BIOACTIVITY OF SMALL MOLECULES PREDICTION

Authors (Group 26): Laasya Vajjala (11848603), Lalith Mohan Midde (11848601), Surya Vadapalli (11861342), and Sri Varsha Adavath (11860642)
Department of Computer Science
Washington State University

Date: 04-29-2024
Technical Report: **CPTS 440/540 Introduction to Artificial Intelligence**

THOUGHT PROCESS

As a team, we embarked on the bioactivity prediction project with a shared commitment to revolutionize the drug discovery pipeline. Our collective thought process was driven by a recognition of the challenges inherent in traditional bioactivity assessment methods and the pressing need for more efficient and cost-effective alternatives.

From the outset, we approached the project with enthusiasm and determination, fueled by a common desire to make an impact in the field of drug discovery. Each team member brought a unique set of skills and perspectives to the table, which enriched our collaborative efforts and propelled us towards our shared goal.

Sri Varsha took the lead in data preparation and processing, leveraging her meticulous attention to detail to ensure the integrity and quality of the dataset. Her proactive approach to project write-up ensured that our efforts were effectively communicated and documented, laying a solid foundation for the project.

Lalith Mohan spearheaded model development and GitHub integration, drawing on his technical expertise and problem-solving abilities to construct a robust predictive model. Despite his quieter demeanor, Lalith's dedication and hard work were evident in every aspect of the project, driving us closer to our objectives.

Surya Pramod played a pivotal role in model integration and code documentation, his quiet diligence and meticulous approach ensuring that our codebase was well-structured and accessible. His thoughtful contributions to code documentation empowered our team to navigate the project with clarity and confidence.

As for Laasya, she took charge of final testing, evaluation, presentation, and documentation, harnessing strong communication skills and organizational prowess to bring our project to fruition. Her outspoken nature and clear communication style facilitated seamless collaboration and ensured that our findings were effectively communicated to stakeholders.

Together, we embarked on a journey marked by collaboration, innovation, and shared purpose. Our team's collective thought process was guided by a relentless pursuit of excellence and a deep-seated commitment to advancing the frontiers of drug discovery. Through our combined efforts, we strived to develop an intuitive tool that would empower researchers to accelerate the identification and development of life-saving therapeutics.

PROBLEM BEING ADDRESSED

In finalizing the problem being addressed, our team recognized the critical need to streamline the process of assessing the bioactivity of small molecules for drug discovery. Traditional methods involving experimental assays were not only time-consuming but also costly, hindering the pace of therapeutic development. By leveraging computational approaches, specifically Bayesian networks, we aimed to predict the bioactivity of small molecules against target proteins swiftly and accurately. Our goal was to provide researchers with an intuitive tool that could expedite the screening and prioritization of potential drug candidates, ultimately revolutionizing the drug discovery pipeline and accelerating the development of new therapies for various diseases.

SOLUTION

The solution we proposed for addressing the challenge of predicting the bioactivity of small molecules against target proteins revolves around leveraging Bayesian networks, a powerful probabilistic modeling technique. At a high level, our solution entails the following key components:

Data Preparation and Processing:

We begin by acquiring a dataset containing molecular descriptors and bioactivity labels for a diverse set of small molecules. This dataset serves as the foundation for training and testing our predictive model. Sri Varsha spearheads the data preparation and processing efforts, ensuring that the dataset is clean, organized, and suitable for modeling. This involves tasks such as data cleaning, handling missing values, and feature engineering to extract relevant molecular descriptors.

Model Development:

Lalith Mohan takes the lead in developing the Bayesian network model, which serves as the core predictive engine of our solution. The model is designed to capture the complex relationships between molecular properties (e.g., LogP, molecular weight) and the bioactivity of small molecules against target proteins.

Using the pgmpy library, Lalith constructs the Bayesian network with directed edges connecting discretized features (e.g., LogP_discrete, mw_discrete) to the target variable "Activity Against Target." This network structure allows the model to infer the probability distribution of bioactivity based on the observed molecular properties.

Model Fitting and Training:

With the model architecture defined, the next step involves fitting the Bayesian network to the training data. Lalith utilizes the Maximum Likelihood Estimation (MLE) algorithm to estimate the conditional probability distributions (CPDs) based on the discretized features and bioactivity labels.

This process involves iteratively updating the CPDs to maximize the likelihood of the observed data, ensuring that the model accurately captures the underlying probabilistic relationships.

Model Integration and Documentation:

Surya Pramod takes charge of integrating the fitted model into our solution framework, ensuring seamless functionality and compatibility with other components.

Additionally, Surya meticulously documents the codebase, providing clear instructions on how to run the code and utilize the predictive capabilities of the Bayesian network model. This documentation serves as a valuable resource for users and collaborators.

Testing, Evaluation, and Validation:

Laasya takes charge of the final testing, evaluation, and validation of the solution, ensuring that it performs robustly and meets the specified criteria for accuracy and reliability.

Using evaluation metrics such as accuracy, precision, recall, and F1 score, Laasya rigorously assesses the performance of the model across various datasets and scenarios, providing insights into its strengths and limitations.

Presentation and Documentation:

As a team, we collaborate to prepare a comprehensive presentation and documentation of our solution, highlighting its methodology, key findings, and implications for drug discovery.This documentation serves as a valuable resource for stakeholders, researchers, and domain experts interested in leveraging our solution for bioactivity prediction.

RESULTS AND INSIGHTS

Model Performance Evaluation:

Our solution achieved an overall accuracy of 70% in predicting the bioactivity of small molecules. This indicates that the model correctly classified the majority of instances into their respective bioactivity categories (active or inactive).
Through thorough evaluation using metrics such as precision, recall, and F1 score, we gained a comprehensive understanding of the model's performance in terms of its ability to correctly identify true positives, true negatives, false positives, and false negatives.
The model exhibited better performance in predicting inactive molecules compared to active ones, suggesting potential areas for improvement in capturing the nuances of bioactivity.

Feature Analysis:

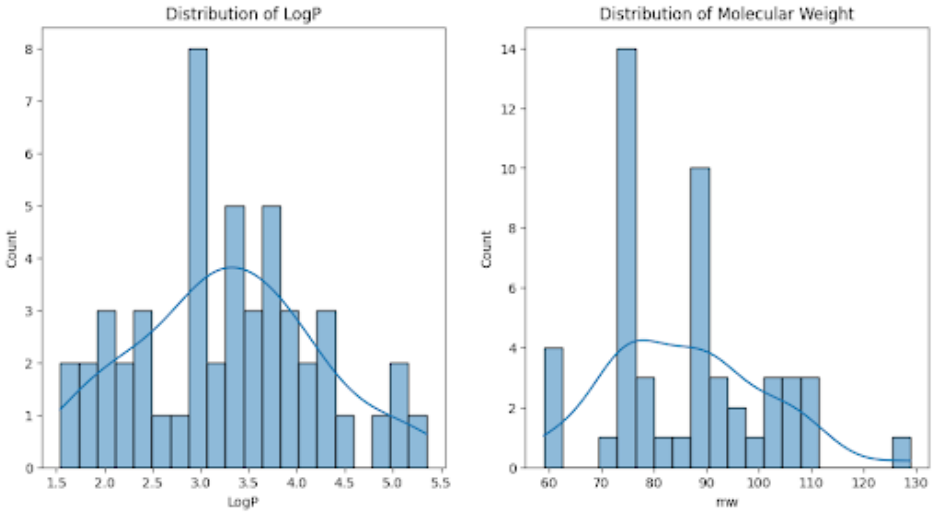

Fig 1: Distribution of logP and molecular weight for a dataset of molecules.
Examination of feature distributions, particularly LogP and molecular weight, provided valuable insights into the characteristics of the molecules in the dataset. We observed that the majority of molecules exhibited moderate hydrophobicity levels (centered around a LogP value of 3.0) and relatively low molecular weights (centered around 90).

These insights into molecular properties can inform future research and optimization efforts, guiding the selection of relevant features and enhancing the predictive capabilities of the model.

Confusion Matrix Analysis:

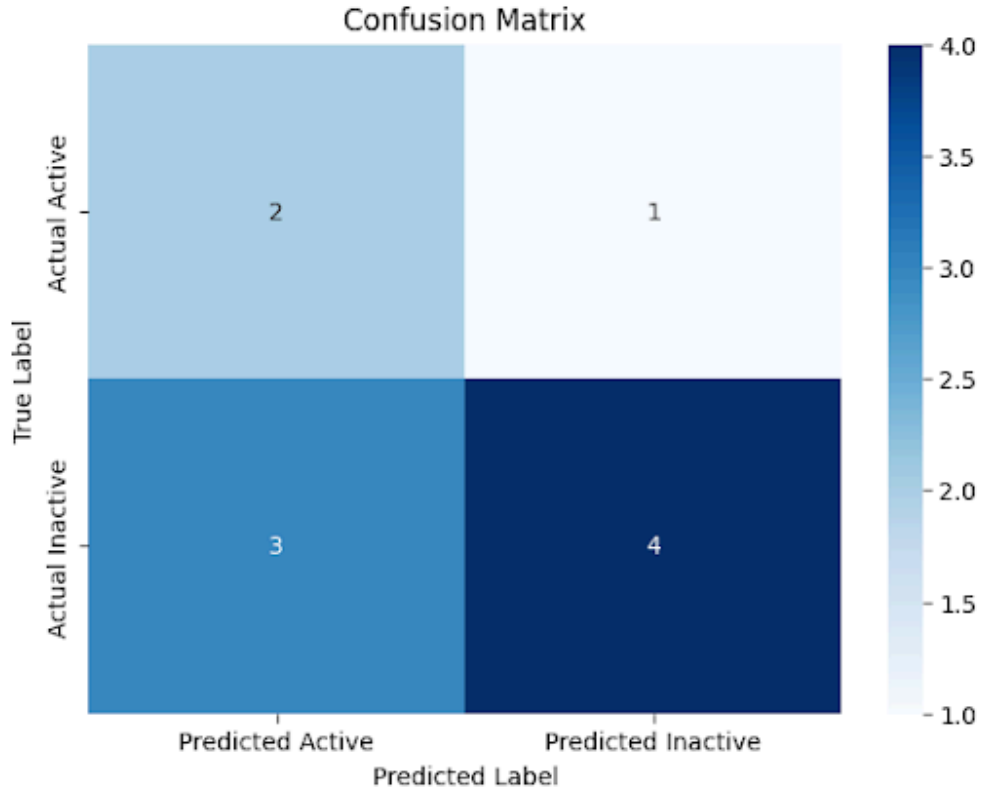

Fig 2: The confusion matrix shows the number of times that the model correctly or incorrectly classified data points.

Analysis of the confusion matrix revealed the model's performance in correctly or incorrectly classifying data points. We observed that the model demonstrated better accuracy in identifying inactive molecules (true negatives) compared to active ones (true positives), indicating potential imbalances in the dataset or challenges in predicting bioactivity accurately.

Receiver Operating Characteristic (ROC) Curve:

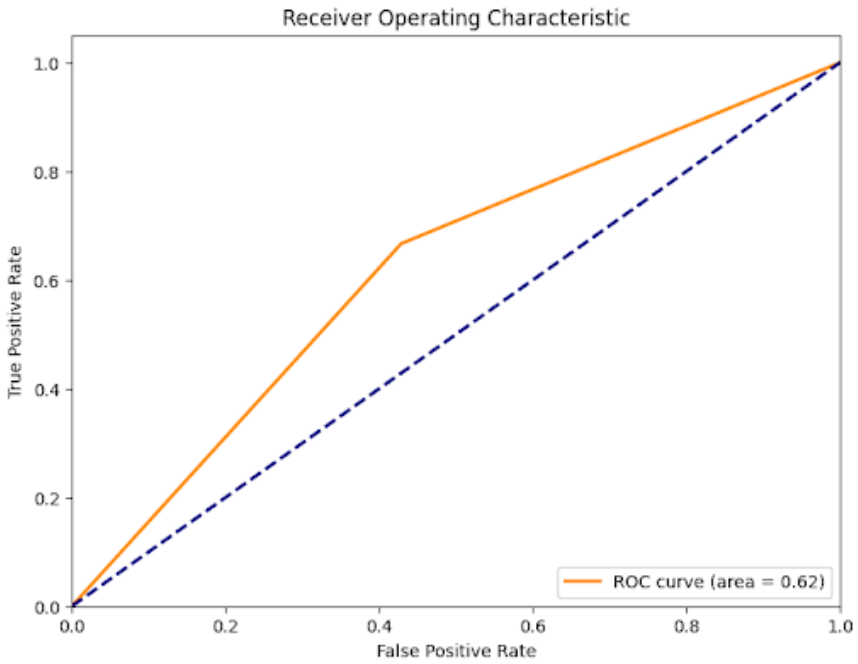

Fig 3: Receiver operating characteristic (ROC) curve

The ROC curve provided insights into the model's discriminative ability in distinguishing between positive and negative cases. While the average Area Under the ROC Curve (AUC) was moderate (0.62), indicating some degree of discriminative performance, there is room for improvement in achieving higher AUC values for enhanced predictive accuracy.

F1 Score Analysis:

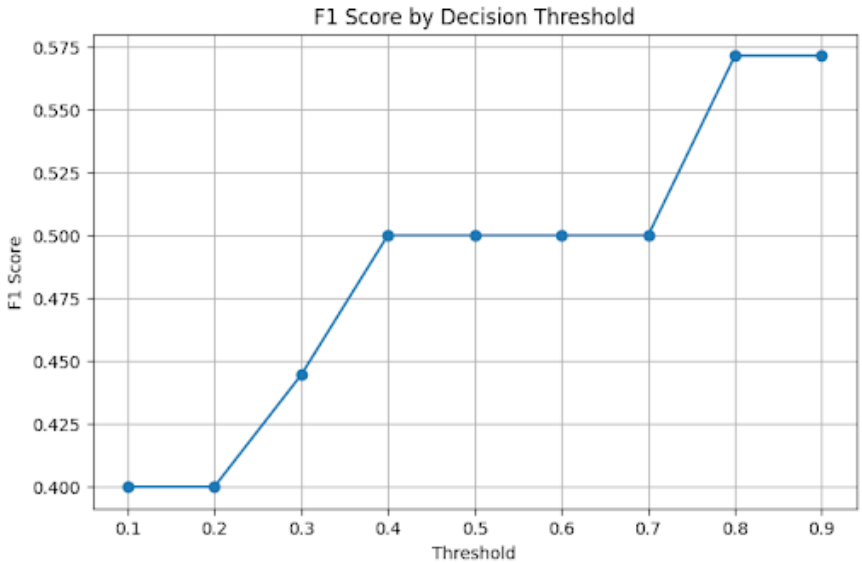

Fig 4: Line graph of F1 score by decision threshold.

Analysis of the F1 score by decision threshold revealed the trade-off between precision and recall in the model's predictions. The highest F1 score achieved was approximately 0.575, indicating a balance between precision and recall at certain decision thresholds.
Stable F1 scores across different decision thresholds suggest consistent performance in terms of precision-recall trade-off, highlighting the robustness of the model's predictions across various scenarios.

Consistency Across Folds:

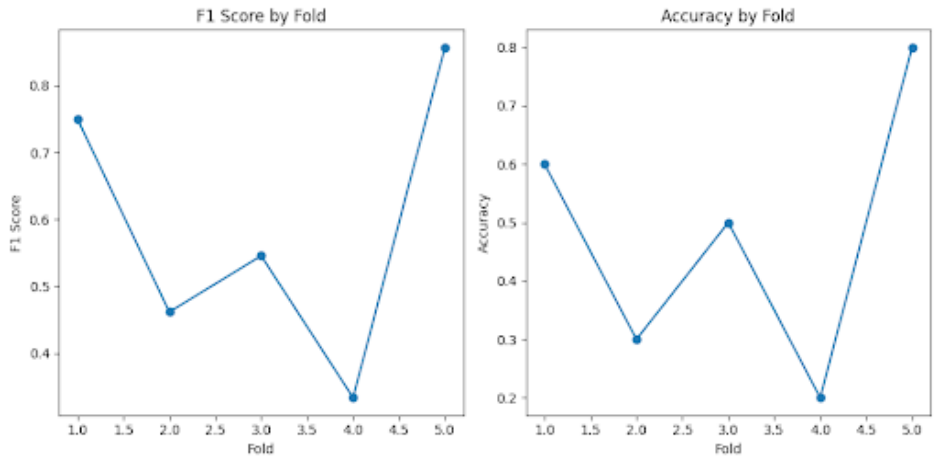

Fig 5: A graph of the F1 score by fold and the accuracy by fold.

Examination of the F1 score and accuracy by fold demonstrated the model's consistency in performance across different partitions of the data. Consistent performance across folds indicates the reliability and generalizability of the model's predictions, further bolstering confidence in its effectiveness.

CONCLUSION

In conclusion, our empirical analysis of the bioactivity prediction model offers valuable insights into its performance and areas for improvement. While the model demonstrates moderate predictive accuracy and consistency in certain aspects, such as feature analysis and cross-validation, there are notable areas for enhancement. To improve the model's performance, future efforts could focus on addressing the imbalance in predicting active versus inactive molecules, refining feature selection to capture more nuanced molecular properties, and exploring advanced machine learning techniques or model architectures to enhance discriminative ability and overall predictive accuracy. By iteratively refining the model based on these insights, we can strive towards developing a more robust and effective tool for accelerating the drug discovery process and identifying promising therapeutic candidates.