Project Title: Survival Prediction for Lung Cancer Patients

Team Members:

Andrei Tsistjakov

Kennar Kahju

Oliver Puusalu

Task 2: Business Understanding Report Identifying Your Business Goals

Background

Lung cancer is a severe and prevalent disease with a significant impact on global health. Our project aims to leverage synthetic data to predict the survival of lung cancer patients over the next five years. Early detection and accurate survival predictions can guide medical professionals in tailoring treatment plans, improving patient outcomes, and optimizing healthcare resource allocation.

Business Goals

- Develop a predictive model for lung cancer survival using synthetic data.
- Enhance the understanding of factors influencing survival rates in lung cancer patients.
- Contribute to advancements in personalized medicine for lung cancer patients.

Business Success Criteria

- Achieve a predictive accuracy of at least 80% in survival predictions.
- Generate insights into the most influential features affecting survival rates.

Assessing Your Situation

Inventory of Resources

Access to synthetic data simulating lung cancer patient profiles.

Expertise in data science, machine learning, and medical domain knowledge within the team.

Requirements, Assumptions, and Constraints

Requirement - Ensure compliance with ethical guidelines in handling patient-related data.

Assumption - Synthetic data accurately reflects real-world lung cancer patient characteristics.

Constraint - Limited access to certain medical databases and real patient data due to privacy and legal considerations.

Risks and Contingencies

Risk - Inaccuracies in synthetic data may impact model performance.

Contingency - Conduct thorough validation and sensitivity analyses to address data quality concerns.

Terminology

Define medical terms and abbreviations to ensure clear communication between data scientists and healthcare professionals.

Costs and Benefits

Costs

Time and Effort - The primary cost may involve the time and effort invested by the team members. Developing a robust predictive model and conducting thorough analyses require dedication and expertise.

Computational Resources - While not necessarily a monetary cost, the use of computational resources, including hardware and software, could be considered a resource investment.

Benefits

Knowledge Advancement - The project contributes to the advancement of knowledge in oncology by identifying key features influencing lung cancer survival. This intellectual benefit adds value to the scientific community.

Skill Development - Team members gain valuable experience and expertise in the intersection of data science and healthcare. The skills acquired can have long-term benefits for both individual team members and the organizations they are associated with.

Social Impact - Although not directly financial, the societal benefit of improving patient outcomes and contributing to personalized medicine for lung cancer patients is substantial. The positive impact on public health and well-being is a meaningful benefit.

Collaboration Opportunities - The project may open avenues for collaboration with healthcare professionals, researchers, and institutions. Building partnerships can lead to future collaborative projects and opportunities.

Defining Your Data-Mining Goals

Data-Mining Goals

- Build a robust predictive model for lung cancer survival.
- Identify key features influencing survival outcomes.
- Evaluate model performance and refine as needed.

Data-Mining Success Criteria

- Achieve a high area under the ROC curve (AUC) for the survival prediction model.
- Uncover actionable insights into the significance of various patient features on survival.
- Demonstrate the model's generalizability through rigorous testing and validation.

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 will 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.

Task 4: Planning your project

Tasks

- Add head rows to tables – Adding head rows to the tables is necessary before merging the data. The tables lack head rows and must be manually added from the Synthea GitHub wiki page[1]. This task holds utmost importance as the data merging process cannot be accomplished without it. On average, one person can complete this task in approximately 3 hours.
- Feature recognition – Extract the pertinent characteristics from the data, isolate them from extraneous information, and amalgamate them into a cohesive data file. It is endorsed that this task is performed collaboratively with the entire team to ensure critical features are not unnoticed. This process may take up to 10 hours per person as a group.
- Cleaning the data – We need to run several tests to identify any outliers, which we can address accordingly. We must also check for errors or duplicate entries that could affect the final result. This task might be somewhat challenging, mainly when dealing with many relevant features or a large amount of data. Also, we have to convert specific values into dummy variables that can be used in machine learning. We estimate that this process will take approximately 5 hours per person.
- Build a model – Building different machine learning models and identifying the best fit for our needs is an important task. It usually takes around 6 hours per person and is best done in a group. We can leverage team members' diverse skills and perspectives to find the most optimal solution by working collaboratively.
- Report the findings – To ensure that the results from different models are organised and easy to access, gathering them in a separate document or file is recommended. This task should be done during the model-building task, which should take 1 hour. Doing so can save time and avoid confusion when analysing the results later.
- Add a readme.md – Once the project is complete, it is crucial to create a readme.md file and upload it to GitHub. This file should provide a clear and concise explanation of the project and instructions on replicating any findings. One person should be allocated around 2 hours to create this document, ensuring it is comprehensive and informative.
- Design the poster – Design a visually appealing poster and include all relevant information. Working in a team should take around four hours to complete.

Methods and Tools

- The Jupyter Notebook is used for the central part of the model building.
- A tool such as Canvas can be used to create the poster.
- Many methods we will use will probably come from the Pandas package. We will likely utilise various techniques from the Pandas package.

[1] https://github.com/synthetichealth/synthea