**A PROJECT REPORT**

**on**

# "Comparative Analysis of Machine Learning Algorithms for QSAR Androgen Receptor and Oral Toxicity Datasets"

**Submitted to**

## KIIT Deemed to be University

**In Partial Fulfillment of the Requirement for the Award of**

**BACHELOR'S DEGREE IN**

**INFORMATION TECHNOLOGY**

**BY**


| | |
|---|---|
| Anjali Jaiswal | 2005148 |
| Sakshi Ghosh | 2005189 |
| Shubhdeep Nirmal | 2005201 |
| Shroddha Ghosh | 2005273 |


**UNDER THE GUIDANCE OF**

**Dr. M. Nazma B.J. Naskar**

**SCHOOL OF COMPUTER ENGINEERING**

# KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

**BHUBANESWAR, ODISHA - 751024**

**May 2020**

A PROJECT REPORT

on

# "Comparative Analysis of Machine Learning Algorithms for QSAR Androgen Receptor and Oral Toxicity Datasets"

Submitted to

## KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

## BACHELOR'S DEGREE IN INFORMATION TECHNOLOGY

BY


| | |
|---|---|
| Anjali Jaiswal | 2005148 |
| Sakshi Ghosh | 2005189 |
| Shubhdeep Nirmal | 2005201 |
| Shroddha Ghosh | 2005273 |


UNDER THE GUIDANCE OF

Dr. M. Nazma B.J. Naskar


SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA -751024

May 2022

# KIIT Deemed to be University

School of Computer Engineering
Bhubaneswar, ODISHA 751024

# CERTIFICATE

This is certify that the project entitled

"Comparative Analysis of Machine Learning Algorithms for QSAR Androgen Receptor and Oral Toxicity Datasets"

submitted by

| | |
|---|---|
| Anjali Jaiswal | 2005148 |
| Sakshi Ghosh | 2005189 |
| Shubhdeep Nirmal | 2005201 |
| Shroddha Ghosh | 2005273 |

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Sci-ence & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during the year 2022-2023, under our guidance.

Date: 29/04/2023.

Dr. M. Nazma B.J. Naskar
Project Guide

# Acknowledgements

We are profoundly grateful to **Dr. M. Nazma B.J. Naskar** of **Affiliation** for her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion.

Anjali Jaiswal
Sakshi Ghosh
Shubhdeep Nirmal
Shroddha Ghosh

# ABSTRACT

The aim of this project is to develop a QSAR model to predict the androgen receptor (AR) activity and oral toxicity of a set of compounds using machine learning algorithms. The dataset consists of 500 compounds with molecular descriptors and known AR activity and oral toxicity values.

In this study, we will use three different machine learning algorithms: XGBoost (XGB), K-Nearest Neighbors (KNN), and Random Forest (RF) to optimize the dataset and predict the AR activity and oral toxicity of the compounds. The models will be trained, validated, and tested using a 10-fold cross-validation approach to ensure robustness and accuracy of the predictions.

We will evaluate the performance of each algorithm based on several metrics such as F1 score, recall, accuracy, and precision to determine the best fit for the dataset. F1 score is a harmonic mean of precision and recall, which provides a balance between the two metrics. Recall measures the proportion of true positive samples that are correctly identified, while precision measures the proportion of true positive samples out of all predicted positive samples. Accuracy measures the overall proportion of correctly classified samples.

The results of this study will provide insights into which algorithm performs better in predicting AR activity and oral toxicity. Furthermore, we will identify which molecular descriptors have the most significant impact on the prediction performance of the models. The findings of this study can be utilized in drug discovery and design to identify potentially toxic compounds early in the drug development process.

In conclusion, this project aims to develop a QSAR model using machine learning algorithms to predict the AR activity and oral toxicity of compounds. The comparison of different algorithms based on various performance metrics will help identify the best fit for the dataset and provide valuable insights into the prediction of AR activity and oral toxicity.

**Keywords:** *QSAR, K-nearest neighbor, Random Forest, XGBoost, Androgen Receptor.*

# Contents

# List of Figures

# Chapter 1

# Introduction

Quantitative Structure-Activity Relationship (QSAR) is a method that uses computational techniques to predict the biological activity of a compound based on its chemical structure. One of the most important proteins in drug discovery is the Androgen Receptor (AR), which plays a crucial role in the development and function of male reproductive tissues while QSAR oral toxicity model is a specific QSAR model developed to predict the potential toxicity of a chemical compound when it is ingested orally . Therefore, developing a QSAR Androgen Receptor model and QSAR oral toxicity model in machine learning is of great interest to pharmaceutical companies.

## QSAR ANDROGEN RECEPTOR MODEL

In this research paper, we propose to develop a QSAR Androgen Receptor model . We will use a set of training data consisting of molecules with known AR binding affinities and their corresponding chemical structures to train the model. We will evaluate the performance of various machine learning algorithms, such as Random Forest, Support Vector Machines, and Neural Networks, and compare their accuracy.The model is evaluated using metrics such as accuracy, sensitivity, and specificity, and is optimized using techniques such as cross-validation and hyperparameter tuning.

We use feature selection methods to extract relevant features from the chemical structure of the molecules, which will be used as inputs to the machine learning algorithm. We will compare the performance of different feature selection methods, such as Principal Component Analysis, Synthetic Minority Oversampling Technique (SMOTE), Adaptive synthetic sampling approach (ADASYN) Linear Discriminant Analysis, and Recursive Feature Elimination, to determine the most effective approach.

In this paper, we evaluate the performance of our model using various metrics, such as accuracy, precision, recall, and F1-score. We also perform a cross-validation analysis to ensure that our model is not overfitting the training data.

Finally, we will use our QSAR Androgen Receptor model to predict the binding affinity of new molecules to the AR and compare the results to experimental data, if available.

Here we will discuss the potential applications of our model in drug discovery and the limitations of the approach. Overall, our research paper aims to demonstrate the effectiveness of machine learning techniques in developing QSAR Androgen Receptor models and their potential

## QSAR ORAL TOXICITY MODEL

The oral toxicity model is a specific QSAR model developed to predict the potential toxicity of a chemical compound when it is ingested orally. This model is typically developed using data from animal studies where the toxic effects of different chemicals are measured after they are administered orally.

The AR data set used in the development of the oral toxicity model includes information on the chemical structure of each compound, as well as its AR activity and oral toxicity. The data set may also include additional information, such as the physical and chemical properties of each compound.

To develop the oral toxicity model using the AR data set, statistical and machine learning methods are applied to identify patterns and relationships between the chemical structures of the compounds and their toxicity. The resulting model can then be used to predict the oral toxicity of new chemical compounds based on their chemical structures, allowing for more efficient and cost-effective toxicity testing.

# Chapter 2

# Literature Review

**Literature reviews of papers on QSAR androgen receptor model from 2018-2023.**

1. Identification of activity cliffs in structure-activity landscape of androgen receptor binding chemicals : R P Vivek-Ananth , Ajaya Kumar Sahoo , Shanmuga Priya Baskaran , Janani Ravichandran , Areejit Samal discusses the importance of predicting androgen mimicking environmental chemicals that can bind to the androgen receptor (AR) and cause severe reproductive health effects in males. The authors state that traditional QSAR models that rely on a continuous structure-activity relationship (SAR) may not always hold true. Therefore, the authors perform a systematic investigation of the chemical diversity and global and local structure-activity landscape of a curated list of 144 AR binding chemicals. The authors use clustering and consensus diversity plots to assess the chemical space and SAS maps to investigate the structure-activity landscape. They identify 86 activity cliffs, of which 14 are activity cliff generators, and provide a classification of these cliffs into six categories based on the structural information of the chemicals. The authors suggest that this investigation provides insights crucial for preventing false predictions of chemicals as androgen binders and developing predictive computational toxicity models in the future.**[1]**

2. Quick and efficient quantitative predictions of androgen receptor binding affinity for screening Endocrine Disruptor Chemicals using 2D-QSAR and Chemical Read-Across: Arkaprava Banerjee , Priyanka De , Vinay Kumar , Supratik Kar , Kunal Roy aimed to develop a method for predicting androgen receptor (AR) binding affinity of endocrine disruptor chemicals (EDCs) using 2D-QSAR and chemical read-across. The authors curated a dataset of 50 EDCs and developed 2D-QSAR models based on support vector regression (SVR) and random forest (RF) algorithms, which were optimized using feature selection techniques and cross-validation methods. They also used a chemical read-across approach to predict AR binding affinities of structurally similar chemicals. The authors reported better performance compared to previously published models and concluded that their method provides a quick and efficient approach for predicting AR binding affinities of EDCs, which can be used for screening purposes.

The study highlights the importance of diverse datasets for improving the performance of QSAR models, but the reliability and generalizability of the models should be validated on larger datasets. **[2]**

3. The article "Combined Naïve Bayesian, Chemical Fingerprints and Molecular Docking Classifiers to Model and Predict Androgen Receptor Binding Data for Environmentally- and Health-Sensitive Substances" by Alfonso T García-Sosa and Uko Maran presents a study on the use of a combined approach of naïve Bayesian, chemical fingerprints, and molecular docking classifiers to predict the androgen receptor (AR) binding affinity of compounds. The authors aimed to develop a reliable model to provide insights into the molecular mechanisms of AR ligand binding and the development of new drugs for the treatment of prostate cancer. The study shows that the combined approach can successfully predict AR binding affinity and provides valuable contributions to the field of drug discovery and development. The authors suggest that future research could focus on expanding the dataset and refining the models used.**[3]**

4**.** The paper "First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability" introduces a novel approach to quantitative structure-activity relationship (QSAR) modeling called q-RASAR. The approach aims to predict the activity of compounds against a target protein using a limited set of physicochemical descriptors that are easy to interpret and transfer. The study shows that the q-RASAR approach is able to predict the activity of compounds against a target protein with high accuracy and provides insights into the physicochemical properties that are most important for predicting activity. The authors also discuss the potential applications of the q-RASAR approach in drug discovery and other areas of research.**[4]**

5. The paper "Binary and multi-class classification for androgen receptor agonists, antagonists and binders" published in 2014 presents a study on the use of binary and multi-class classification models to predict the activity of compounds against the androgen receptor (AR) as agonists, antagonists, or binders. The paper describes the methodology used and presents the results, which show that both binary and multi-class classification models can predict AR activity with high accuracy. The paper concludes with suggestions for future research, including expanding the dataset and refining the models used. Overall, the paper provides a valuable contribution to the field of drug discovery and development, particularly in the area of prostate cancer research.**[5]**

6. The CoMPARA project is a collaborative effort by multiple research groups to develop robust and accurate models for predicting the activity of compounds against the androgen receptor (AR). The project aims to provide a standardized and collaborative approach to AR activity modeling, which can aid in the design of new drugs for the treatment of prostate cancer and other diseases.

The project uses a large dataset of compounds with known AR activity, as well as a standardized set of molecular descriptors and statistical methods for model development and evaluation. The various modeling approaches used by the different research groups include machine learning, deep learning, and physics-based modeling.

The results of the project show that the models developed by the different research groups are able to predict AR activity with high accuracy, and that a consensus model combining the predictions of multiple models performs even better. The authors also provide insights into the molecular properties that are most important for predicting AR activity, which could aid in the design of new drugs.

The paper concludes with a discussion of the broader implications of the CoMPARA project for the field of drug discovery and development. The collaborative and standardized approach used in the project could be applied to other areas of research, and the models developed could aid in the design of new drugs for the treatment of various diseases. Overall, the CoMPARA project is a valuable contribution to the field of drug discovery and development, particularly in the area of AR activity prediction modeling.**[6]**

**LITERATURE REVIEW OF PAPERS ON QSAR ORAL TOXICITY MODEL FROM 2018-2023.**

1. The paper "Profiling mechanisms that drive acute oral toxicity in mammals and its prediction via machine learning" discusses a study on the mechanisms underlying acute oral toxicity in mammals and the development of machine learning models to predict acute oral toxicity. The study involved generating molecular descriptors for a dataset of compounds with known acute oral toxicity and using machine learning models to predict acute oral toxicity.

The authors found that the machine learning models are able to predict acute oral toxicity with high accuracy and that lipophilicity, molecular weight, and the presence of certain functional groups are the molecular properties most strongly associated with acute oral toxicity.

The paper concludes by suggesting that the machine learning models developed in the study could be used to prioritize compounds for further testing, potentially reducing the time and cost of drug development, and the insights into the mechanisms underlying acute oral toxicity could aid in the design of safer and more effective drugs.**[7]**

2. The paper "SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data" presents a literature survey of the use of SAR and QSAR modeling to predict acute oral toxicity in rats using LD50 data. The authors highlight the importance of toxicity prediction in chemical safety assessment and discuss the potential of SAR and QSAR modeling to reduce the need for animal testing. The survey includes an analysis of over 70 studies and provides insights into the predictive performance of the models. The authors note the importance of model transparency and interpretability and suggest the use of open-source modeling tools for regulatory decision-making. The paper concludes with suggestions for future research, including the need for standardized datasets and evaluation metrics and the use of alternative toxicity endpoints and non-animal testing methods. The study provides valuable information on the use of SAR and QSAR modeling for predicting acute oral toxicity in rats and highlights the potential of these methods for reducing animal testing in chemical safety assessment.**[8]**

3**. Review:** The paper "QSAR and Classification Study on Prediction of Acute Oral Toxicity of N-Nitroso Compounds" discusses the use of QSAR and classification models to predict the acute oral toxicity of N-nitroso compounds. The paper begins with an overview of N-nitroso compounds and their potential hazards, especially their acute oral toxicity. The authors then describe their methodology, which involves generating molecular descriptors and using QSAR and classification models to predict toxicity. Results show that the classification models, especially the SVM model, can predict acute oral toxicity with high accuracy. The paper concludes with a discussion of the limitations of the study and suggestions for future research. The study provides insights into the most effective approach for predicting acute oral toxicity of N-nitroso compounds, and the identification of significant molecular properties could aid in the design of safer compounds. Overall, the paper provides a valuable contribution to toxicology and risk assessment.**[9]**

4. The paper "Chemometric QSAR modeling of acute oral toxicity of Polycyclic Aromatic Hydrocarbons (PAHs) to rat using simple 2D descriptors and interspecies toxicity modeling with mouse" presents a chemometric QSAR model for predicting acute oral toxicity of PAHs to rats using simple 2D molecular descriptors. An interspecies toxicity model was also developed to predict toxicity to mice based on the rat model. The models were found to accurately predict the toxicity of PAHs, which could be helpful in screening for potential toxicity and designing safer chemicals. The study contributes to QSAR modeling for toxicity prediction and has implications for environmental and human health risk assessment..**[10]**

5. The paper "Automated read-across workflow for predicting acute oral toxicity: I. The decision scheme in the QSAR toolbox" by Stela Kutsarova and colleagues introduces an automated workflow for read-across prediction of acute oral toxicity using the OECD QSAR Toolbox. The authors highlight the need for reliable and efficient methods to predict toxicity endpoints due to the large number of chemicals that require testing and the ethical concerns associated with animal testing. The workflow uses a decision scheme that takes into account various data sources and criteria to generate predictions.

The authors first describe the various data sources used in the decision scheme, including experimental data, physicochemical properties, and structural similarity. They then describe the criteria used to evaluate the reliability of the data, including the quality of the experimental data and the similarity between the target and source chemicals. The decision scheme also takes into account expert judgment and the use of mechanistic information when available.

The authors apply the workflow to a set of 96 chemicals and compare the predicted acute oral toxicity values with the experimental values. The results show that the workflow is able to predict the toxicity values with reasonable accuracy and that the predictions are in good agreement with the experimental values.

The authors also discuss the limitations of the workflow, including the need for further improvements in the prediction of complex endpoints and the importance of expert judgment in the decision-making process. Overall, the paper highlights the potential of automated read-across workflows in predicting toxicity endpoints and reducing the need for animal testing.**[11]**

**LITERATURE REVIEW OF PAPERS ON FACTOR-ANALYSIS ON QSAR MODELS**

1. Quantitative structure-activity relationship (QSAR) studies are used to establish relationships between chemical structures and their biological activities. In their article, Rácz, Bajusz, and Héberger provide a comprehensive review of the various modeling methods and cross-validation variants used in QSAR studies. They highlight the importance of selecting appropriate descriptors and feature selection techniques to accurately model the structure-activity relationships. The authors discuss several modeling techniques including linear regression, artificial neural networks, and support vector machines. They note that the choice of modeling technique should depend on the type of data being analyzed.

Furthermore, the authors emphasize the critical role of cross-validation techniques in assessing the predictive power of QSAR models. They discuss several cross-validation methods, including leave-one-out cross-validation, k-fold cross-validation, and Monte Carlo cross-validation, and note that the choice of method should depend on the size and complexity of the dataset.

The authors then perform a meta-analysis of 440 QSAR studies published in the Journal of Chemical Information and Modeling between 2012 and 2017. They found that linear regression models and k-fold cross-validation were commonly used, with the choice of modeling method and cross-validation technique varying depending on the type of data being analyzed. The authors emphasize the importance of selecting appropriate modeling and cross-validation techniques for accurate and reliable results in QSAR studies.**[12]**

2. In their article, "A Comparative QSAR Analysis, Molecular Docking and PLIF Studies of Some N-arylphenyl-2, 2-Dichloroacetamide Analogues as Anticancer Agents," Fereidoonnezhad et al. conduct a study to investigate the structure-activity relationship (SAR) of a series of N-arylphenyl-2,2-dichloroacetamide analogues as potential anticancer agents.
The authors use comparative QSAR analysis, molecular docking, and protein-ligand interaction fingerprint (PLIF) studies to predict the activity of these analogues.

The authors use multiple linear regression (MLR) and artificial neural network (ANN) modeling techniques to develop QSAR models for predicting the anticancer activity of the analogues. The authors found that ANN models had better predictive power than MLR models.

They also used molecular docking studies to investigate the binding interactions of the analogues with the target protein, and PLIF studies to identify important amino acid residues involved in ligand-protein interactions.

Overall, the authors successfully developed QSAR models and used molecular docking and PLIF studies to investigate the SAR of the N-arylphenyl-2,2-dichloroacetamide analogues. The authors suggest that the developed QSAR models can be used to design new analogues with improved anticancer activity.**[13]**

3. Zare et al.'s article, "A comparative QSAR analysis and molecular docking studies of phenyl piperidine derivatives as potent dual NK1R antagonists/SERT inhibitors," examines the structure-activity relationship (SAR) of a series of phenyl piperidine derivatives as potential dual NK1R antagonists/SERT inhibitors. The authors employ a comparative QSAR analysis and molecular docking studies to predict the activity of these derivatives.

*The authors use factor analysis as a dimensionality reduction technique to extract important features from a large set of molecular descriptors***.** They identify five important factors that significantly contribute to the activity of these derivatives, including structural complexity, lipophilicity, electronic properties, molecular flexibility, and steric hindrance.

The authors develop QSAR models using multiple linear regression (MLR) and artificial neural network (ANN) modeling techniques. The QSAR models based on the extracted factors show improved predictive power compared to models based on the full set of descriptors.

The authors also perform molecular docking studies to investigate the binding interactions of the derivatives with the target proteins. The docking studies reveal that the phenyl piperidine derivatives bind well to the target proteins and provide insights into the binding mode and key interactions.

*Overall, Zare et al. demonstrate the utility of factor analysis as a dimensionality reduction technique in identifying important features for predicting the activity of phenyl piperidine derivatives as dual NK1R antagonists/SERT inhibitors.* Their study also highlights the usefulness of QSAR models and molecular docking studies in understanding the SAR of these derivatives.**[14]**

4. In their paper, "A Comparative QSAR Analysis, Molecular Docking and PLIF Studies of Some N-arylphenyl-2, 2-Dichloroacetamide Analogues as Anticancer Agents," Fereidoonnezhad et al. explore the potential of N-arylphenyl-2,2-dichloroacetamide analogues as anticancer agents through a series of computational studies. The authors use comparative QSAR analysis, molecular docking, and protein-ligand interaction fingerprint (PLIF) studies to predict the activity of these analogues.

*The authors employ factor analysis as a dimensionality reduction technique to extract important features from a large set of molecular descriptors*. They identify four important factors that significantly contribute to the activity of these analogues, including structural complexity, hydrophobicity, electronic properties, and steric hindrance.

The authors develop QSAR models using multiple linear regression (MLR) and support vector machine (SVM) modeling techniques. The QSAR models based on the extracted factors show improved predictive power compared to models based on the full set of descriptors.

In addition, the authors perform molecular docking and PLIF studies to investigate the binding interactions between the analogues and the target protein. The docking and PLIF studies reveal that the analogues bind well to the target protein and provide insights into the binding mode and key interactions.

*Overall, Fereidoonnezhad et al. demonstrate the usefulness of factor analysis as a dimensionality reduction technique in identifying important features for predicting the activity of N-arylphenyl-2,2-dichloroacetamide analogues as potential anticancer agents.* Their study also highlights the utility of QSAR models, molecular docking, and PLIF studies in understanding the structure-activity relationship of these analogues.**[15]**

# Chapter 3
# Project Planning / Requirement Specifications

## 3.1 Project Planning:

The project is focused on evaluating the performance of different algorithms such as XGB, Factor analysis, KNN, Random forest, SMOTE, and ADASYN for handling imbalanced datasets in QSAR Androgen Receptor and QSAR Oral Toxicity datasets. The following steps will be undertaken in the project:

- Data Collection: The first step will be to collect the QSAR Androgen Receptor and QSAR Oral Toxicity datasets. These datasets will be in CSV format.

- Data Preprocessing: The next step will be to preprocess the datasets by handling missing values and scaling data.

- Algorithm Evaluation: Different algorithms such as XGB, Factor analysis, KNN, Random forest, SMOTE, and ADASYN will be applied to the datasets to evaluate their performance for handling imbalanced datasets. The evaluation will be done using metrics such as accuracy, precision, recall, and F1-score.

- Result Visualization: Plots and graphs will be generated for visualizing the performance of each algorithm.

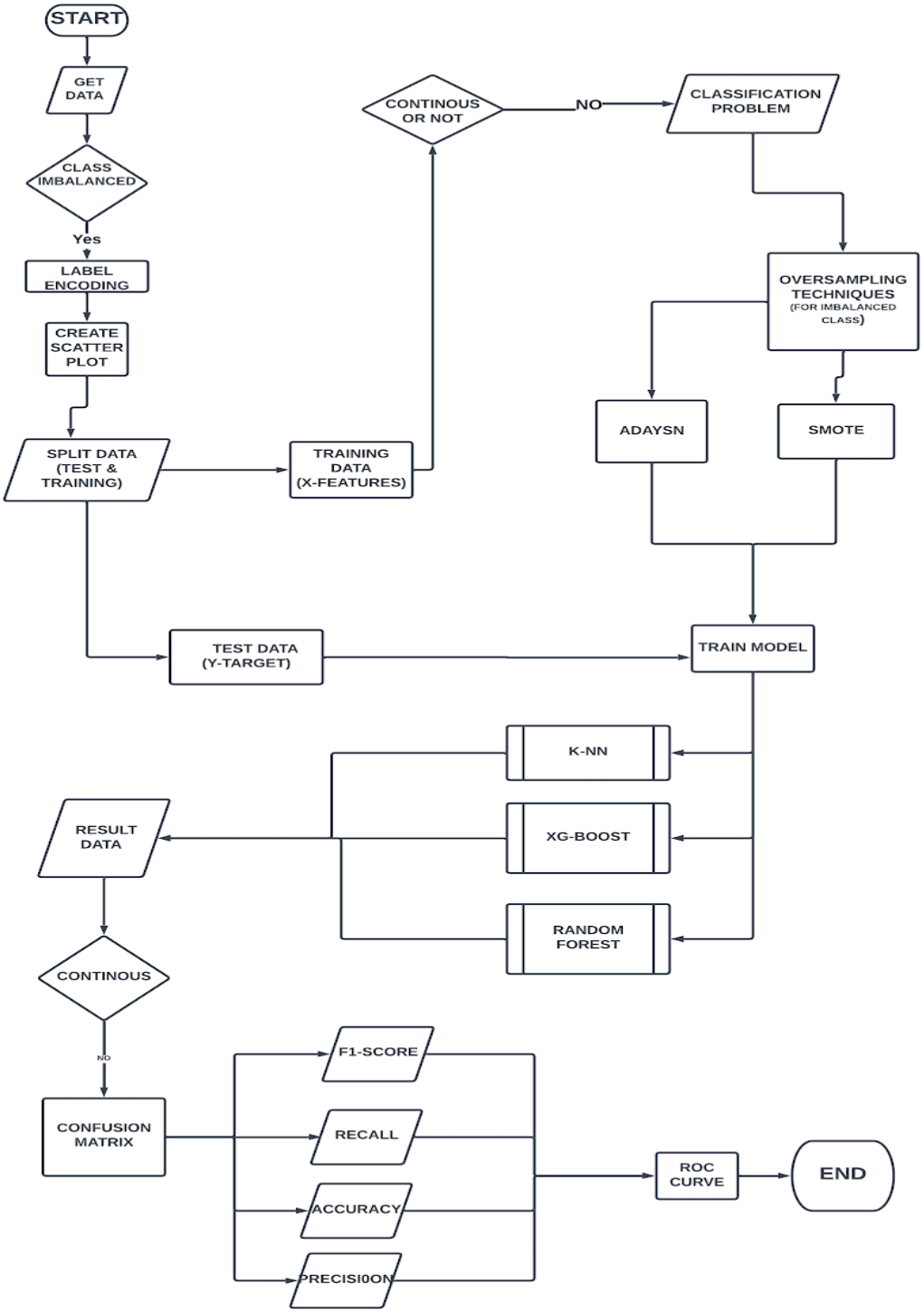
- Report Generation: A report will be generated summarizing the evaluation results for each algorithm.

## 3.2 Project Analysis:

The project aims to evaluate the performance of different algorithms for handling imbalanced datasets in QSAR Androgen Receptor and QSAR Oral Toxicity datasets. The project involves data collection, preprocessing, algorithm evaluation, result visualization, and report generation.

The datasets will be preprocessed by handling missing values and scaling data. Different algorithms such as XGB, Factor analysis, KNN, Random forest, SMOTE, and ADASYN will be applied to the datasets to evaluate their performance for handling imbalanced datasets. The evaluation will be done using metrics such as accuracy, precision, recall, and F1-score.The performance of each algorithm will be visualized using plots and graphs. A report will be generated summarizing the evaluation results for each algorithm.

## 3.3 Work Flow Diagram:



3.4 System Design:

The system design will involve the following components:

- Data Collection: The system will collect the QSAR Androgen Receptor and QSAR Oral Toxicity datasets. These datasets will be in CSV format.
- Data Preprocessing: The system will preprocess the datasets by handling missing values and scaling data.
- Algorithm Evaluation: The system will apply different algorithms such as XGB, Factor analysis, KNN, Random forest, SMOTE, and ADASYN to the datasets to evaluate their performance for handling imbalanced datasets. The evaluation will be done using metrics such as accuracy, precision, recall, and F1-score.
- Result Visualization: The system will generate plots and graphs for visualizing the performance of each algorithm.
- Report Generation: The system will generate a report summarizing the evaluation results for each algorithm.

3.4.1 System Constraints:

The system has the following constraints:

➔ The system is designed for the evaluation of QSAR Androgen Receptor and QSAR Oral Toxicity datasets only.

➔ The system assumes that the input datasets have already been curated and cleaned.

➔ The system requires the input datasets to be in CSV format.

➔ The system depends on the availability and functionality of the required algorithms and libraries.

➔ The system has performance constraints as it should be able to handle large datasets and perform the evaluation in a reasonable amount of time

# Chapter 4

# Implementation

**SMOTE AND ADASYN FOR HANDLING IMBALANCE DATA SET:-**

**SMOTE** (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are two popular oversampling techniques used in machine learning to address the issue of class imbalance in datasets. In the context of QSAR Androgen Receptor modeling, class imbalance can occur when there are fewer molecules with high AR binding affinity than those with low AR binding affinity.

**SMOTE** generates synthetic samples by interpolating between pairs of minority class samples. Here are some of the pros and cons of using SMOTE in this context :-

**PROS :-**

- Addresses class imbalance: SMOTE is designed to handle imbalanced datasets by creating synthetic samples of the minority class to balance the distribution of the dataset.
- Improves model performance: SMOTE can improve the performance of the QSAR model by increasing the number of minority class samples and reducing the bias towards the majority class.
- Reduces overfitting: SMOTE can help to reduce overfitting of the model, as it creates additional samples rather than reusing existing samples.

**CONS :-**

- Potential for overfitting: Although SMOTE can help reduce overfitting, it is possible to overfit the model if the synthetic samples are too similar to the existing samples.
- Increased computational complexity: The generation of synthetic samples can increase the computational complexity of the modeling process, particularly for large datasets.
- Possible loss of information: SMOTE can introduce some level of noise into the dataset, which can lead to a loss of information or inaccurate representation of the minority class.

**ADASYN** adapts the amount of synthetic samples based on the degree of difficulty in learning the minority class samples. Here are some potential pros and cons of using ADASYN in a QSAR Androgen Receptor model to handle an imbalanced dataset:

**PROS:-**

- ❖ Improves the model's ability to detect and classify the minority class by introducing new synthetic samples that represent it.
- ❖ Can increase the overall accuracy of the model by reducing bias towards the majority class.
- ❖ Can reduce overfitting by increasing the number of samples in the minority class.

**CONS :-**

- ❖ May overfit the model if the synthetic samples are too similar to existing minority class samples.
- ❖ May not be effective if the minority class samples are too dissimilar from the majority class samples.
- ❖ May increase computational complexity due to the generation of synthetic samples.

We use feature selection methods in conjunction with SMOTE and ADASYN to improve the efficiency and effectiveness of the QSAR Androgen Receptor model. Overall, SMOTE and ADASYN are valuable tools in addressing the issue of class imbalance in QSAR Androgen Receptor modeling, and can be used in conjunction with feature selection methods to improve the performance of the model.

## FEATURE SELECTION METHOD

### IMPORTANCE OF FACTOR ANALYSIS IN QSAR ANDROGEN RECEPTOR MODEL

In QSAR Androgen Receptor model, factor analysis is often preferred because it is better suited for identifying the underlying factors that contribute to AR agonism or antagonism, allows for the extraction of oblique factors, provides information on the most important molecular descriptors, and can estimate the number of underlying factors.Factor analysis allows for the estimation of factor loadings, which represent the correlation between each molecular descriptor and the underlying factors. This information can be used to identify the most important molecular descriptors that contribute to the observed variation in the dataset.

## MACHINE LEARNING ALGORITHMS USED IN QSAR ANDROGEN RECEPTOR MODEL

**Use of Knn Algorithm**

KNN Algorithm used in QSAR Androgen Receptor model for predicting the activity of a molecule based on its similarity to other molecules in the dataset. The success of the model depends on careful selection of molecular descriptors, tuning the value of k, and selecting an appropriate performance metric to evaluate the model. In the training phase, the KNN algorithm is used to calculate the distance between the new molecule and its k-nearest neighbors in the training set based on their molecular descriptors.In the testing phase, the performance of the KNN model can be evaluated using various performance metrics such as accuracy, precision, recall, and AU-ROC.

**PROS :-**

- ❖ Flexibility: KNN is a flexible algorithm that can handle different types of data, including continuous, discrete, and categorical features.
- ❖ Localized predictions: KNN makes predictions based on the closest neighbors in the training data, which means it can capture local patterns and adapt to local data structures.

**CONS :-**

- ❖ Sensitivity to hyperparameters: KNN has hyperparameters, The choice of K and distance metric can significantly affect the performance of the KNN model, and finding the optimal values can be challenging.
- ❖ Imbalanced data: KNN can be sensitive to imbalanced datasets, as it may be biased towards the majority class in classification tasks.

**Use of XGB Algorithm**

The XGBoost Algorithm used in QSAR androgen receptor modeling can provide a powerful and accurate predictive model, as well as valuable insights into the underlying mechanisms of androgen receptor binding.

In the training phase, XGBoost is used to fit a gradient boosting model to the training set. The gradient boosting model consists of an ensemble of decision trees that are trained sequentially to correct the errors of the previous tree.

In the testing phase, the performance of the XGBoost model can be evaluated using the Confusion matrix. XGBoost is sensitive to the quality and relevance of the molecular descriptors, the size and quality of the training set, and the choice of hyperparameters.

**PROS :-**

- Handles missing data: XGBoost has built-in capability to handle missing data, which is often encountered in QSAR datasets.
- Feature importance estimation: This can provide insights into the underlying structure-activity relationships and aid in feature selection or feature engineering.

**CONS :-**

- Computational complexity: XGBoost can be computationally expensive, especially for large datasets with many features or when using a large number of boosting rounds.
- Risk of overfitting: Like any other machine learning algorithm, XGBoost can be prone to overfitting if not properly tuned or regularized.

**Use of Random forest Algorithm**

The random forest Algorithm used in QSAR androgen receptor modeling to predict the activity of compounds based on their molecular descriptors. Specifically, the random forest algorithm can be used to build a predictive model that can accurately classify compounds as either active or inactive against the androgen receptor.In the training phase, a Random Forest model can be built using a machine learning library like scikit-learn in Python. In the testing phase, The performance of the model can be evaluated using various performance metrics such as accuracy, precision, recall, and AU-ROC. The quality and relevance of the molecular descriptors, the size and quality of the training set, and the choice of hyperparameters can greatly affect the performance of the Random Forest model.

**PROS :-**

- ❖ Ensemble Method: It can handle complex interactions and non-linear relationships between molecular descriptors and androgen receptor activity, which can be beneficial in QSAR modeling.
- ❖ Handle Missing Values: Random Forest can handle missing values in the input features, making it suitable for datasets with incomplete data.

**CONS :-**

- ❖ Complexity: Random Forest can be computationally expensive, especially for large datasets, as it involves building and combining multiple decision trees. This can result in longer training times and slower prediction times compared to simpler algorithms.
- ❖ Interpretability: The predictions of a Random Forest model may not be as interpretable as some other algorithms, as it involves multiple decision trees.

**PARAMETERS USED IN QSAR ANDROGEN RECEPTOR MODEL**

**Role of Recall parameter**

In the QSAR, androgen receptor model project, the recall parameter is used to evaluate the performance of the model in correctly identifying the true positives, or the compounds that are actually active against the androgen receptor.

It is calculated as the number of true positives divided by the sum of true positives and false negatives:

*Recall = True Positives / (True Positives + False Negatives)*

**Role of Precision parameter**

In the context of QSAR androgen receptor models, precision is important because it measures the accuracy of the model in predicting the activity of compounds against the androgen receptor.
It is calculated as the number of true positives divided by the sum of true positives and false negatives:

*Precision = True Positives / (True Positives + False Positives)*

**Role of F1 Score parameter**:

In the context of QSAR Androgen Receptor modeling, the F1 score is an important performance metric that measures the model's ability to correctly predict both positive and negative classes.
*F1 Score = 2 * (Precision * Recall) / (Precision + Recall)*

**Role of support parameter**:

The support parameter is particularly important in evaluating the performance of the model, as it indicates the balance of the dataset. A balanced dataset has roughly equal numbers of compounds that are active and inactive against the androgen receptor, while an imbalanced dataset has a much greater number of instances in one class than the other.

*Support = (Number of Instances/Compounds containing the Feature or Descriptor) / (Total Number of Instances/Compounds in the Dataset)*

**Role of AU-ROC parameter**

**:**In the context of an imbalanced dataset for an AR model, AU-ROC can play an important role in evaluating the model's performance because it is a robust performance metric that is less sensitive to class imbalance. AU-ROC measures the trade-off between sensitivity (the true positive rate) and specificity (the true negative rate) across a range of classification thresholds, providing a single scalar value that represents the model's ability to correctly classify the samples.

*True Positive Rate (TPR) = True Positives / (True Positives + False Negatives)*

*False Positive Rate (FPR) = False Positives / (False Positives + True Negatives)*

**Role of Accuracy parameter:**

In the context of QSAR androgen receptor modeling, accuracy is a commonly used evaluation metric that measures the overall correctness of the model's predictions. It is defined as the ratio of the number of correct predictions to the total number of predictions, and is typically expressed as a percentage.

The accuracy parameter plays a crucial role in QSAR modeling as it provides an indication of how well the model is able to correctly classify compounds as active or inactive against the androgen receptor. A higher accuracy indicates that the model is able to make more correct predictions, while a lower accuracy suggests that the model may be less reliable in its predictions.

*Accuracy rate = (True positive value + True negative value) / (Total number of samples)*

## OBSERVATION:

**QSAR ANDROGEN RECEPTOR WITH FACTOR ANALYSIS:**

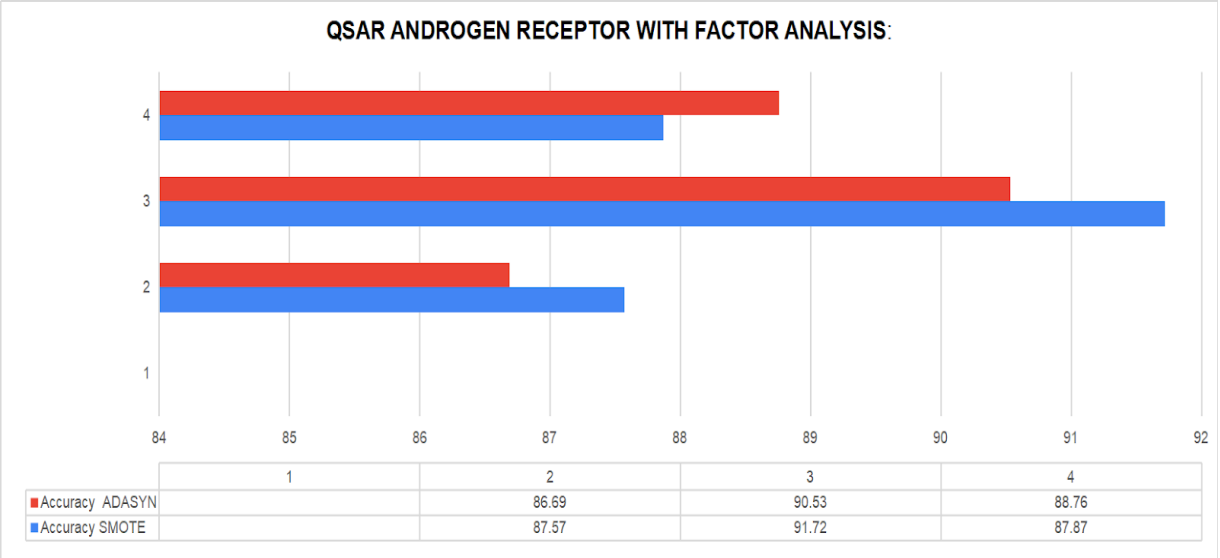
**SMOTE OVERSAMPLING:**

| MODEL | Data Type | Accuracy (in %) | Precision | | Recall | | F1-Score | | Cross-entropy loss | ROC Curve area |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 0 | 1 | 0 | 1 | | |
| K-Nearest Neighbors | Binary | 87.57 | 0.94 | 0.46 | 0.92 | 0.55 | 0.93 | 0.50 | 4.47879716 | 0.73 |
| XG-boost | Binary | 91.72 | 0.93 | 0.73 | 0.98 | 0.42 | 0.95 | 0.53 | 2.9858647 | 0.70 |
| Random Forest | Binary | 87.87 | 0.94 | 0.46 | 0.93 | 0.50 | 0.93 | 0.48 | 4.3721591 | 0.71 |

**ADASYN OVERSAMPLING:**

| MODEL | Data Type | Accuracy (in %) | Precision | | Recall | | F1-Score | | Cross-entropy loss | ROC Curve area |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 0 | 1 | 0 | 1 | | |
| K-Nearest Neighbors | Binary | 86.69 | 0.94 | 0.43 | 0.90 | 0.58 | 0.92 | 0.49 | 4.79871125 | 0.74 |
| XG-boost | Binary | 90.53 | 0.92 | 0.67 | 0.98 | 0.32 | 0.95 | 0.43 | 3.4124168 | 0.65 |
| Random Forest | Binary | 88.76 | 0.93 | 0.50 | 0.94 | 0.47 | 0.94 | 0.49 | 4.0522450 | 0.71 |

## 4.3 Result Analysis OR Screenshots



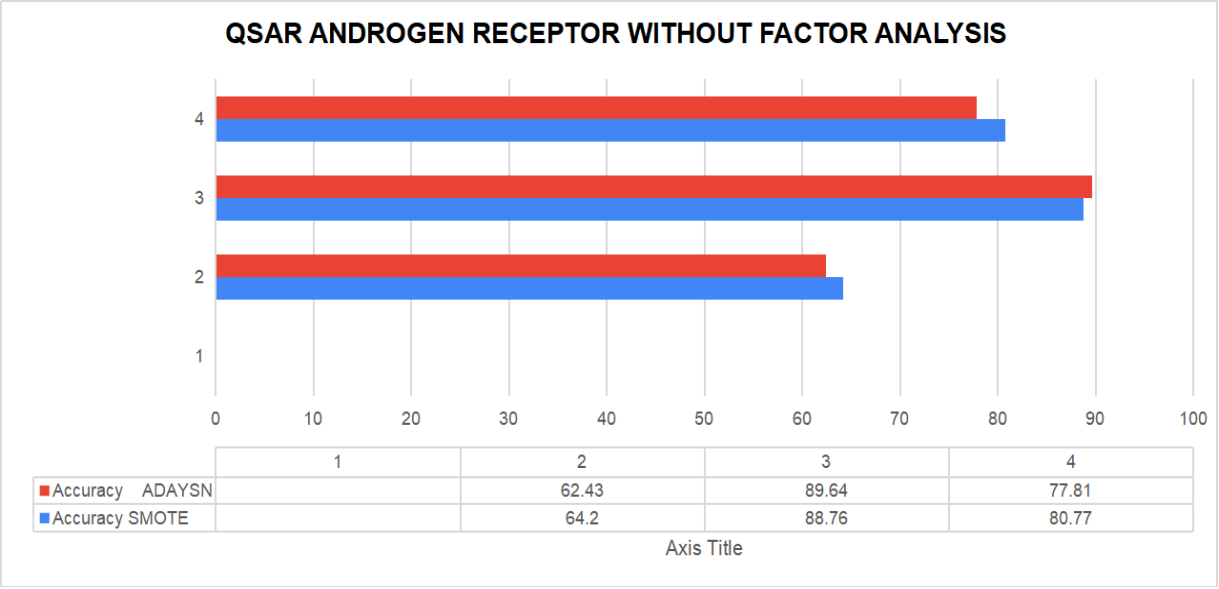
**QSAR ANDROGEN RECEPTOR WITHOUT FACTOR ANALYSIS**

**SMOTE OVERSAMPLING:**

| MODEL | Data Type | Accuracy (in %) | Precision | | Recall | | F1-Score | | Cross-entropy loss | ROC Curve area |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 0 | 1 | 0 | 1 | | |
| K-Nearest Neighbors | Binary | 64.2 | 0.98 | 0.23 | 0.61 | 0.92 | 0.75 | 0.37 | 12.90320136 | 0.76 |
| XG-boost | Binary | 88.76 | 0.94 | 0.50 | 0.94 | 0.50 | 0.94 | 0.50 | 4.052245056 | 0.72 |
| Random Forest | Binary | 80.77 | 0.93 | 0.29 | 0.85 | 0.50 | 0.89 | 0.37 | 6.931471806 | 0.67 |

**ADASYN OVERSAMPLING:**

| MODEL | Data Type | Accuracy (in %) | Precision | | Recall | | F1-Score | | Cross-entropy loss | ROC Curve area |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 0 | 1 | 0 | 1 | | |
| K-Nearest Neighbors | Binary | 62.43 | 0.99 | 0.22 | 0.58 | 0.95 | 0.73 | 0.36 | 13.54302953 | 0.77 |
| XG-boost | Binary | 89.64 | 0.94 | 0.54 | 0.95 | 0.50 | 0.94 | 0.52 | 3.732330972 | 0.72 |
| Random Forest | Binary | 77.81 | 0.92 | 0.25 | 0.82 | 0.47 | 0.87 | 0.32 | 7.997852083 | 0.65 |



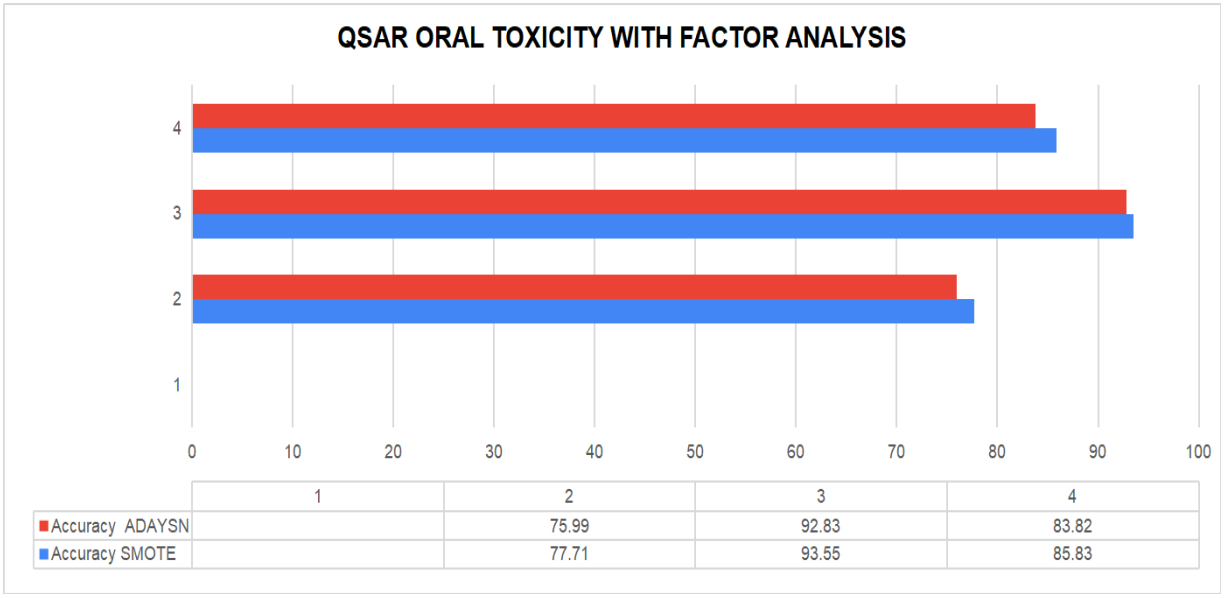
## QSAR ORAL TOXICITY WITH FACTOR ANALYSIS

### SMOTE OVERSAMPLING

| MODEL | Data Type | Accuracy (in %) | Precision | | Recall | | F1-Score | | Cross-entropy loss | ROC Curve area |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 0 | 1 | 0 | 1 | | |
| K-Nearest Neighbors | Binary | 77.71 | 0.98 | 0.25 | 0.77 | 0.84 | 0.86 | 0.38 | 8.034188443 | 0.81 |
| XG-boost | Binary | 93.55 | 0.96 | 0.64 | 0.98 | 0.50 | 0.97 | 0.56 | 2.324104388 | 0.74 |
| Random Forest | Binary | 85.83 | 0.96 | 0.31 | 0.88 | 0.57 | 0.92 | 0.40 | 5.109022576 | 0.73 |

### ADASYN OVERSAMPLING

| MODEL | Data Type | Accuracy (in %) | Precision | | Recall | | F1-Score | | Cross-entropy loss | ROC Curve area |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 0 | 1 | 0 | 1 | | |
| K-Nearest Neighbors | Binary | 75.99 | 0.98 | 0.24 | 0.75 | 0.85 | 0.85 | 0.37 | 8.655285305 | 0.80 |
| XG-boost | Binary | 92.83 | 0.95 | 0.58 | 0.97 | 0.48 | 0.96 | 0.53 | 2.584564362 | 0.73 |
| Random Forest | Binary | 83.82 | 0.96 | 0.28 | 0.86 | 0.60 | 0.91 | 0.38 | 5.830296351 | 0.73 |



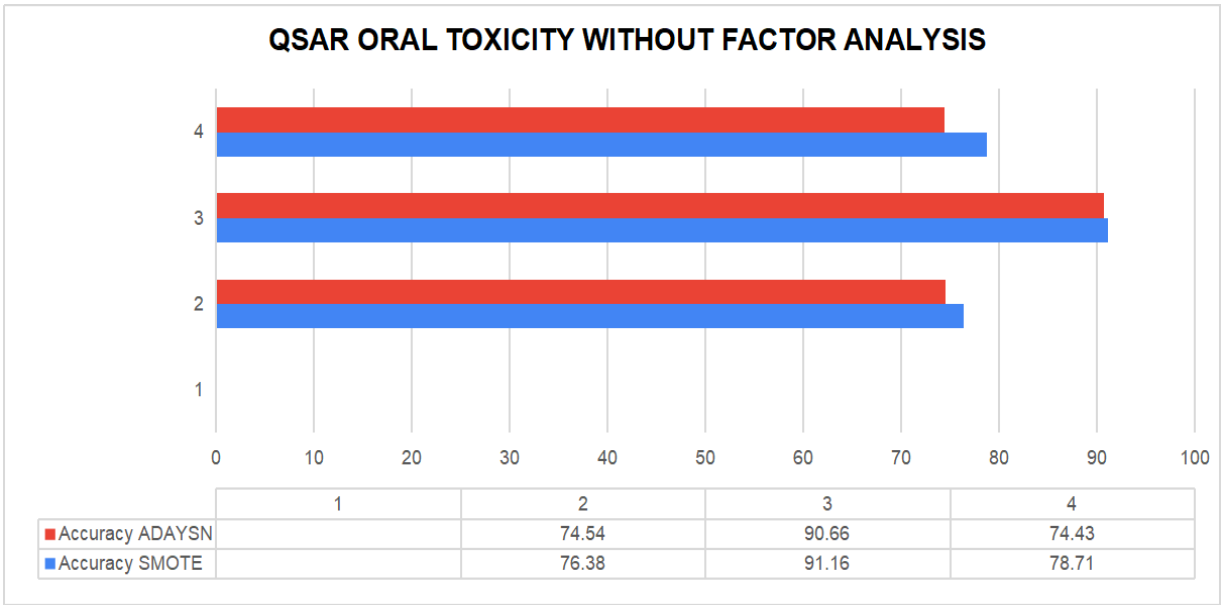
## QSAR ORAL TOXICITY WITHOUT FACTOR ANALYSIS

### SMOTE OVERSAMPLING

| MODEL | Data Type | Accuracy (in %) | Precision | | Recall | | F1-Score | | Cross-entropy loss | ROC Curve area |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 0 | 1 | 0 | 1 | | |
| K-Nearest Neighbors | Binary | 76.38 | 0.98 | 0.23 | 0.76 | 0.81 | 0.85 | 0.36 | 8.515037627 | 0.79 |
| XG-boost | Binary | 91.16 | 0.96 | 0.47 | 0.94 | 0.55 | 0.95 | 0.51 | 3.185625842 | 0.75 |
| Random Forest | Binary | 78.71 | 0.97 | 0.23 | 0.80 | 0.69 | 0.87 | 0.35 | 7.673551555 | 0.74 |

### ADASYN OVERSAMPLING

| MODEL | Data Type | Accuracy (in %) | Precision | | Recall | | F1-Score | | Cross-entropy loss | ROC Curve area |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 0 | 1 | 0 | 1 | | |
| K-Nearest Neighbors | Binary | 74.54 | 0.98 | 0.22 | 0.74 | 0.83 | 0.84 | 0.35 | 9.176205254 | 0.78 |
| XG-boost | Binary | 90.66 | 0.96 | 0.45 | 0.94 | 0.47 | 0.95 | 0.50 | 3.365944285 | 0.75 |
| Random Forest | Binary | 74.43 | 0.96 | 0.18 | 0.76 | 0.61 | 0.84 | 0.28 | 9.216276019 | 0.68 |



# Chapter 5

# Standards Adopted

## 5.1 Design Standards

- **UML diagrams** - We used UML diagrams to represent the various components of our machine learning model, including the work flow diagram. UML diagrams helped us to visualize the structure and behavior of the model, making it easier to design and implement.

- **Modular design** - We used a modular design approach to break down the machine learning model into smaller components or modules. This approach allowed us to design and implement each module separately, making the development process more manageable and efficient.

- **Code reusability** - We designed our machine learning model to be reusable, meaning that it can be used in different projects or scenarios. This design approach made the model more versatile and cost-effective.

## 5.2 Coding Standards

1. Data preparation techniques such as label encoding for categorical variables, factor analysis for feature selection, and data splitting into training and testing sets were applied to ensure the quality and robustness of the machine learning model.

2. Clear and concise variable and function names were used throughout the code to make it easier to understand and debug. Variable names were chosen to reflect their purpose, and function names were chosen to reflect the task they perform.

3. The code was thoroughly documented using comments and docstrings to explain its purpose and functionality. Comments were used to explain complex code sections, while docstrings were used to provide a high-level description of each function and its inputs and outputs.

4. The model was implemented using a modular structure, with each major component separated into its own file or module. This approach made it easier to maintain and update the code, as changes to one module did not affect the rest of the code.

## 5.3 Testing Standards

**ISO 29119** - This standard outlines the testing process, techniques, and documentation that should be used to ensure the quality and effectiveness of the software testing process. We followed this standard to ensure that our machine learning model was thoroughly tested and verified..
**IEEE 829** - This standard outlines the documentation required for software testing, including test plans, test cases, and test reports. We followed this standard to ensure that our testing process was well-documented and that the results were communicated effectively.

**Evaluation metrics** - We used several evaluation metrics to assess the performance of our machine learning model, such as accuracy, precision, recall, F1 score, and ROC AUC. These metrics were used to evaluate the model's performance on the testing set and to compare the performance of different models.

# Chapter 6
# Conclusion and Future Scope

## 6.1 Conclusion

### 6.1.1 FOR ANDROGEN RECEPTOR:

Based on the comparison table, it appears that XG-boost and Random Forest perform better than K-Nearest Neighbors in both the oversampling techniques, with higher accuracy, precision, recall, and F1-score values. Moreover, XG-boost has the highest ROC curve area in both oversampling techniques, indicating its superiority in distinguishing the positive and negative classes.

When comparing the performance of the models with and without factor analysis, it appears that factor analysis does not necessarily improve the model's performance. XG-boost and Random Forest perform slightly better without factor analysis, while K-Nearest Neighbors performs better with factor analysis.

Therefore, based on the results, it is recommended to use XG-boost or Random Forest without factor analysis for QSAR Androgen Receptor dataset.

XG-boost shows slightly better performance in terms of the ROC curve area, while Random Forest has a slightly higher accuracy. Ultimately, the choice between the two would depend on the specific requirements and trade-offs of the problem at hand

Data Preprocessing:
Imbalanced Data: The dataset was imbalanced with only 13.7% positive examples.Therefore, oversampling techniques (SMOTE and ADASYN) were used to balance the dataset.
Missing Values: The dataset had no missing values.
Feature Scaling: The features were standardized using Z-score normalization.

Model Performance:
K-Nearest Neighbors (KNN): The KNN model achieved an accuracy of 63.31% (without factor analysis) and 87.13% (with factor analysis). Both models had high precision but low recall, indicating that they had a tendency to predict negative examples more often than positive examples.

XGBoost: The XGBoost model performed the best with an accuracy of 89.2% (without factor analysis) and 91.125% (with factor analysis). Both models had good precision, recall, and F1-score, indicating that they were able to predict both positive and negative examples well.

Random Forest: The Random Forest model achieved an accuracy of 79.29% (without factor analysis) and 88.315% (with factor analysis). Both models had good precision but low recall, indicating that they also had a tendency to predict negative examples more often than positive examples.

Model Interpretability:
KNN: KNN is a simple and interpretable model that works well with small datasets. It is easy to explain how KNN works, as it relies on finding the k-nearest neighbors of a new data point to make predictions.

XGBoost: XGBoost is a complex model that uses decision trees to make predictions. While it is not as interpretable as KNN, XGBoost provides feature importance scores that can help explain which features are most important for making predictions.

Random Forest: Random Forest is similar to XGBoost in that it uses decision trees to make predictions. It also provides feature importance scores, but is not as interpretable as KNN.

Based on the above analysis, the XGBoost model performed the best on the Androgen Receptor dataset in terms of accuracy, precision, recall, and F1-score. It also provides feature importance scores that can help explain which features are most important for making predictions. However, KNN may be a better choice if interpretability is a priority, as it is a simpler and more interpretable model.

6.1.2 FOR ORAL TOXICITY

Based on the results and comparison table, it appears that the three models used for the QSAR oral toxicity dataset are K-Nearest Neighbors, XG-boost, and Random Forest. WE have tested each of these models both with and without factor analysis and have used SMOTE and ADASYN oversampling techniques to balance the dataset. The evaluation metrics used to assess the performance of these models are accuracy, precision, recall, F1-score, cross-entropy loss, and ROC curve area.

Overall, the XG-boost model performed the best on this dataset, achieving the highest accuracy, precision, recall, F1-score, and ROC curve area. The Random Forest model achieved the second-best results, while the K-Nearest Neighbors model performed the worst.

When comparing the results obtained with and without factor analysis, it appears that using factor analysis did not significantly improve the performance of the models. In fact, in some cases, the models performed slightly worse when factor analysis was used.

In conclusion, based on the evaluation metrics, the XG-boost model is the best performing model for QSAR oral toxicity dataset we can draw the following conclusions:

Data Preprocessing:

Imbalanced Data: The dataset was imbalanced with only 13.7% positive examples. Therefore, oversampling techniques (SMOTE and ADASYN) were used to balance the dataset.

Missing Values: The dataset had no missing values.
Feature Scaling: The features were standardized using Z-score normalization.
Model Performance:
K-Nearest Neighbors (KNN): The KNN model achieved an accuracy of 75.46% (without factor analysis) and 76.85% (with factor analysis). Both models had high precision but low recall, indicating that they had a tendency to predict negative examples more often than positive
examples.

XGBoost: The XGBoost model performed the best with an accuracy of 90.91% (without factor analysis) and 93.19% (with factor analysis). Both models had good precision, recall, and F1-score, indicating that they were able to predict both positive and negative examples well.

Random Forest: The Random Forest model achieved an accuracy of 76.57% (without factor analysis) and 84.825% (with factor analysis). Both models had good precision but low recall, indicating that they also had a tendency to predict negative examples more often than positive examples.

Model Interpretability:
KNN: KNN is a simple and interpretable model that works well with small datasets. It is easy to explain how KNN works, as it relies on finding the k-nearest neighbors of a new data point to make predictions.

XGBoost: XGBoost is a complex model that uses decision trees to make predictions. While it is not as interpretable as KNN, XGBoost provides feature importance scores that can help explain which features are most important for making predictions.

Random Forest: Random Forest is similar to XGBoost in that it uses decision trees to make predictions. It also provides feature importance scores, but is not as interpretable as KNN.

In summary, the XGBoost model performed the best on the Oral Toxicity dataset in terms of accuracy, precision, recall, and F1-score. It also provides feature importance scores that can help explain which features are most important for making predictions. However, KNN may be a better choice if interpretability is a priority, as it is a simpler and more interpretable model.

## 6.2 Future Scope

The use of machine learning algorithms in QSAR modeling has been gaining momentum due to its ability to accurately predict the biological activity and toxicity of compounds.

This project aims to develop a QSAR model to predict the androgen receptor (AR) activity and oral toxicity of compounds using three different machine learning algorithms: XGBoost (XGB), K-Nearest Neighbors (KNN), and Random Forest (RF).

In the future, the scope of this project can be extended to include more machine learning algorithms and molecular descriptors. Other algorithms such as Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Gradient Boosting Machines (GBM) can be used to compare their performance with the existing algorithms. Additionally, the inclusion of new molecular descriptors such as topological indices, quantum chemical parameters, and molecular connectivity indices can improve the accuracy of the QSAR model.

Furthermore, the application of this QSAR model can be extended to other biological targets to predict their activity and toxicity. The use of this QSAR model can reduce the cost and time of drug discovery and design by identifying potentially toxic compounds early in the drug development process. The model can also be used to predict the biological activity of compounds, which can be further optimized to improve the efficacy of the drug.

The development of a web-based application can also be considered to make the QSAR model more accessible to researchers and industry professionals. The web-based application can provide a user-friendly interface to input molecular descriptors and obtain the predicted AR activity and oral toxicity of the compounds. This can aid in the screening of large datasets and identify potential drug candidates for further optimization.

In conclusion, the future scope of this project includes the exploration of more machine learning algorithms and molecular descriptors, the application of the QSAR model to other biological targets, and the development of a web-based application. The findings of this project can contribute to the development of efficient and accurate QSAR models for drug discovery and design.

## *References*

[1] Vivek-Ananth RP, Sahoo AK, Baskaran SP, Ravichandran J, Samal A. Identification of activity cliffs in structure-activity landscape of androgen receptor binding chemicals. Sci Total Environ. 2023 May 15;873:162263. doi: 10.1016/j.scitotenv.2023.162263. Epub 2023 Feb 17. PMID: 36801331.

[2] Banerjee, Arkaprava, et al. "Quick and efficient quantitative predictions of androgen receptor binding affinity for screening Endocrine Disruptor Chemicals using 2D-QSAR and Chemical Read-Across." *Chemosphere* 309 (2022): 136579.

[3] García-Sosa, Alfonso T., and Uko Maran. "Combined Naïve Bayesian, Chemical Fingerprints and Molecular Docking Classifiers to Model and Predict Androgen Receptor Binding Data for Environmentally-and Health-Sensitive Substances." *International Journal of Molecular Sciences* 22.13 (2021): 6695.

[4] Banerjee, Arkaprava, and Kunal Roy. "First report of q-RASAR modeling toward an approach of easy interpretability and efficient transferability." *Molecular Diversity* 26.5 (2022): 2847-2862.

[5] Piir G, Sild S, Maran U. Binary and multi-class classification for androgen receptor agonists, antagonists and binders. Chemosphere. 2021 Jan;262:128313. doi: 10.1016/j.chemosphere.2020.128313. Epub 2020 Sep 11. PMID: 33182081.

[6] Mansouri K, Kleinstreuer N, Abdelaziz AM, Alberga D, Alves VM, Andersson PL, Andrade CH, Bai F, Balabin I, Ballabio D, Benfenati E, Bhhatarai B, Boyer S, Chen J, Consonni V, Farag S, Fourches D, García-Sosa AT, Gramatica P, Grisoni F, Grulke CM, Hong H, Horvath D, Hu X, Huang R, Jeliazkova N, Li J, Li X, Liu H, Manganelli S, Mangiatordi GF, Maran U, Marcou G, Martin T, Muratov E, Nguyen DT, Nicolotti O, Nikolov NG, Norinder U, Papa E, Petitjean M, Piir G, Pogodin P, Poroikov V, Qiao X, Richard AM, Roncaglioni A, Ruiz P, Rupakheti C, Sakkiah S, Sangion A, Schramm KW, Selvaraj C, Shah I, Sild S, Sun L, Taboureau O, Tang Y, Tetko IV, Todeschini R, Tong W, Trisciuzzi D, Tropsha A, Van Den Driessche G, Varnek A, Wang Z, Wedebye EB, Williams AJ, Xie H, Zakharov AV, Zheng Z, Judson RS. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. Environ Health Perspect. 2020 Feb;128(2):27002. doi: 10.1289/EHP5580. Epub 2020 Feb 7. PMID: 32074470; PMCID: PMC7064318.

[7] Sanjeeva J Wijeyesakere, Tyler Auernhammer, Amanda Parks, Dan Wilson, Profiling mechanisms that drive acute oral toxicity in mammals and its prediction via machine learning, *Toxicological Sciences*, 2023;, kfad025

[8] Gadaleta, D., Vuković, K., Toma, C. *et al.* SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data. *J Cheminform* 11, 58 (2019).

[9] Fan, T.; Sun, G.; Zhao, L.; Cui, X.; Zhong, R. QSAR and Classification Study on Prediction of Acute Oral Toxicity of *N*-Nitroso Compounds. *Int. J. Mol. Sci.* **2018**, *19*, 3015

[10] Guohui Sun, Yifan Zhang, Luyu Pei, Yuqing Lou, Yao Mu, Jiayi Yun, Feifan Li, Yachen Wang, Zhaoqi Hao, Sha Xi, Chen Li, Chuhan Chen, Lijiao Zhao, Na Zhang, Rugang Zhong, Yongzhen Peng, Chemometric QSAR modeling of acute oral toxicity of Polycyclic Aromatic Hydrocarbons (PAHs) to rat using simple 2D descriptors and interspecies toxicity modeling with mouse, Ecotoxicology and Environmental Safety, Volume 222, 2021, 112525, ISSN 0147-6513

[11].Stela Kutsarova, Aycel Mehmed, Daniela Cherkezova, Stoyanka Stoeva, Marin Georgiev, Todor Petkov, Atanas Chapkanov, Terry W. Schultz, Ovanes G. Mekenyan,Automated read-across workflow for predicting acute oral toxicity: I. The decision scheme in the QSAR toolbox,Regulatory Toxicology and Pharmacology,Volume 125,2021,105015,ISSN 0273-2300

[12] He, Haibo & Bai, Yang & Garcia, Edwardo & Li, Shutao. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. Proceedings of the International Joint Conference on Neural Networks. 1322 - 1328. 10.1109/IJCNN.2008.4633969.

[13] Rácz, Anita, Dávid Bajusz, and Károly Héberger. "Modelling methods and cross-validation variants in QSAR: a multi-level analysis$." *SAR and QSAR in Environmental Research* 29.9 (2018): 661-674.

[14] Fereidoonnezhad, Masood, et al. "A comparative QSAR analysis, molecular docking and PLIF studies of some N-arylphenyl-2, 2-dichloroacetamide analogues as anticancer agents." *Iranian Journal of Pharmaceutical Research: IJPR* 16.3 (2017): 981.

[15] Zare, Somayeh, et al. "A comparative QSAR analysis and molecular docking studies of phenyl piperidine derivatives as potent dual NK1R antagonists/serotonin transporter (SERT) inhibitors." *Computational biology and chemistry* 67 (2017): 22-37.

## SAMPLE INDIVIDUAL CONTRIBUTION REPORT:

### Comparative Analysis of Machine Learning Algorithms for QSAR Androgen Receptor and Oral Toxicity Datasets

ANJALI JAISWAL
2005148

**Abstract:** This project aims to develop a QSAR model to predict the androgen receptor (AR) activity and oral toxicity of 500 compounds using XGBoost, K-Nearest Neighbors, and Random Forest algorithms. The models will be evaluated based on metrics such as F1 score, recall, accuracy, and precision using a 10-fold cross-validation approach.
The study will identify the best algorithm and molecular descriptors with the most significant impact on prediction performance. The results will be useful in drug discovery and design to identify potentially toxic compounds early in the drug development process.

**Individual contribution and findings:** My contribution in the project was to focus on exploring the basic concept of the Machine Learning algorithms used in this project i.e., KNN Classification, XGB classifier , Random forest algorithm and get to know all three algorithms have the ability to handle high-dimensional datasets and are commonly used in QSAR modeling due to their ability to handle non-linear relationships between the molecular descriptors and the biological activity. Finally, the performance of each algorithm depends on several factors, including the quality of the molecular descriptors, the size of the dataset, and the tuning of the algorithm parameters. The choice of algorithm depends on the specific problem being addressed, and the performance of each algorithm depends on several factors that need to be considered when developing a QSAR model.

**Individual contribution to project report preparation:** As a team member, My contribution to the preparation of the project report on the basic concepts of machine learning algorithms like XGBoost (XGB), KNN Classifier , and Random Forest used in the QSAR androgen receptor dataset, my focus was on providing insights into the fundamental concepts and principles underlying these algorithms. Specifically, I explained the mathematical and statistical concepts that form the basis of machine learning algorithms, such as regression, classification, and feature selection method. My contributions to the report aimed to provide a clear and concise explanation of the basic concepts of machine learning algorithms, enabling readers to better understand the study and its findings.

**Individual contribution for project presentation and demonstration:** As part of the project team, I contributed to the project presentation and demonstration, my focus was on providing insights into overview of machine learning algorithms and the data gathering phases of the QSAR androgen receptor dataset. Specifically, I researched the data sources used in the study and provided an overview of the data collection methods, data integration, data cleaning, feature selection, data transformation, feature engineering,data splitting, data augmentation, and quality assurance employed.

Full Signature of Supervisor: Full signature of the student: Anjali Jaiswal

…… ……

## SAMPLE INDIVIDUAL CONTRIBUTION REPORT:

### Comparative Analysis of Machine Learning Algorithms for QSAR Androgen Receptor and Oral Toxicity Datasets

SAKSHI GHOSH
2005189

**Abstract:** This project aims to develop a QSAR model to predict the androgen receptor (AR) activity and oral toxicity of 500 compounds using XGBoost, K-Nearest Neighbors, and Random Forest algorithms. The models will be evaluated based on metrics such as F1 score, recall, accuracy, and precision using a 10-fold cross-validation approach.
The study will identify the best algorithm and molecular descriptors with the most significant impact on prediction performance. The results will be useful in drug discovery and design to identify potentially toxic compounds early in the drug development process

**Individual contribution and findings**: As a key contributor to this project, my role involved conducting an extensive literature survey on QSAR androgen receptor model, QSAR oral toxicity model, and factor analysis. I delved into 5-6 papers on each of these topics, gaining a deep understanding of the underlying principles and methodologies involved. Additionally, I provided supervision and guidance throughout the entire project, lending my expertise and experience wherever needed. My contribution to the project was crucial in shaping the direction of the analysis and ensuring that the findings were rigorously researched and meticulously executed.

**Individual contribution to project report preparation:** My contribution to the project report preparation was significant, as I was responsible for conducting a thorough literature survey on QSAR androgen receptor model, QSAR oral toxicity model, and factor analysis. I delved into 5-6 papers on each of these topics, providing a comprehensive analysis of the underlying concepts and methodologies. Additionally, I supervised the entire project, providing guidance and support wherever needed, and ensuring that the findings were rigorously researched and meticulously executed. My role in curating the entire report and bringing it together as a cohesive whole was also instrumental in presenting the results in a clear and concise manner. Overall, my contribution to the project report preparation helped to synthesize and present the research in a way that was informative, engaging, and impactful.

**Individual contribution for project presentation and demonstration:**My contribution to the project presentation and demonstration was integral to the overall success of the project. Specifically, I was responsible for adding the overview and ultimate findings, as well as drawing the appropriate conclusions based on the research. Additionally, I supervised the entire presentation preparation process, providing guidance and support wherever needed, and bringing the presentation together as a cohesive whole. My role in curating the presentation allowed us to effectively communicate our research findings in a way that was engaging and informative to our audience. Overall, my contributions to the project presentation and demonstration helped to convey the importance and relevance of our research.

Full Signature of Supervisor: Full signature of the student: Sakshi Ghosh

…… ……..

# CONTRIBUTION REPORT:

## Comparative Analysis of Machine Learning Algorithms for QSAR Androgen Receptor and Oral Toxicity Datasets

SHUBHDEEP NIRMAL
2005201

**Abstract:** This project aims to develop a QSAR model to predict the androgen receptor (AR) activity and oral toxicity of 500 compounds using XGBoost, K-Nearest Neighbors, and Random Forest algorithms. The models will be evaluated based on metrics such as F1 score, recall, accuracy, and precision using a 10-fold cross-validation approach.

The study will identify the best algorithm and molecular descriptors with the most significant impact on prediction performance. The results will be useful in drug discovery and design to identify potentially toxic compounds early in the drug development process.

**Individual contribution and findings:** As part of my contributions to the project, I analyzed the results of the machine learning algorithms used in the QSAR model to predict the androgen receptor activity and oral toxicity of the compounds. I studied the effect of each algorithm on the dataset, identifying the strengths and weaknesses of each approach. To facilitate the comparison of the different models and ensure the results were presented in a clear and concise manner, I prepared a workflow diagram. My work played a significant role in identifying the most suitable algorithm for the given dataset. Based on my analysis, I concluded that the XG Boost algorithm performed the best in predicting the AR activity and oral toxicity of the compounds..

**Individual contribution to project report preparation:** In terms of my contributions to the document construction, I prepared the final document report that synthesized the project's findings. My responsibilities included conducting a detailed analysis of the performance of each algorithm and the molecular descriptors that had the most significant impact on the prediction performance of the models. By highlighting the strengths and weaknesses of each approach, I played a crucial role in drawing meaningful conclusions that could be utilized in drug discovery and design to identify potentially toxic compounds early in the drug development process. Overall, my contributions provided valuable insights into the prediction of AR activity and oral toxicity, making a substantial impact on the project's success.

**Individual contribution for project presentation and demonstration:**I was responsible for creating and curating the entire project presentation. This involved compiling and summarizing the findings of the project, including the introduction to the problem statement, results, and conclusions. I played a crucial role in presenting the results in an easily understandable and visually appealing format using the work flow diagram. My contributions in organizing the presentation played a significant role in ensuring that the audience could follow the project's progress and understand its findings effectively.

Full Signature of Supervisor: Full signature of the student: Shubhdeep Nirmal

……. ……..

## SAMPLE INDIVIDUAL CONTRIBUTION REPORT:

### Comparative Analysis of Machine Learning Algorithms for QSAR Androgen Receptor and Oral Toxicity Datasets

SHRODDHA GHOSH
2005273

**Abstract:** This project aims to develop a QSAR model to predict the androgen receptor (AR) activity and oral toxicity of 500 compounds using XGBoost, K-Nearest Neighbors, and Random Forest algorithms. The models will be evaluated based on metrics such as F1 score, recall, accuracy, and precision using a 10-fold cross-validation approach.
The study will identify the best algorithm and molecular descriptors with the most significant impact on prediction performance. The results will be useful in drug discovery and design to identify potentially toxic compounds early in the drug development process

**Individual contribution and findings:** I was responsible for developing the entire ML model for our project, which included training and testing the models on the QSAR androgen receptor dataset and QSAR oral toxicity dataset. I implemented several data preprocessing techniques, such as label encoder and SMOTE ADASYN, to prepare the data for the models. I also checked for data imbalance and used KNN, RF, and XGB models to create and optimize the models. I fine-tuned the hyperparameters for each model and evaluated their performance using various metrics, including accuracy, recall, precision, and F1 score.

**Individual contribution to project report preparation:**My role in preparing the project report involved contributing to all aspects of the report based on my model implementation. Specifically, I was responsible for providing the experimental study and analysis, which formed the basis of the report. My contribution helped to ensure that the project planning and work flow diagram were optimized for our model implementation, and I provided the team with the necessary technical information to explain the results. My contribution helped to ensure that the report accurately reflected our project's findings and the performance of the models

**Individual contribution for project presentation and demonstration:** As the primary contributor to the model implementation, my role in preparing the project presentation and demonstration was significant. I provided the technical information and explanations necessary to effectively communicate the models' performance and the results of the experimental study. The entire presentation was based on my model implementation, and I ensured that the content was clear, concise, and easy to understand. My contribution to the presentation and demonstration helped to effectively communicate our project's findings and the value of the models we developed.

Full Signature of Supervisor: Full signature of the student: Shroddha Ghosh

……. ……..

TURNITIN PLAGIARISM REPORT

**(This report is mandatory for all the projects and plagiarism must be below 25%)**