Enhancing Cancer Prognosis Through Early Detection and Diagnostic Insights

Carolina Perez, Cody Folgmann, Dain Kim

Description

This Project will focus on predicting the survival outcomes of cancer patient based on early detection and diagnostic data. We will investigate what factors of early detection can be linked to improved survival rates for patients using data, such as tumor size, biopsy results, and patient demographics. From these we can attempt to construct a predictive model in order to assess survival probability.

The Data Set

URL of DataSet:

Brain Tumor Prediction Dataset:

- **URL**: https://www.kaggle.com/datasets/ankushpanday1/brain-tumor-prediction-dataset
- **Description**: Contains data related to patients diagnosed with brain tumors, including lifestyle habits and treatment details.

Prior Work

Cancer Research is extensive, and their have been previous works in early detection. Some Previous Models that have been created include:

- *Breast Cancer Survival Prediction Model (Saker, 2020)*
  A Model to breast cancer survival rates using clinical and genetic data. It helps to assess the likelihood of survival based on various factors such as tumor size, genetic mutations, and patient health history. This model aims to provide personalized survival predictions.
- *Lung Cancer Survival Prediction Model (Y Wu, 2021)*
  Focused on early detection through lung cancer screenings, this model predicts survival outcomes by analyzing patient data such as age, smoking history, and tumor characteristics. It highlights how early screening can lead to earlier treatment and improved survival rates, especially in high-risk populations.
- *Socio-Economic Factors and Cancer Survival (Jones et al., 2018)*
  This model incorporates socio-economic data such as income, education, and healthcare access to predict cancer survival rates. By processing these socio-economic factors, it shows how a patient's environment and access to resources can significantly impact their survival chances.

Proposed Work

- **Data Cleaning**
  Handling Missing Values: Identify and impute or remove missing data to ensure completeness and accuracy.
  Removing Duplicates: Detect and eliminate duplicate records to maintain data integrity.

- **Data Pre-Processing**
  Normalization and Standardization: Scale numerical attributes to bring them to a common scale, ensuring consistency.
  Encoding Categorical Variables: Convert categorical data into numerical formats using techniques like one-hot encoding or label encoding.

- **Data Integration**
  Merging Data from Multiple Sources: Combine datasets from different sources to create a unified dataset.
  Transforming Data into a Consistent Format: Ensure all data is in a standardized format to facilitate analysis.

- **Modeling**
  Model Selection and Training: Choose appropriate machine learning algorithms and train models on the pre-processed data.

List of Tools

- **Python** : data manipulation, analysis, and modeling. (Libraries: Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn.)
- **Jupyter Notebooks** : interactive data analysis and visualization.
- **Power BI/Tableau** : offer drag-and-drop interfaces and various visualization options
- **SQL** : querying and managing relational databases
- **Excel**: Use Excel's charting capabilities to create visually appealing graphs and dashboards

Evaluation

We can evaluate the effectiveness of ot prectivice model by comparing them against actual data from the dataset.

- Accuracy - The percentage of correct predictions for patient survival
- Classification Evaluation- How does the model classify survival vs non-survival cases?
- Model Comparison - how does this model compare results to other existing models?
- Importance analysis - analyze the importance of different features in our predictive model for cancer survival outcomes, and what techniques can be used to determine which variables contribute the most to the model's predictions?