Blue text on a black background

Description automatically generated

Data Mining

Course Project

2023/2024

Group 10

André Borgas, 60436 - 12 Hours

Bernardo Godinho, 51497 - 12 Hours

Pedro Sousa, 54729 - 12 Hours

### Report on Protein-Chemical Activity Prediction Using Matrix Factorization and Molecular Fingerprints

**Introduction:**

The task is to predict molecular activity using machine learning techniques. Molecular activity prediction plays a crucial role in drug discovery and development processes by identifying potential candidates for further experimental testing. The report leverages molecular fingerprints, which are representations of molecular structures, and applies various machine learning algorithms to predict the activity of molecules.

The code for this report can be seen in <https://github.com/bgbigberna/DP_2324/blob/main/Project_RandomForest/PD_2324_G10.ipynb>

**Data Loading and Preprocessing:**

We begin with loading the necessary libraries such as pandas and pickle. Then read two CSV files containing training and test data for molecular activity prediction. Additionally, we load a pickled dictionary containing molecular fingerprints. In the image you can observe the first rows of the files, namely the Activity Train Dataset, Activity Test Blanked Dataset and Molecular Fingerprint Dataset.

Uma imagem com texto, captura de ecrã

Descrição gerada automaticamenteThe training and test datasets are then merged with the molecular fingerprints based on the molecule IDs. Leading spaces in molecule IDs are also removed to ensure consistency.

Uma imagem com texto, captura de ecrã, Tipo de letra

Descrição gerada automaticamente

**Feature Engineering:**

The molecular fingerprints are processed to ensure a fixed length using a function, and then get converted into 2D.

By converting the molecular fingerprints into a 2D array format, the data is structured in a standardized manner, which is essential for compatibility with many machine learning algorithms. Most machine learning algorithms in libraries like scikit-learn expect input data to be in the form of a 2D array. (This was an issue faced by the early code that was solved by the use of ChatGPT)

**Modeling:**

It is then applied a K-means clustering to identify patterns in the molecular fingerprints. Additionally, it is used Singular Value Decomposition (SVD) for dimensionality reduction. It is also calculated the cosine similarity between molecular fingerprints.

Uma imagem com texto, Tipo de letra, captura de ecrã, file

Descrição gerada automaticamente

Calculating cosine similarity allows us to measure the similarity between the molecular fingerprints of molecules in the training and test datasets. This information plays a significant role in analyzing the similarity between molecular structures in different datasets, which can inform subsequent modeling decisions and enhance the predictive performance of the models. Random Forest, Linear Regression, Decision Tree, Gradient Boosting are the machine learning algorithms used, Linear Regression and Decision Tree were addressed in class and the new ones are suggestions from ChatGPT.

Uma imagem com texto, Tipo de letra, captura de ecrã

Descrição gerada automaticamente

Random Forest Regressor, the machine learning algorithm used, is trained and evaluated for activity prediction.

**Evaluation and Prediction:**

Finally, the trained model is used to predict the activity of molecules in the test dataset. The predictions are saved to a CSV file for further analysis. Below it is possible to visualize a sample of that file generated (the file submited will only have the first 3 columns of the image).

Uma imagem com texto, captura de ecrã, preto e branco

Descrição gerada automaticamente

**Conclusion:**

In conclusion, the project aims to address task of predicting molecular activity by leveraging machine learning techniques and molecular fingerprints.

Through the utilization of advanced algorithms such as K-means clustering, Singular Value Decomposition (SVD), and machine learning regressors, the report demonstrates an approach towards modeling and predicting molecular activities. By clustering molecular fingerprints, reducing dimensionality, and training predictive models, the project strives to uncover patterns and relationships within the molecular data, thereby enhancing the efficiency and accuracy of activity predictions.

The calculated cosine similarity between molecular fingerprints further enriches the analysis by providing insights into the structural similarity between molecules in the training and test datasets. This information can aid in identifying structurally similar molecules and extrapolating activity predictions based on molecular similarity.