DSC-680 Project-2 Proposal

Topic: Lung Cancer prediction and analysis

Developing a predictive model for lung cancer using machine learning techniques to aid early diagnosis and treatment planning.

Business Problem: The project aims to address the critical issue of late-stage diagnosis of lung cancer by creating a tool that can predict the likelihood of lung cancer based on patient data, thus improving early detection and increasing survival rates.

Lung cancer remains one of the leading causes of cancer-related deaths worldwide, primarily due to its late-stage diagnosis when treatment options are limited and less effective. Early detection significantly improves prognosis and survival rates, as treatments are more effective in the early stages of the disease. However, lung cancer's symptoms often appear late, and current screening methods, such as low-dose CT scans, are not universally applied due to cost, accessibility, and radiation exposure concerns.

This project aims to address these challenges by developing a predictive model that leverages patient data to assess the likelihood of lung cancer. By incorporating a variety of data points such as patient demographics, medical history, smoking habits, genetic markers, and imaging data, the model can identify individuals at high risk of developing lung cancer. This tool can assist healthcare providers in making more informed decisions about which patients should undergo further diagnostic testing or receive preventive care.

Datasets: The primary dataset will be obtained from publicly available medical databases such as the SEER (Surveillance, Epidemiology, and End Results) program and Kaggle lung cancer datasets. The data will include patient demographics, medical history, smoking habits, imaging data, and pathology reports.

Methods: The project will utilize a variety of machine learning techniques including logistic regression, decision trees, random forests, and neural networks. Feature selection methods will be used to identify the most significant predictors. Model evaluation will be performed using cross-validation, and performance metrics such as accuracy, precision, recall, and AUC-ROC will be used to assess the models.

Ethical Considerations: Potential ethical concerns include patient privacy and data security, especially when dealing with sensitive medical records. Ensuring informed consent and data anonymization will be crucial. Bias in data and model outcomes, which could lead to disparities in diagnosis, must be addressed by careful validation and bias mitigation strategies.

Challenges/Issues: Anticipated challenges include dealing with imbalanced datasets, as lung cancer cases might be significantly fewer than non-cancer cases. Ensuring the model's generalizability to different populations and dealing with missing or inconsistent data will also be challenging.

Imbalanced Datasets: A significant challenge is dealing with imbalanced datasets where non-cancer cases far outnumber cancer cases. This imbalance can bias the model towards predicting non-cancer outcomes, resulting in high accuracy but poor sensitivity for detecting actual cancer cases. Techniques like oversampling, undersampling, and SMOTE are necessary but can introduce issues like overfitting.

Model Generalizability: Ensuring the model generalizes well to different populations is critical. Training data may not represent broader demographics, leading to biased predictions. Cross-validation and external validation with diverse datasets help mitigate this issue but require extensive and varied data sources.

Missing or Inconsistent Data: Medical datasets often have missing or inconsistent data due to incomplete records or varying data collection methods. Handling this involves sophisticated imputation techniques and rigorous data preprocessing, which can be time-consuming and complex.

Data Privacy and Security: Handling sensitive patient data necessitates stringent privacy and security measures. Ensuring compliance with data protection regulations and implementing data anonymization, encryption, and secure storage protocols is crucial to protect patient confidentiality.

Interpretability: Complex models like neural networks offer high accuracy but lack interpretability, making it hard for healthcare providers to trust and understand predictions. Using explainability techniques like SHAP and LIME is essential to build trust and facilitate model adoption in clinical settings.

References: I will refere below link.

1. National Cancer Institute's SEER program: https://seer.cancer.gov/

2. Kaggle lung cancer datasets: https://www.kaggle.com/datasets