Colorectal Cancer Dataset

Rishikesh Kumar - Morelli Davide

Seminar on
Artificial
Intelligence 2

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

WHAT WE DID SO FAR...

WHAT WE DID SO FAR...

DATA EXPLORATION (EDA)

For the EDA insights we'll switch to the notebook:

We'll return here for the Data Preprocessing part

DATA PREPROCESSING

No missing values

No wrong values

Numerical attributes

AGE - TUMOR SIZE - HEALTHCARE COST - INCIDENCE RATE - MORTALITY RATE

Binary attributes

GENDER - FAMILY HISTORY - SMOKING - ALCOHOL CONSUMPTION - DIABETES - IBD - GENETIC MUTATION - EARLY DETECTION - SURVIVAL 5 YEARS - MORTALITY - URBAN OR RURAL - ECONOMIC CLASSIFICATION - INSURANCE STATUS - SURVIVAL PREDICTION

Categorical attributes

CANCER STAGE - OBESITY BMI - DIET RISK - PHYSICAL ACTIVITY - SCREENING HISTORY - TREATMENT - HEALTHCARE ACCESS

DATASET STANDARDIZATION & CLEANING

Binary values

ENCODED USING:

Binary encoding

Categorical values

ENCODED USING:

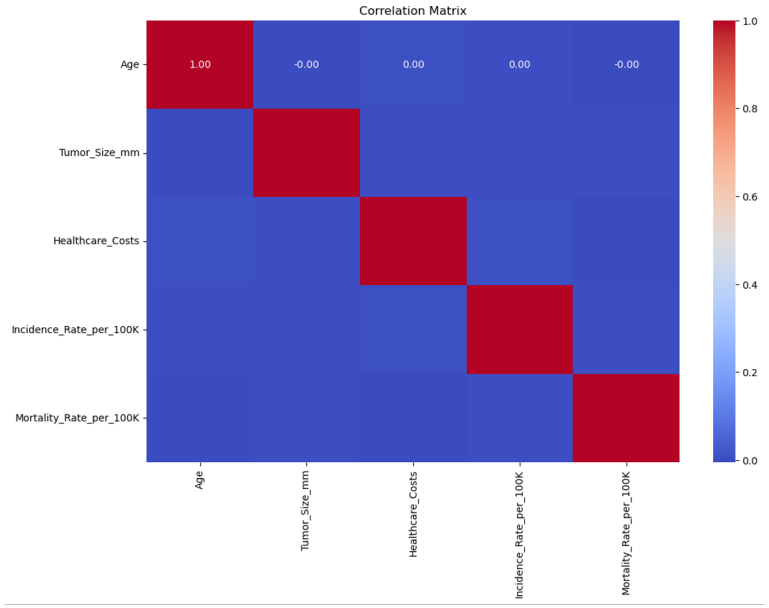
One-hot encoding

No duplicate rows to eliminate

No outliers (IQR)

FEATURE SELECTION

Feature selection using **correlation matrix** and **Chi-squared method**



The numerical fields have **very low correlation**

However low correlation can be beneficial for methods like **bagging ensembles** or **boosting algorithms**

FEATURE SELECTION

Using chi-squared analysis we managed to identify the “worst” and the “best” features:

BEST	WORST
Genetic mutation - IBD - Age	Tumor size - Alcohol consumption - Urban or Rural

However low correlation can be beneficial for methods like **bagging ensembles** or **boosting algorithms**

SUGGESTIONS FOR THE NEXT STEPS?

Thank you for your attention.