Project F2: Predictive Model LCSF1

Lung Cancer Survival Forecasting for 1 year

Team Member: **Shao Ci Weng**
Teeli Kuri
Jan Petersen

Link to GitHub: https://github.com/OJPOJ/ItDS-project

Task 2. Business understanding

- Identifying your business goals
    - Background

The project aims to harness synthetic medical data for predicting 1-year mortality in lung cancer patients, thus addressing a critical need in healthcare for accurate and early patient prognosis.

    - Business goals

Our team's goal is to aid healthcare providers in making informed decisions by accurately forecasting patient outcomes, improving treatment planning, and potentially enhancing patient survival rates.

    - Business success criteria

Our team project's success will be evaluated based on the predictive model's accuracy, specifically achieving an AUC-ROC value above 0.75, and its applicability in clinical settings. Additionally, whether we were able to identify some key observations, drugs, treatments, etc. that influence the survival rate of the patients.

- Assessing your situation
    - Inventory of resources

Dataset: Our project uses a synthetic dataset that includes patient IDs, treatment IDs, and treatment times. This data is critical for training and validating our predictive model.
Technical resources: Advanced platform and libraries for developing, training, and evaluating machine learning models. Moreover, tools and software for data cleaning, transformation, and analysis.
Human resources: Professionals skilled in data analysis, machine learning modelling, also need to cooperate with experts in the field of oncology to validate the clinical relevance of the model, and help in interpreting the data in a healthcare context.

    - Requirements, assumptions, and constraints

1. **Requirements:** Completion within a set timeframe and compliance with data privacy standards.
2. **Assumptions:** The synthetic dataset accurately reflects real-world scenarios, and the findings will be generalizable to actual world data.
3. **Constraints:** Resource limitations, particularly in computing power and data access.

- Risks and contingencies

1. **Risk:** Potential overfitting of the model, data inaccuracies.
2. **Contingencies:** Implementation of rigorous validation procedures, continuous engagement with team members and healthcare experts for model relevance, and ensuring a robust data cleaning process.

- Terminology

Developing a shared understanding of key terms in both data science and oncology is essential for effective communication among our team members and healthcare professionals.

- Costs and benefits

1. **Costs:** Time and resources for data processing, model development, and testing.
2. **Benefits**: Improved patient care and potential insights for future medical research.

- Defining your data-mining goals
  - Data-mining goals

1. To develop a predictive model that can accurately forecast 1-year mortality in lung cancer patients based on treatment data.
2. To achieve and surpass an AUC-ROC value of 0.75.
3. Finding some treatments, observations, etc. which can be connected to mortality in one year.

- Data-mining success criteria

Our team will try different models. Splitting the data into training and testing sets, with the model being trained on the training set and evaluated on the testing set. The model's performance will be measured quantitatively using the AUC-ROC, Accuracy, Precision, and Recall metrics. Moreover, success will also be evaluated based on the model's clinical relevance and applicability.

Task 3. Data understanding

- Gathering data
  - Outline data requirement

The project necessitates comprehensive records of lung cancer patients. The dataset forms the backbone of the analysis, enabling a detailed exploration of treatment trajectories and patient outcomes. Moreover, it should encompass a wide range of medical interventions and the patient survival status of these treatments over time.

- Verify data availability

The dataset has been provided by the responsible lecturer, granting us access to the essential lung cancer data for this project. Its availability ensures that our team has a reliable foundation of information to base our analysis and modelling.

- Define selection criteria

1. **Relevance:** The dataset's focus should be squarely on lung cancer treatments, particularly within the first year post-diagnosis. This period is critical for understanding the initial response to treatments and the early progression of the disease.
2. **Recency:** It is crucial that the dataset accurately reflects the current treatment methods and practices in oncology. This ensures that the insights gained and models developed are applicable to nowadays clinical settings and patient care strategies.

- Describing data

SUBJECT_ID: Each patient is assigned a unique identifier, allowing for the tracking of individual treatment journeys. This ID is essential for maintaining patient confidentiality while enabling detailed analyses at the individual level.
DEFINITION_ID: Type of medical intervention.(Reflects the nature of the medical interventions, categorized into various types like drugs, procedures, conditions, etc.)
TIME: The timeline is crucial for understanding the sequence and timing of treatments, providing insights into how treatment strategies evolve over the course of the disease (Starting from the time of diagnosis).

- Exploring data

1. Utilizing descriptive statistics, the distribution, and variation in treatment types and their timings will be thoroughly examined. This includes analysing the frequency of different treatments and the typical timelines for their administration.
2. Identify prevalent treatment patterns and their relationship to patient outcomes.
3. This represents the year when each medical event occurred, starting from the time of diagnosis. The timeline is crucial for understanding the sequence and timing of treatments, providing insights into how treatment strategies evolve over the course of the disease.

- Verifying data quality

Ensuring high data quality is critical to the credibility and usefulness of our project results. We must accurately and consistently classify the type of medical intervention indicated by DEFINITION_ID in the dataset. This means that each intervention should be properly labelled and grouped to reflect the true nature of the treatment it represents.
Secondly, despite its synthetic dataset, it must closely resemble real-world lung cancer treatment scenarios. This includes checking whether the type of treatment, the frequency of use, and the order in which it is given to patients are similar to what would occur in actual medical practice. This consistency with real-world data ensures that findings and models developed from this dataset are relevant and applicable to real-world clinical situations, thereby increasing the overall value and applicability of the project.

Task 4. Planning your project

Plan Outline: (Proposal)

1. Data Analysis

Task Description: Analyse the synthetic dataset to understand the distribution and variation of treatment types and timelines.
Methods/Tools: Descriptive statistics, data visualization tools (e.g., Python libraries like Matplotlib, Seaborn).
Hours per Team Member: Shao Ci Weng - 5 hours, Teeli Kuri - 5 hours, Jan Petersen - 5 hours.

2. Data Refactoring

Task Description: Refactor and clean the dataset to ensure consistency and accuracy in the treatment data.
Methods/Tools: Data cleaning tools (e.g., Python Pandas).
Hours per Team Member: Shao Ci Weng - 2 hours, Teeli Kuri - 2 hours, Jan Petersen - 2 hours.

3. Feature Engineering

Task Description: Develop and select relevant features from the dataset that could influence the survival rate.
Methods/Tools: Feature selection techniques (e.g., principal component analysis, Lasso and Ridge regression), iterative methods, one-hot or statistical methods.
Hours per Team Member: Shao Ci Weng - 3 hours, Teeli Kuri - 3 hours, Jan Petersen - 3 hours.

4. Splitting the training data for model Assessment

Task Description: Divide the dataset into training and testing sets, ensuring a balanced representation of patient IDs.
Methods/Tools: Data splitting techniques (e.g., sklearn's train_test_split).

5. Model Training

Task Description: Train different models to predict 1-year mortality in lung cancer patients.
Methods/Tools: Machine learning algorithms (e.g., Random Forest, Logistic Regression, AdaBoost, XGBoost, Naive Bayes, Support Vector Machines, etc).
Hours per Team Member: Shao Ci Weng - 10 hours, Teeli Kuri - 10 hours, Jan Petersen - 10 hours.

6. Model Assessment

Task Description: Validate the models using cross-validation and compare their performance with the business goals.

Methods/Tools: Cross-validation techniques, performance metrics (e.g., AUC-ROC).
Hours per Team Member: Shao Ci Weng - 5 hours, Teeli Kuri - 5 hours, Jan Petersen - 5 hours.

7. Documentation and Reporting

Task Description: Document the project findings, model insights, and recommendations. Add all the code, as well as readme files describing what the code does and how to run it on GitHub. Prepare a visually appealing and understandable poster.
Methods/Tools: GitHub, documentation tools, data visualization for the poster.
Hours per Team Member: Shao Ci Weng - 5 hours, Teeli Kuri - 5 hours, Jan Petersen - 5 hours.

Additional Comments:

Revision and Feedback: Hold regular team meetings for progress review and feedback.
Flexibility: Be prepared to adjust tasks and hours based on ongoing findings and team discussions.