Kegg Metabolic Reaction Network

# ABSTRACT

The objective of this project is to develop a regression model for predicting the the neighborhood connection based on various network features. The data used for this analysis is sourced from the KEGG database and is represented in the form of a reaction network.

The project begins by importing the necessary libraries and loading the data into a pandas DataFrame. The 'Neighborhood Connectivity' column is identified as the target variable, and the remaining columns are selected as independent variables.

To evaluate the performance of the regression model, the dataset is split into training and testing sets. The scikit-learn library is utilized to perform this data splitting task. The training set is used to train the regression model, and the testing set is used to assess the model's predictive capabilities.

Although the provided code snippet focuses on data preprocessing and data splitting, it serves as an initial step towards building a regression model for predicting metabolic pathway density. Further steps involving the selection of an appropriate regression algorithm, model training, evaluation, and interpretation of the results are not included in the provided code snippet.

The successful development of a regression model for predicting metabolic pathway density can provide valuable insights into the organization and characteristics of these pathways.

# KEYWORDS

1. **Regression modeling**: A statistical approach used to establish a relationship between dependent and independent variables. In this project, regression modeling is employed to predict the density of metabolic pathways based on various network features.
2. **Network analysis**: Involves the study of complex systems represented as networks or graphs. In this project, network analysis is used to model metabolic pathways as either reaction networks or relation networks and extract meaningful insights from the data.
3. **KEGG database**: Short for Kyoto Encyclopedia of Genes and Genomes, KEGG is a comprehensive bioinformatics database that provides information on genes, pathways, diseases, and other biological entities. The project utilizes data from the KEGG database to construct and analyze metabolic pathway networks.
4. **NeighborhoodConnectivity prediction**: The aim of this project is to develop a regression model that can predict the neighborhood connection.it provides information about how well connected the nodes are to their immediate neighbor.
5. **Network modeling**: Involves the representation of metabolic pathways as networks or graphs, where compounds and genes are nodes, and reactions or relations are edges. Network modeling allows for the exploration and analysis of the interconnectedness and relationships within metabolic pathways.

# IMPORTANT LINKS

**Git-Hub :** https://github.com/AmulPandey/Machine\_Learning\_Project

**Website : https://sites.google.com/view/machine-learning-model/home**

# INTRODUCTION

The above project aims to develop a regression model to predict the density of metabolic pathways using network-based analysis and machine learning techniques. The project utilizes data from the KEGG database, which provides information about genes, compounds, and biological pathways.

The target variable in this project is the **NeighborhoodConnectivity**, which represents the information about how well connected the nodes are to their immediate neighbor .

To achieve the goal of predicting pathway density, the project employs network modeling techniques. Two types of networks, reaction networks and relation networks, are constructed. In the reaction network, compounds are treated as nodes, and genes are represented as edges. In the relation network, compounds and products are considered as edges, while enzymes and genes are represented as nodes.

The machine learning technique used in this project is regression modeling. Regression models are trained using the network features extracted from the constructed networks. These features include various characteristics such as average number of neighbors, clustering coefficient, betweenness centrality, and others. The regression models learn the relationships between these features and the target variable .

Data preprocessing steps are performed, including handling missing values using imputation techniques. The dataset is split into training and testing sets to evaluate the performance of the regression models. Performance metrics such as mean squared error or R-squared can be used to assess the accuracy of the predictions.

The ultimate objective of this project is to develop a regression model that can effectively predict the density of metabolic pathways. By understanding the factors that contribute to pathway density, researchers can gain insights into the organization and behavior of metabolic networks, which has implications in fields such as systems biology, drug discovery, and metabolic engineering.