MSBD566 – Predictive Modeling and Analysis
Name: Esmeralda Garcia
Date: October 22, 2025
Assignment: Midterm Project

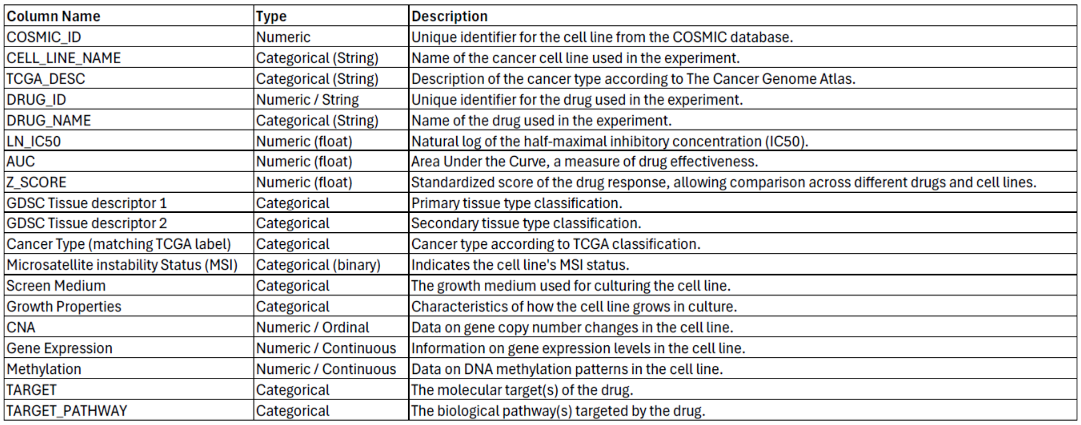
Project Description

This midterm project was designed to predict how cancer cell lines will respond to various anticancer drugs using a publicly available dataset from Kaggle called "Genomics of Drug Sensitivity in Cancer (GDSC). I will be designating 3 categories for prediction, Very Potent, Average Sensitivity, and Resistant, based on the genomic and experimental features. I believe this prediction is of great importance for the field of precision and personalized medicine. This project used Python and scikit-learn to train a classification model to identify patterns in data and predict drug responses with acceptable accuracy.

Data Description

The dataset, Genomics of Drug Sensitivity in Cancer (GDSC), used for this midterm project was downloaded from Kaggle. It contains information related to the different drug response of almost 1000 human cancer cell lines. The cell lines were exposed to almost 300 different anti-cancer drugs, the combinations resulted in over 240,000 individual observations. 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 were exposed to the drug for 72 hours and its viability was then measured.

Variables in my data:



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.

I selected the target to be the original LN_IC50 variable. This variable was categorized into three biologically meaningful categories:

Category	LN_IC50 Range	Meaning
Very Potent	**< 0**	**Highly sensitive**
Average Sensitivity	**0 – 2.3**	**Moderate Response**
Resistant	**> 2.3**	**Low Sensitivity**

The dataset contained some missing values, mainly in the cancer cell type. To simplify preprocessing, rows with any missing values in selected features were removed, resulting in a clean dataset for my training and evaluation.

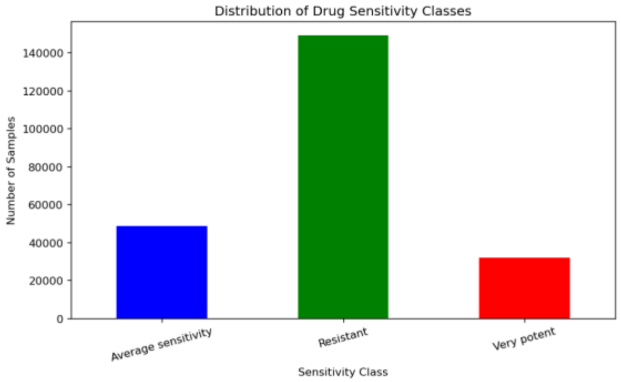
The limitation of this data set is that the results reflect that observed in vitro and not with actual patients. Results will absolutely vary in the context of the human bod and other biological processes involved.

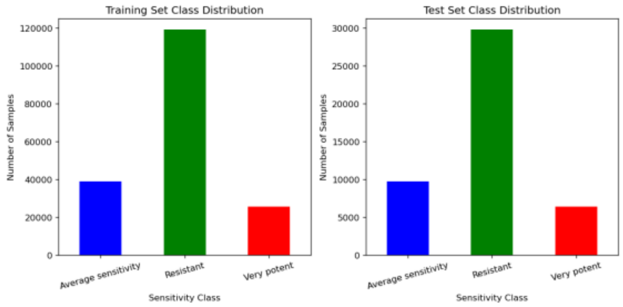
Method and Analysis

The project followed a structured machine-learning workflow:

1. Data Preprocessing:
   The selected features were separated into numerical and categorical groups. Numerical features (AUC, Z_SCORE, CNA, Gene Expression, and Methylation) were standardized using *StandardScaler* to ensure they had similar ranges. Categorical variables (tissue type, target pathway, MSI status, screen medium, and growth properties) were encoded using *OneHotEncoder*, this is a data processing tool that converts text categories into a numerical format like 0 and 1.

2. Train–Test Split:
   The data was split into 80% training and 20% testing subsets using stratified sampling to keep the class proportions. This ensured that all three drug-response classes were represented fairly in both datasets.



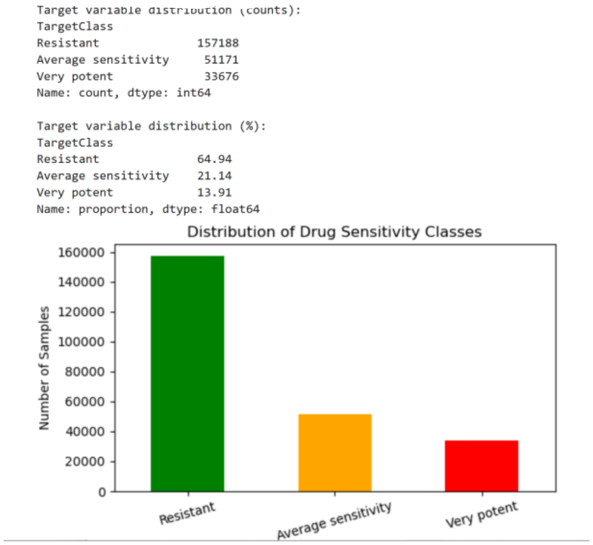


3. Model Training:
   A Logistic Regression classifier was used with a one-vs-rest strategy to distinguish between the multiple classes. The model was built with a scikit-learn pipeline, which automatically applies preprocessing and training in a single step. Logistic regression was chosen for its interpretability, simplicity, and ability to provide class probabilities.

4. Model Evaluation:
   The trained model predicted the response class on the test set. Performance was measured using standard classification metrics like precision, recall, F1-score, and accuracy, as well as visualization tools like the confusion matrix and ROC curves. These visualizations helped identify which classes the model performed well on and where it made mistakes.

5. Visualizations:
   Several plots were created to better understand the data and results.

   - A bar chart of class distribution showed that “Resistant” samples were most common.



```
Target variable distribution (counts):
TargetClass
Resistant              157188
Average sensitivity     51171
Very potent             33676
Name: count, dtype: int64

Target variable distribution (%):
TargetClass
Resistant              64.94
Average sensitivity    21.14
Very potent            13.91
Name: proportion, dtype: float64
```

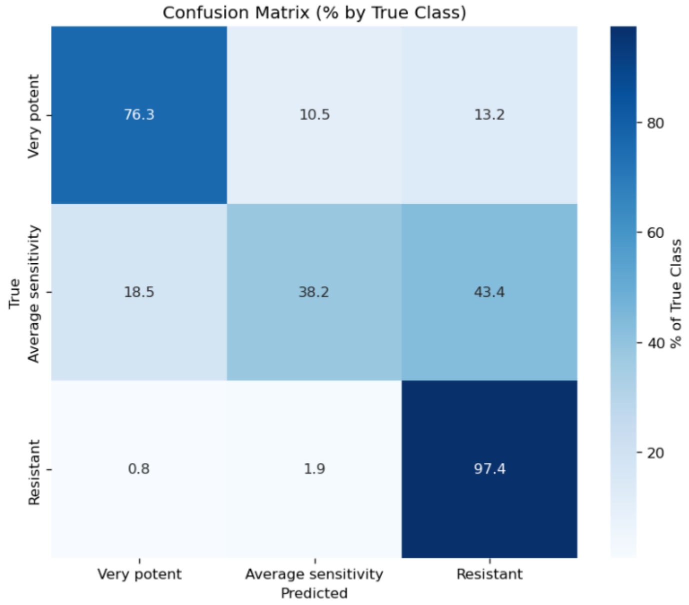
- ○ A plot of missing data identified which columns required cleaning.
- ○ A bar chart of the Top 5 Very Potent Drugs showed which compounds were most effective across cell lines.
- ○ A confusion matrix heatmap showed the percentage of correct and incorrect predictions for each class.
- ○ ROC curves displayed the trade-off between sensitivity and specificity for each of the three classes.

Evaluation

The logistic model achieved an overall accuracy of approximately 81.8%, showing strong predictive performance for a biological classification problem.

- Resistant samples were predicted with the highest precision of 0.85 and recall of 0.97, meaning the model correctly identified most resistant cases with few false positives.
- Very Potent samples achieved a moderate F1-score of 0.73, showing the model could reliably identify highly sensitive drug–cell combinations.
- Average Sensitivity was the most difficult class, with lower recall of 0.38 because this middle group overlaps biologically with both potent and resistant responses.

The confusion matrix confirmed that most misclassifications occurred between the Average Sensitivity class and the other two classes, this was expected due to gradual biological transitions rather than the sharp boundaries between drug-response levels.



Overall, the logistic regression model performed well and provided results that could easily be interpreted by clinicians when determining what drug to administer per cell type. My future work would revolve around increasing the accuracy, this can be achieved using non-linear models like Random Forest or XGBoost.