Research Project Proposal

Yuval Arbel and Itamar Cohen

Table of Contents

1. Project's Goal 2
1.1 What element in the DS pipeline are you trying to improve? 2
1.2 Why does the chosen element need improvement? 2
1.3 What are the desired results? 2
1.4. What is the relation/connection to the material we learn in class? 3
2. The Solution 4
2.1 Initial Ideas for a Solution 4
2.2 Propose Ways to Measure the Solution 4
3. Related Work 5
4. Experiments Plan 6
4.1 Datasets 6
4.2 Testing Plan 6

TDS Project: Part 3 - Research Project

1. Project's Goal

1.1 What element in the DS pipeline are you trying to improve?

This project focuses on **enhancing Pattern Mining techniques** for lung cancer severity prediction. Specifically, we aim to apply **Frequent Pattern Mining and Association Rule Mining** to uncover hidden relationships between symptoms, lifestyle choices, genetic predispositions, and cancer severity levels.

Current predictive models rely heavily on direct correlations, but **hidden multi-variable relationships** could provide **new insights into disease progression and risk factors.**

1.2 Why does the chosen element need improvement?

Pattern mining in medical datasets is **underutilized**, especially for conditions like lung cancer where multiple factors interact in complex ways. Traditional statistical methods often fail to capture these relationships.

By **improving the way patterns are detected and analyzed**, we aim to:

- **Discover critical symptom combinations** that strongly indicate disease severity.
- **Refine predictive models** by generating **new pattern-based features** that improve accuracy.
- **Support medical decision-making** by identifying **actionable insights** into patient risk levels.

1.3 What are the desired results?

- **Identify Hidden Symptom Patterns**: Detect recurring symptom and lifestyle factor combinations that frequently appear in patients with different lung cancer severity levels.
- **Enhance Feature Engineering**: Transform discovered symptom-risk associations into new predictive features, providing deeper insights into how multiple factors interact in disease progression.
- **Improve Predictive Accuracy**: Strengthen classification models by integrating mined patterns, making severity predictions more precise and reliable.
- **Assist in Early Detection**: Use mined patterns to identify high-risk patients based on lifestyle and medical history, enabling earlier interventions and treatment plans.
- **Increase Explainability in Diagnosis**: Provide clear, interpretable associations between symptoms and severity, making model outputs more understandable for medical professionals.

1.4. What is the relation/connection to the material we learn in class?

- **Pattern Mining**: Uses **Apriori and FP-Growth** to discover meaningful symptom-risk relationships.
- **Predictive Model Analytics**: Refines feature selection and improves classification performance using mined patterns.
- **Data Exploration**: Utilizes EDA techniques to analyze symptom distributions and correlations before applying pattern mining.

2. The Solution

2.1 Initial Ideas for a Solution

- **Idea 1: Association Rule Mining for Lung Cancer Severity Prediction**
  - We will apply **association rule mining** to the lung cancer dataset to identify frequently co-occurring symptoms and risk factors.
  - The goal is to extract rules such as:
    "If a patient has Chronic Cough and Exposure to Air Pollution, then there is a high probability of developing Severe Lung Cancer."
  - These rules will help in understanding which symptom-risk combinations contribute most to cancer severity.
- **Idea 2: Frequent Pattern Mining for Feature Engineering**
  - We will use **Frequent Pattern Growth (FP-Growth) or Apriori Algorithm** to find recurring symptom clusters.
  - New **pattern-based features** will be created, such as:
    "High-Risk Indicator" (if a patient exhibits both Genetic Risk and Chronic Lung Disease), or "Environmental Risk Score" (if Air Pollution and Passive Smoking co-occur).
  - These engineered features will then be integrated into predictive models to improve classification accuracy.

2.2 Propose Ways to Measure the Solution

To assess the effectiveness of our approach, we will use:

- **Support, Confidence, and Lift** (Evaluating Association Rules)
  - **Support**: Frequency of a symptom-risk pattern in the dataset.
  - **Confidence**: Probability of severe cancer given a specific pattern.
  - **Lift**: Measures how much a pattern increases the likelihood of severity compared to random chance.
- **Model Performance** (Assessing the impact of mined patterns)
  - **Accuracy** & **Recall**: Evaluating classification improvements.
  - **Feature Contribution**: Checking how much mined patterns influence predictions.

3. Related Work

- Identifying HotSpots in Lung Cancer Data Using Association Rule Mining

    - **Authors:** Ankit Agrawal & Alok Choudhary
    - **Source:** IEEE International Conference on Data Mining Workshops
    - **Summary:** This study applies **association rule mining** to SEER lung cancer data to detect **hotspots—patient groups with distinct survival patterns**. The **HotSpot algorithm** identifies key factors (e.g., **age, tumor grade, lymph nodes**) linked to survival, providing **interpretable insights** for improving prognosis.

- Comparison of the C4.5 and a NaiveBayes Classifier for the Prediction of Lung Cancer Survivability

    - **Authors:** George Dimitoglou, James A. Adams, and Carol M. Jim
    - **Source:** SEER Data Study
    - **Summary:** This paper compares **C4.5 decision trees and Naive Bayes classifiers** in predicting **lung cancer survivability** using **15 years of SEER data**. C4.5 performed slightly better, but both methods required **strong feature selection** and **domain knowledge** to improve accuracy

4. Experiments Plan

4.1 Datasets

- **Cancer Patients and Air Pollution: A New Link**
  - **Description:** Patient data including demographics, environmental exposure, lifestyle habits, and medical history.
  - **Relevance:** Ideal for applying association rule mining to uncover links between risk factors and cancer severity.
  - **Link:** Cancer Patients and Air Pollution

- **Lung Cancer Dataset**
  - **Description:** Includes age, smoking, and alcohol consumption data.
  - **Relevance:** Helps identify frequent lifestyle patterns contributing to lung cancer severity.
  - **Link:** Lung Cancer Dataset

- **Lung Cancer Risk Dataset**
  - **Description:** Covers various risk factors influencing lung cancer.
  - **Relevance:** Enables discovering associations between risk factors and severity for better model explainability.
  - **Link:** Lung Cancer Risk Dataset

- **Lung Cancer Prediction**
  - **Description:** Demographic, medical history, treatment, and outcome data.
  - **Relevance:** Useful for mining symptom-risk factor patterns to improve early detection and accuracy.
  - **Link:** Lung Cancer Prediction

4.2 Testing Plan

1. **Discover Frequent Patterns** – Utilize association rule mining to find connections between symptoms, lifestyle factors, and lung cancer severity.
2. **Enhance Data Representation** – Transform identified patterns into new features to improve predictive model performance.
3. **Assess Model Impact** – Compare model effectiveness before and after incorporating mined patterns using **Accuracy, Precision, and Recall** metrics.
4. **Validate Across Datasets** – Test the approach on multiple lung cancer datasets to ensure reliability and consistency.