**Executive Summary**

This project aims to enhance our understanding of the factors influencing liver cancer outcomes by leveraging a comprehensive dataset containing demographic, epidemiological, and healthcare-related variables. By conducting a thorough EDA and implementing robust data preparation and feature engineering techniques, I seek to identify key predictors and interactions that drive liver cancer incidence and survival rates. The insights derived from this analysis will not only support the development of accurate predictive models but also inform targeted intervention strategies and policy decisions in public health. The approach encompasses detailed visualizations, statistical analyses, and data transformations to ensure the data is ready for subsequent machine learning applications.

**Problem Statement / Research Objectives**

Liver cancer remains a significant global health challenge, with multifactorial causes ranging from lifestyle habits to healthcare access. The primary problem addressed in this project is to predict liver cancer outcomes using a dataset that includes variables such as Population, Incidence\_Rate, Mortality\_Rate, Age, Alcohol\_Consumption, Smoking\_Status, and healthcare access indicators among others. The research objectives are as follows:

1. To perform an in-depth exploratory analysis of the dataset to understand the distribution, relationships, and interactions between variables.
2. To identify key risk factors and quantify their influence on liver cancer outcomes.
3. To engineer features and transform variables (e.g., scaling continuous variables and encoding categorical variables) to prepare the dataset for predictive modeling.
4. To provide actionable insights that can be used by healthcare professionals and policymakers to design targeted intervention strategies.

**Exploratory Data Analysis**

The EDA phase of this project involved a comprehensive examination of both numerical and categorical features. Key steps included:

* Data Analysis and Visualization: Histograms, boxplots, and scatter plots were utilized to assess the distribution of continuous variables such as Population, Incidence\_Rate, Mortality\_Rate, Age, and Survival\_Rate. This visual inspection highlighted the range of values and helped detect potential outliers using the Interquartile Range (IQR) method – 0.
* Assumptions and Unit of Analysis: I assumed that each record in the dataset represents a distinct observation that may encompass both individual-level characteristics (e.g., Age, Gender) and aggregate measures (e.g., Population).
* Feature Interactions: Correlation analyses and bivariate visualizations (such as scatter and box plots) were employed to explore interactions between variables. For instance, relationships between healthcare access measures and survival rates were examined, as were potential interactions among lifestyle factors such as Alcohol\_Consumption and Smoking\_Status.

**Data Preparation / Feature Engineering**

In this stage, the focus was on transforming and encoding variables to make the dataset suitable for modeling:

* Handling of Categorical Variables: Categorical variables were split into two groups: ordinal and nominal. For ordinal features—such as Alcohol\_Consumption, Healthcare\_Access, Preventive\_Care, Obesity, and Seafood\_Consumption—I applied an OrdinalEncoder with predefined orders to preserve the inherent ranking. The remaining nominal variables (including Country, Region, Gender, Smoking\_Status, Hepatitis\_B\_Status, Hepatitis\_C\_Status, Diabetes, Rural\_or\_Urban, Herbal\_Medicine\_Use, Screening\_Availability, Treatment\_Availability, Liver\_Transplant\_Access, and Ethnicity) were transformed using scikit-learn’s OneHotEncoder to create binary features while avoiding multicollinearity.
* Variable Transformations and Data Scaling: Given the vast differences in scale among continuous variables (for example, Population versus Age or Incidence\_Rate), a StandardScaler was used to standardize these features to a mean of zero and unit variance. This scaling step is crucial for ensuring that the model does not disproportionately favor features with larger numerical ranges and also helps in achieving faster convergence in many machine learning algorithms.
* Assumptions and Tests: Throughout the data preparation process, underlying assumptions—such as the appropriateness of the IQR method for outlier detection or the correctness of ordinal mappings—were critically examined. A copy of the original dataset was maintained before scaling to preserve the raw data for reference and reproducibility, ensuring transparency in the preprocessing workflow.

Data: <https://www.kaggle.com/datasets/ankushpanday1/liver-cancer-predictions>