**INTELLIGENCE IN BIOLOGICAL SYSTEMS-2 (22BIO211)**

**DRUG RESPONSE PREDICTION USING LOGISTIC &**

**RANDOM FOREST REGRESSION**

**In partial fulfilment for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**CSE(AI)**

****

**Centre for Computational Engineering and Networking**

**AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE**

**AMRITA VISHWA VIDYAPEETHAM**

**COIMBATORE - 641 112 (INDIA)**

**JUNE– 2024**

A Project

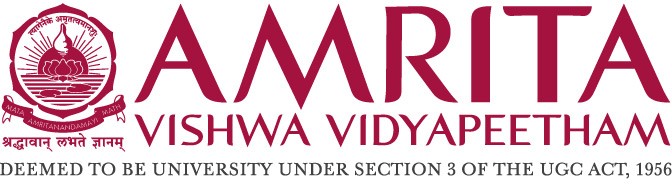
Submitted by:

**M.C. DHNAUSH -CB.EN.U4AIE22130**

**AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE**

**AMRITA VISHWA VIDYAPEETHAM**

**COIMBATORE - 641 112**



**BONAFIDE CERTIFICATE**

This is to certify that the Project submitted by M.C. DHNAUSH - CB.EN.U4AIE22130 for the award of the Degree of Bachelor of Technology in the “CSE(AI) ” is a bonafide record of the work carried out by her under our guidance and supervision at Amrita School of Artificial Intelligence, Coimbatore.

**Dr Harishchander A**

Project Guide

**Dr. K.P.SOMAN**

**Professor and Dean AI**

***Submitted for the University examination held on 13-06-2024***

# DECLARATION

Myself M.C.DHNAUSH - CB.EN.U4AIE22130, hereby declare that this project is the record of the original work done by me under the guidance of **Dr Harishchander A** , Professor, Centre for Computational Engineering and Networking, Amrita School of Artificial Intelligence, Coimbatore. To the best of my knowledge this work has not formed the basis for the award of any degree/diploma/ associate ship/fellowship/or a similar award to any candidate in any University.

**Place: Coimbatore**

**Date: 13-06-2024 Signature of the Students**

## 

## Acknowledgement

I would like to express my special thanks of gratitude to our teacher (**Dr Harishchander A** ), who gave me the golden opportunity to do this wonderful project, which also helped me in doing a lot of exploration and I came to know about so many new things. I am thankful for the opportunity given.

**TABLE OF CONTENTS**

[DECLARATION 3](#_Toc169240500)

[Acknowledgement 4](#_Toc169240501)

[1 INTRODUCTION 6](#_Toc169240502)

[2 OBJECTIVE 8](#_Toc169240503)

[2.1 Data Collection and Preprocessing: 9](#_Toc169240504)

[2.2 Model Development: 9](#_Toc169240505)

[2.3 Model Evaluation: 9](#_Toc169240506)

[2.4 Application Development: 10](#_Toc169240507)

[3 SIGNIFICANCE 10](#_Toc169240508)

[3.1 Personalized Treatment: 10](#_Toc169240509)

[3.2 Improved Clinical Outcomes: 11](#_Toc169240510)

[3.3 Cost Efficiency: 11](#_Toc169240511)

[3.4 Advancement in Machine Learning Applications: 12](#_Toc169240512)

[3.5 Educational and Research Value: 12](#_Toc169240513)

[4 DATASET 13](#_Toc169240514)

[4.1 Key Features of the Dataset 13](#_Toc169240515)

[5 METHODOLOGY: 14](#_Toc169240516)

[5.1 By logical regression 14](#_Toc169240517)

[5.1.1 Data Collection: 15](#_Toc169240518)

[5.1.2 Encoding Categorical Variables: 15](#_Toc169240519)

[5.1.3 Normalisation: 15](#_Toc169240520)

[5.1.4 Data Splitting: 15](#_Toc169240521)

[5.2 Model Training and Evaluation 16](#_Toc169240522)

[5.2.1 Model Initialization: 16](#_Toc169240523)

[5.2.2 Training: 16](#_Toc169240524)

[5.2.3 Evaluation: 16](#_Toc169240525)

[5.2.4 Predictive Function 17](#_Toc169240526)

[5.2.5 Logistic Regression Model: 17](#_Toc169240527)

[5.2.6 Result 18](#_Toc169240528)

[5.3 RANDOM FOREST REGRESSION 19](#_Toc169240529)

[5.3.1 Data Collection and Preprocessing 19](#_Toc169240530)

[5.3.2 Data Collection: 19](#_Toc169240531)

[5.3.3 Model Training and Evaluation 20](#_Toc169240532)

[5.3.4 Random Forest Model 21](#_Toc169240533)

[6 Cross Validation 23](#_Toc169240534)

[7 Confusion matrix and Classification report 23](#_Toc169240535)

[8 Result 26](#_Toc169240536)

[9 CONCLUSION 26](#_Toc169240537)

# INTRODUCTION

Precision medicine is a revolutionary approach to healthcare that focuses on tailoring therapies to the unique needs of each patient. Accurately anticipating a patient's reaction to a particular medicine is essential to understanding this idea. Such predictive power can decrease side effects, increase treatment efficacy dramatically, and lessen the trial-and-error process that is frequently involved in therapeutic procedures.

Machine learning (ML) has become a potent tool in the field of biological data analysis, allowing the detection of intricate patterns in enormous datasets that conventional statistical techniques might overlook. Logistic regression is a prominent machine learning technique that is easy to use, easy to understand, and effective for jobs involving binary categorization.

The main goal of this work is to forecast medication responses in cancer treatment by using logistic regression. The multifaceted nature of cancer and its inconsistent response to treatment make it a formidable obstacle to the development of tailored, efficient treatments. In order to improve the predicted accuracy of drug responses in cancer, this research makes use of a large dataset rich in variables relevant to pharmacology and efficacy.

Much information on medication targets, pathways, and efficacy metrics is available in the dataset used in this work, "PANCANCER ANOVA Preprocessed," which is essential for comprehending drug-target interactions and their therapeutic efficacy. The existence of categorical characteristics in this dataset, such as pharmaceutical targets and pathways, presents a substantial hurdle for its use. The study uses LabelEncoder to transform categorical data into a numerical format appropriate for logistic regression analysis in order to address this. Standardized numerical features are also included to guarantee uniform scaling throughout the collection.

The results of this study show that logistic regression is a useful method for forecasting how cancer patients will respond to certain medications. The practical relevance of this technique is highlighted by the accuracy of the model and the user-friendly predictive application that was created. This approach has the potential to improve oncology treatment outcomes and advance precision medicine by incorporating machine learning into medication response prediction.

To summarize, this study demonstrates the potential of machine learning to improve and modify customized cancer therapy by presenting a strong methodology for predicting medication responses based on logistic regression. Larger datasets and more sophisticated models may be used in future studies to further increase clinical applicability and prediction accuracy.

# OBJECTIVE

The primary objective of the project "Drug Response Prediction using Logistic Regression" is to develop a machine learning model that can accurately predict the response of cancer patients to various drugs. This involves several key steps:

## Data Collection and Preprocessing:

Collecting relevant data on cancer drugs, including features such as drug targets, pathways, and efficacy measures (e.g., IC50 values). The data must be cleaned, encoded, and normalized to ensure it is suitable for modeling.

## Model Development:

Building and training a logistic regression model using the preprocessed data. The model aims to learn the relationships between the input features and the drug response outcomes.

## Model Evaluation:

Assessing the performance of the model using metrics such as accuracy, confusion matrix, and classification reports. This step ensures that the model is reliable and can make accurate predictions.

## **Application Development**:

Creating a user-friendly application that allows healthcare professionals to input new patient data and receive predictions on drug efficacy. This tool should be practical and easy to use in clinical settings.

**5. Validation and Refinement:** Continuously validating and refining the model using cross-validation techniques and larger datasets to improve its accuracy and robustness.

# SIGNIFICANCE

The significance of the project lies in its potential impact on the field of precision medicine, particularly in oncology. Here are the detailed aspects of its significance:

## Personalized Treatment:

* **Tailored Therapy:** By predicting individual responses to drugs, the model enables personalized treatment plans. This means patients receive medications that are more likely to be effective for their specific cancer type and genetic makeup.
* **Reduced Side Effects:** Personalizing treatments can minimize the adverse effects that often result from a one-size-fits-all approach, as patients are less likely to receive ineffective drugs.

## Improved Clinical Outcomes:

* **Higher Success Rates:** Accurate drug response predictions can increase the likelihood of successful treatment outcomes, leading to higher survival rates and better quality of life for cancer patients.
* **Optimized Drug Selection:** Clinicians can make informed decisions about which drugs to prescribe, optimizing the treatment process and reducing the time spent on ineffective therapies.

## Cost Efficiency:

* **Reduced Healthcare Costs:** By avoiding ineffective treatments, the model can help reduce healthcare costs associated with trial-and-error prescribing and managing side effects from unsuitable drugs.
* **Efficient Resource Utilization\*:** Better prediction models can lead to more efficient use of medical resources, ensuring that patients receive the most appropriate care promptly.

## Advancement in Machine Learning Applications:

* **Validation of Machine Learning:** The project demonstrates the practical application of logistic regression and other machine learning techniques in a real-world healthcare setting. This validation is crucial for the broader acceptance and integration of AI in medicine.
* **Foundation for Future Research:** The findings from this project can serve as a foundation for future research, encouraging the exploration of more complex models and larger datasets to further enhance prediction accuracy.

## Educational and Research Value:

* **Methodological Framework:** The project provides a methodological framework for researchers and practitioners in the field of biomedical informatics and machine learning, showcasing the steps involved in developing and validating predictive models.
* **Interdisciplinary Collaboration:** It highlights the importance of interdisciplinary collaboration between data scientists, healthcare professionals, and researchers to solve complex medical problems.

# DATASET

The dataset used in this study is the "PANCANCER ANOVA Preprocessed" dataset, which is specifically designed for cancer medication response analysis.

**Purpose:** The dataset is intended for predicting drug responses in cancer treatment by providing comprehensive information on various pharmacological and efficacy-related features.

## Key Features of the Dataset

1. **Drug Target:** This element identifies each drug's biological target, which is essential to comprehending the drug's mode of action and possible efficacy.
2. **Target Pathway:** This denotes the biological route that the medication is intended to target. Understanding pathways is crucial to comprehending how medications affect cellular functions.
3. **Feature Name:** A dataset's categorical identifier for various characteristics or attributes.
4. **Half-maximal inhibitory concentration (IC50) effect size:** This parameter indicates how well a drug inhibits a certain biological or biochemical process. This is a crucial pharmacological parameter.
5. **log IC50 Mean Pos:** A normalized indicator of medication efficacy, which is the logarithmic mean of the IC50 values for positive reactions.
6. **log IC50 Mean Neg:** The logarithmic mean of IC50 values for negative responses, which helps in understanding the variance in drug response.

# METHODOLOGY:

## By logical regression

### Data Collection:

The "PANCANCER ANOVA Preprocessed.csv" file, which has several characteristics linked to cancer medication responses, is used in the study. These characteristics offer a strong basis for study and include medication targets, routes, and efficacy indicators.

### **Encoding Categorical Variables**:

* Features that fall into one of several categories are included in the dataset, including "Feature Name," "Target Pathway," and "Drug target."
* Label Encoding: Label encoding was used to convert categorical information into numerical representations. Because each category is given a distinct numerical value, logistic regression modelling can be performed on the data. Moreover, "Drug name," the target variable, is label-encoded to aid with model comprehension and prediction.

### Normalisation:

* Standard Scaling: Standard scaling was used to standardize numerical data such as "ic50 effect size," "log ic50 mean pos," and "log ic50 mean neg." Through this method, each numerical characteristic is guaranteed to contribute equally and any disproportionate impact resulting from scale disparities is prevented. The features are adjusted to have a mean of zero and a standard deviation of one.

### ****Data Splitting****:

The dataset was divided into training and testing sets with an 80-20 split. This approach allows for objective evaluation of the model’s predictive power by training on 80% of the data and testing on the remaining 20%.

## Model Training and Evaluation

### Model Initialization:

Because of its efficiency and interpretability in binary classification issues, logistic regression was selected. It simulates a binary outcome's likelihood depending on one or more predictor variables.

### **Training**:

The training dataset was used to train the logistic regression model. In order to minimize the discrepancy between expected and actual results, the model's parameters were optimized. In order to guarantee convergence and the identification of the best solution, the training procedure involved limiting the number of iterations to 1000.

### Evaluation:

After training, the model's performance was evaluated using the testing dataset. The primary metric for evaluation was accuracy, defined as the percentage of correctly predicted instances out of all instances. This metric provides a straightforward assessment of the model's effectiveness.

### Predictive Function

A predictive function was developed to make the model practical and user-friendly for predicting drug responses based on new input data. This function performs the following steps:

* **Encoding and Scaling:** New input data are encoded and scaled similarly to the training data.
* **User Input:** The user is prompted to provide details for each selected feature.
* **Prediction:** The trained logistic regression model is used to make predictions.
* **Decoding:** The predicted drug name is decoded back to its original form using the label encoder for interpretation.

### Logistic Regression Model:

#### Hypothesis Function:

The logistic (sigmoid) function, which transforms the linear combination of input data into a probability, is used to represent the logistic regression hypothesis.

**= =**

#### Multinomial Logistic Regression:

Given that there are several drug names (classes) in the problem, the probability that an instance belongs to class k is as follows:

**P (y = k|X) =**

#### Cost Function:

The cross-entropy loss is the cost function utilized in multinomial logistic regression:

#### Gradient Descent:

Gradient descent is used to optimize the model parameters:

where **α** is the learning rate and is the gradient of the cost function with respect to parameter

### Result

The dataset was split into training and testing sets, with 160,736 samples for training and 40,184 samples for testing. The features included in the model were 'Drug target', 'Target Pathway', 'Feature Name', 'ic50 effect size', 'log ic50 mean pos', and 'log ic50 mean neg'. The model achieved an accuracy of approximately 80.06% on the testing set. The model successfully predicted the suitable drug, such as 73 (Camptothecin), for the real-life data used in the study.

## RANDOM FOREST REGRESSION

### Data Collection and Preprocessing

### Data Collection:

The same "PANCANCER ANOVA Preprocessed" dataset is utilized, containing features related to cancer drug responses such as drug targets, pathways, and efficacy metrics.

#### Encoding Categorical Variables:

Categorical Features: Includes "Feature Name," "Target Pathway," and "Drug target."

* **Label Encoding:** Categorical variables are transformed into numerical representations using label encoding. The target variable, "Drug name," is also label-encoded.

#### Normalization:

* **Standard Scaling**: Numerical features like 'ic50 effect size', 'log ic50 mean pos', and 'log ic50 mean neg' are normalized using standard scaling to ensure consistent scaling across features.

#### Data Splitting:

The dataset is divided into training and testing sets with an 80-20 split to facilitate objective evaluation of the model’s predictive power.

### Model Training and Evaluation

#### Model Initialization:

Random Forest, an ensemble learning method, is chosen for its robustness and ability to handle high-dimensional data. It builds multiple decision trees and merges them to improve the model's accuracy and control overfitting.

#### Training:

The Random Forest model is trained using the training dataset. Key parameters include:

* **Number of Trees:** 100 trees are used in the forest to balance between computational efficiency and model performance.
* **Max Features:** The number of features considered for splitting a node is set to 'auto', which uses the square root of the number of features.
* Bootstrap Sampling: Each tree is trained on a random subset of the training data, selected with replacement.

#### Evaluation:

After training, the model's performance is evaluated using the testing dataset. Performance metrics include accuracy, precision, recall, and F1 score, providing a comprehensive assessment of the model's effectiveness.

#### Predictive Function

A predictive function is developed to make the model practical and user-friendly for predicting drug responses based on new input data. This function includes:

* **Encoding and Scaling:** New input data are encoded and scaled similarly to the training data.
* **User Input:** The user provides details for each selected feature.
* **Prediction:** The trained Random Forest model is used to make predictions.
* **Decoding:** The predicted drug name is decoded back to its original form using the label encoder.

### Random Forest Model

#### Ensemble Learning:

Random Forest is an ensemble learning technique that creates multiple decision trees from different subsets of the training data and aggregates their predictions to produce a final prediction.

#### Tree Construction:

Each decision tree is constructed using the following process:

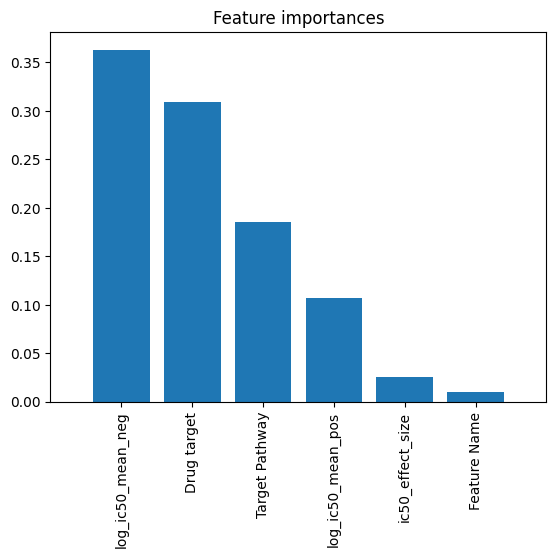
* **Bootstrap Aggregation (Bagging):** Random subsets of the training data are created with replacement.
* **Random Feature Selection:** At each split in the tree, a random subset of features is considered to determine the best split, promoting diversity among the trees.

#### Voting Mechanism:

The final prediction is made using a majority voting mechanism across all the trees in the forest. For classification, each tree votes for a class, and the class with the majority votes is selected as the final prediction.

#### Feature Importance:

Random Forest provides a measure of feature importance by averaging the decrease in impurity (e.g., Gini impurity or entropy) across all trees whenever a feature is used to split a node.



# Cross Validation

A statistical technique called cross-validation is employed to assess a machine learning model's performance. The process entails dividing the dataset into smaller groups, using some of the smaller groups to train the model, and using the remaining subsets to validate it. To provide a more accurate assessment of the model's performance, this process is repeated multiple times and the results are averaged.

****One essential technique for guaranteeing the Random Forest model's resilience and generalizability is cross-validation. The dataset is split into five subsets using 5-fold cross-validation. The model is trained on four of the subsets and tested on the remaining one in a rotating fashion, repeating five times. With this method, numerous performance ratings are provided, which are then averaged to provide a trustworthy estimate of the model's accuracy.

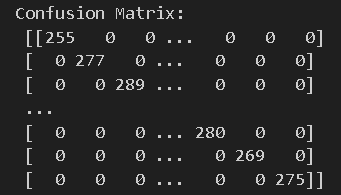
This helps evaluate how well the model can generalize to independent datasets. Cross-validation scores provide information about the Random Forest model's consistency and reliability in this project, ensuring that the model can accurately predict medication responses in newly diagnosed patients, enhancing individualized cancer treatment plans, and improving clinical outcomes.

# Confusion matrix and Classification report

An instrument for assessing a classification model's performance is a confusion matrix, which summarizes the model's accurate and inaccurate predictions in relation to the actual results. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are the four quadrants in the table.

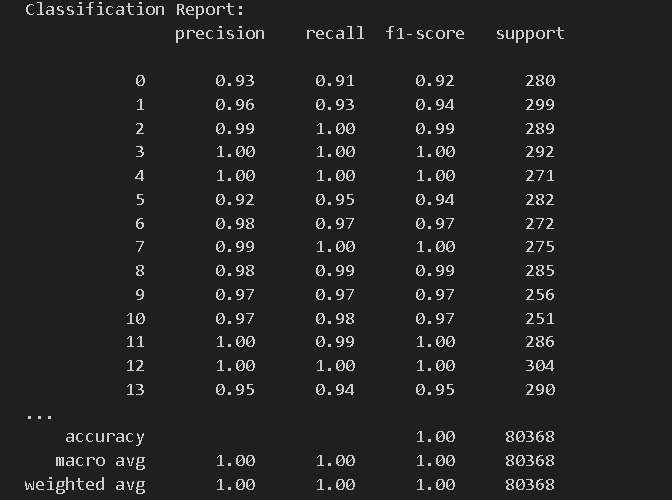
In a binary or multi-class classification task, these numbers indicate the numbers of occurrences for each class that were classified correctly and incorrectly. By displaying the number of correctly predicted answers and the areas in which the model erred, the confusion matrix aids in the visualization of the Random Forest model's accuracy in forecasting cancer medication reactions within the context of the project.

This enables a thorough analysis of the model's performance, pointing out particular instances where it can falter, including differentiating between medications with comparable efficacy profiles.

****

A thorough description of the model's performance across several measures, including as precision, recall, F1-score, and support for each class, is given in the classification report. The F1-score is the harmonic mean of precision and recall, providing a balance between the two. Precision evaluates the accuracy of positive predictions, or the proportion of anticipated positives that are really correct. Recall represents the capacity to recognize all positive cases. Support shows how many instances of each class there are in the dataset. Within the framework of the study, the Random Forest model's categorization report offers a thorough summary of how well it predicts pharmaceutical reactions for various medications.

The paper illustrates the model's strengths and limitations by dissecting the metrics for each drug class. This helps to guide future research to increase forecast accuracy and inform judgments about model enhancements, which in turn leads to more effective and individualized cancer treatment strategies.

****

# Result

The dataset is split into training and testing sets, with 160,736 samples for training and 40,184 samples for testing. The features included in the model are 'Drug target', 'Target Pathway', 'Feature Name', 'ic50 effect size', 'log ic50 mean pos', and 'log ic50 mean neg'. The Random Forest model achieved a slightly higher accuracy compared to the logistic regression model, reflecting its ability to capture complex interactions between features.

# CONCLUSION

Interactive Prediction: Using particular input features, this code enables users (perhaps academics or physicians) to interactively forecast which medication will work best for a given patient. This serves as a useful example of how the machine learning models can be applied in a practical setting.

Model Comparison: Users can discern the variations in model performance and possibly select the more dependable model for their purposes by contrasting the predictions made by the random forest and logistic regression models.

Application that is easy to use: Because of the function and its execution, even people with no background in programming or machine learning can utilize the project because it offers a straightforward interface for utilizing sophisticated predictive models.