Task 3: Data Understanding within CRISP-DM

Gathering Data

Outline Data Requirements

The project aims to predict the five-year death probability in patients already diagnosed with lung cancer. Therefore, the data required would include patient demographics, medical history, details of the lung cancer diagnosis (stage, type, etc.), treatment details, and survival data.

The data requirements for this project are determined by the problem at hand - predicting the five-year death probability in lung cancer patients. Therefore, we need data that includes patient demographics (age, gender, etc.), medical history (comorbidities, previous treatments, etc.), details of the lung cancer diagnosis (stage, type, etc.), treatment details (surgery, chemotherapy, radiation therapy, etc.), and survival data (date of death if applicable, survival time, etc.).

Verify Data Availability

We plan to use synthetic data generated using Synthea, a synthetic patient population generator. Synthea can generate realistic synthetic patient data, including demographics, medical history, and detailed health records. Synthea can generate realistic synthetic patient data, including demographics, medical history, and detailed health records, making it a suitable source for our data needs.

Define Selection Criteria

The selection criteria for this project would be patients diagnosed with lung cancer. We would filter the synthetic data generated by Synthea to include only these patients. This ensures that our dataset is relevant to the problem we are trying to solve.

Describing Data

The dataset would include various fields related to the patients’ demographics (age, gender, etc.), medical history (comorbidities, previous treatments, etc.), lung cancer diagnosis (date of diagnosis, stage, type, etc.), treatment details (surgery, chemotherapy, radiation therapy, etc.), and survival data (date of death if applicable, survival time, etc.). Each field in the dataset would be described in detail, including its type (numerical, categorical, date, etc.) and the meaning of its values. This step is crucial for understanding the structure of the data and how different fields relate to the problem at hand.

Exploring Data

Exploratory data analysis would be performed to understand the distribution of each variable, identify outliers, and discover relationships between variables. This would involve calculating descriptive statistics, creating visualizations, and performing correlation analysis. For example, we might explore the relationship between the stage of lung cancer at diagnosis and the five-year survival rate. This step helps us gain insights into the data and generate hypotheses for further analysis.

Verifying Data Quality

The quality of the synthetic data would be verified by checking for missing values, duplicate records, inconsistent records, and outliers. Any issues identified would be addressed appropriately. For example, missing values might be imputed using appropriate strategies, and outliers might be investigated to determine if they represent errors or genuine extreme values.

As a result of the Data Understanding phase, we would have a dataset of synthetic patients diagnosed with lung cancer, understood the meaning of all fields in the dataset, and decided which parts of the data to use for the machine learning model. Some data cleaning might be performed during this task to facilitate data understanding, even though data cleaning is typically part of the data preparation step in CRISP-DM. The next steps would involve preparing the data for machine learning, developing the predictive model, and evaluating its performance.