Sex:

79 male entries and 46 female entries.

In conclusion:

The mean and median BMI are a little concerning, as a normal BMI should fall into the range 18.5 to 24.9. The oldest “child” in this dataset is aged 16. There are a few too many NA-s in the dataset as 7 of 126 patients don’t have a weight+height combination or a BMI. The only issue with the data seems to be that the male and female ratio is not a perfect 50%.

Diagnosis table:

51% of the patients’ diagnosis is COPD, 21% are healthy and 28% have a different disease diagnosis.

Verifying data quality:

The data seems to be of good quality and it seems to represent the patients well. The data should be enough to achieve the project’s goal.

Task 4. Planning your project (0.25 points)

Project Plan

Clean and Prepare a Unified Database

Duration: 8 hours
Team Members: Kaarel Tamuri (2 hours), Stina-Marie Maripuu (6 hours)
Tools and Methods: Python for data cleaning and preprocessing.

Find out how recognize .WAV Sound Files

Duration: 10 hours
Team Members: Kaarel Tamuri (8 hours) Stina-Marie Maripuu (2 hours)
Tools and Methods: Python libraries for audio file processing.

Model to Analyze Breathing Sounds

Duration: 12 hours
Team Members: Kaarel Tamuri (10 hours), Stina-Marie Maripuu (2 hours)
Tools and Methods: Possibly using libraries like Scikit-learn, to analyze breathing length, crackles, and wheezes.

Predictive Model for Respiratory Diseases

Duration: 12 hours
Team Members: Kaarel Tamuri (3 hours), Stina-Marie Maripuu (9 hours)
Tools and Methods: Building on the previous task, using machine learning algorithms to correlate breathing sound analysis with respiratory disease diagnosis.

Create a Project Poster

Duration: 8 hours
Team Members: Kaarel Tamuri (4 hours), Stina-Marie Maripuu (4 hours)
Tools and Methods: Design tools like Canva