MSBD566-Predictive Modeling and Analytics

Assignment 2

1. Name: Esmeralda Garcia
   Course Name: MSBD566 Predictive Modeling and Analytics
   Date: October 2025
   Assignment Title: Dataset Description of GDSC Drug Sensitivity in Cancer

2. Data Title: Genomics of Drug Sensitivity in Cancer (GDSC)

3. Source: Kaggle
   "Genomics of Drug Sensitivity in Cancer (GDSC)" from Samira Alipour
   https://www.kaggle.com/datasets/samiraalipour/genomics-of-drug-sensitivity-in-cancer-gdsc?resource=download

4. Data Description: The dataset has information related to the different drug response of various genetic cell lines. The target of interest in the dataset is LN_IC50 which would tell us if the various cell lines are sensitive to the drug administered. Cancer cell lines are exposed to the drug for 72 hours and its viability is then measured. This dataset includes over 1000 cell lines and hundreds of anti-cancer drugs.

5. Data Dictionary:

Column Name	Type	Description
COSMIC_ID	Numeric	Unique identifier for the cell line from the COSMIC database.
CELL_LINE_NAME	Categorical (String)	Name of the cancer cell line used in the experiment.
TCGA_DESC	Categorical (String)	Description of the cancer type according to The Cancer Genome Atlas.
DRUG_ID	Numeric / String	Unique identifier for the drug used in the experiment.
DRUG_NAME	Categorical (String)	Name of the drug used in the experiment.
LN_IC50	Numeric (float)	Natural log of the half-maximal inhibitory concentration (IC50).
AUC	Numeric (float)	Area Under the Curve, a measure of drug effectiveness.
Z_SCORE	Numeric (float)	Standardized score of the drug response, allowing comparison across different drugs and cell lines.
GDSC Tissue descriptor 1	Categorical	Primary tissue type classification.
GDSC Tissue descriptor 2	Categorical	Secondary tissue type classification.
Cancer Type (matching TCGA label)	Categorical	Cancer type according to TCGA classification.
Microsatellite instability Status (MSI)	Categorical (binary)	Indicates the cell line's MSI status.
Screen Medium	Categorical	The growth medium used for culturing the cell line.
Growth Properties	Categorical	Characteristics of how the cell line grows in culture.
CNA	Numeric / Ordinal	Data on gene copy number changes in the cell line.
Gene Expression	Numeric / Continuous	Information on gene expression levels in the cell line.
Methylation	Numeric / Continuous	Data on DNA methylation patterns in the cell line.
TARGET	Categorical	The molecular target(s) of the drug.
TARGET_PATHWAY	Categorical	The biological pathway(s) targeted by the drug.

6. Data Disclaimers:
   - Data resulted from cell lines not human patients.
   - Some columns contain missing or null values so data will be cleaned and normalized.