Colorectal Cancer Survival Prediction using Machine Learning

Repo: https://github.com/sandi148/colorectal-cancer-survival-prediction

Video: sandy 211002137-20250604_231426-Meeting Recording.mp4

Deployment: https://colorectal-cancer-survival-prediction-rybrucaoafe8jpmxpozqxn.streamlit.app/

Table of Contents

- Project Overview
- Dataset
- Data Exploration
- Data Preprocessing
- Feature Selection
- Visualizing
- Modeling
- Evaluation
- Deployment

Overview

This project aims to build machine learning models to predict 5-year survival among patients with colorectal cancer based on clinical, lifestyle, and demographic factors. It demonstrates the application of classification algorithms, ensemble learning, and performance evaluation techniques in healthcare contexts. Early prediction of colorectal cancer survival can help doctors tailor personalized treatment strategies, reduce mortality, and improve patient quality of life. The workflow covers data cleaning, feature engineering, feature selection, outlier removal, model training, evaluation, and ensemble learning.

Dataset

- Rows: 167,497 patients
- Columns: 28 features including demographics, clinical data, lifestyle factors, and economic indicators.
- Target Variable: `Survival_5_years` (Yes/No)

- Colorectal Cancer Global Dataset & Predictions

1. Data Exploration

- Data loaded into a pandas DataFrame for exploration and preprocessing.

- Summary statistics

2. Data Preprocessing

- Missing Values: Checked and handled as appropriate.
- Outlier Removal: Outliers in numerical features removed using the IQR method
- Check for duplicate rows
- Encoding: Label Encoding for binary categorical features.
  Strip Whitespace from Ordinal Columns
  Ordinal encoding applied to categorical features (e.g., Healthcare_Access, Physical_Activity, Diet_Risk).
  Label Encoding of Categorical Columns

3. Feature Selection and Engineering

- Feature Selection: Applied `SelectKBest` with `f_classif` and `chi2` to identify the
- Top Features Identified: [`'Family_History', 'Alcohol_Consumption', 'Mortality', 'Insurance_Status', 'Urban_or_Rural', 'Country', 'Incidence_Rate_per_100K', 'Mortality_Rate_per_100K', 'Healthcare_Access', 'Obesity_BMI', 'Screening_History', 'Treatment_Type', 'Tumor_Size_mm'`]
- feature engineering: Tumor_Size_Category
- Saving Cleaned Data: Cleaned DataFrame saved as cleaned_selected_data.csv

4. Visualizing

- Visualizing data distributions for selected features
- the correlation matrix by heatmap
- Visualize target class distribution
- ROC Curves for all models
- Confusion Matrix for all models

5. Modeling Pipeline

- Feature and Target Definition
- Data Splitting
- Feature Scaling
- Handling Class Imbalance: `from imblearn.over_sampling import SMOTE`

6. Models and Ensemble Methods Used

- Logistic Regression (baseline)
- K-Nearest Neighbors (KNN) (with different k values)
- Random Forest Classifier
- STACKING classifier (Using LR, KNN, NB)
- Bagging (with Logistic Regression)
- Decision Tree Classifier (with different max depths)
- XGBoost Classifier
- AdaBoost Classifier
- Voting Classifier (ensemble)
- Ensemble Model Training and Evaluation
- Cross Validation for each model

7. Deployment using an AdaBoostClassifier with a DecisionTree base estimator for predicting 5-year survival in colorectal cancer patients