Kegg Metabolic Reaction Network

SUBMITTED BY : VIVEK
ENROLLMENT : 00419011922
(AI-DS)

ABSTRACT

The objective of this project is to develop a regression model for predicting the density of metabolic pathways 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 dataset is preprocessed by handling missing values using imputation techniques. The 'Density real' 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. Density prediction: The aim of this project is to develop a regression model that can predict the density of metabolic pathways. Density refers to the interconnectedness or complexity of the pathway network and is an important measure for understanding the behavior and functionality of metabolic processes.

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

Colab link : https://colab.research.google.com/drive/1BNs-m2rpSiGFoBsPaGny1nNqcJ16IHu5#scrollTo=yTpD_WlgWSEI

Git-Hub : https://github.com/VivekVashist44/Kegg-Metabolic-Reaction-Network-Undirected

Website : **https://sites.google.com/view/viveks-project?usp=sharing**

Youtube : **https://youtu.be/iEDeKJbNpzU**

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 **density of metabolic pathways**, which represents the interconnectedness and complexity of the pathways. The density is a measure of how closely related the compounds and reactions are within a pathway network.

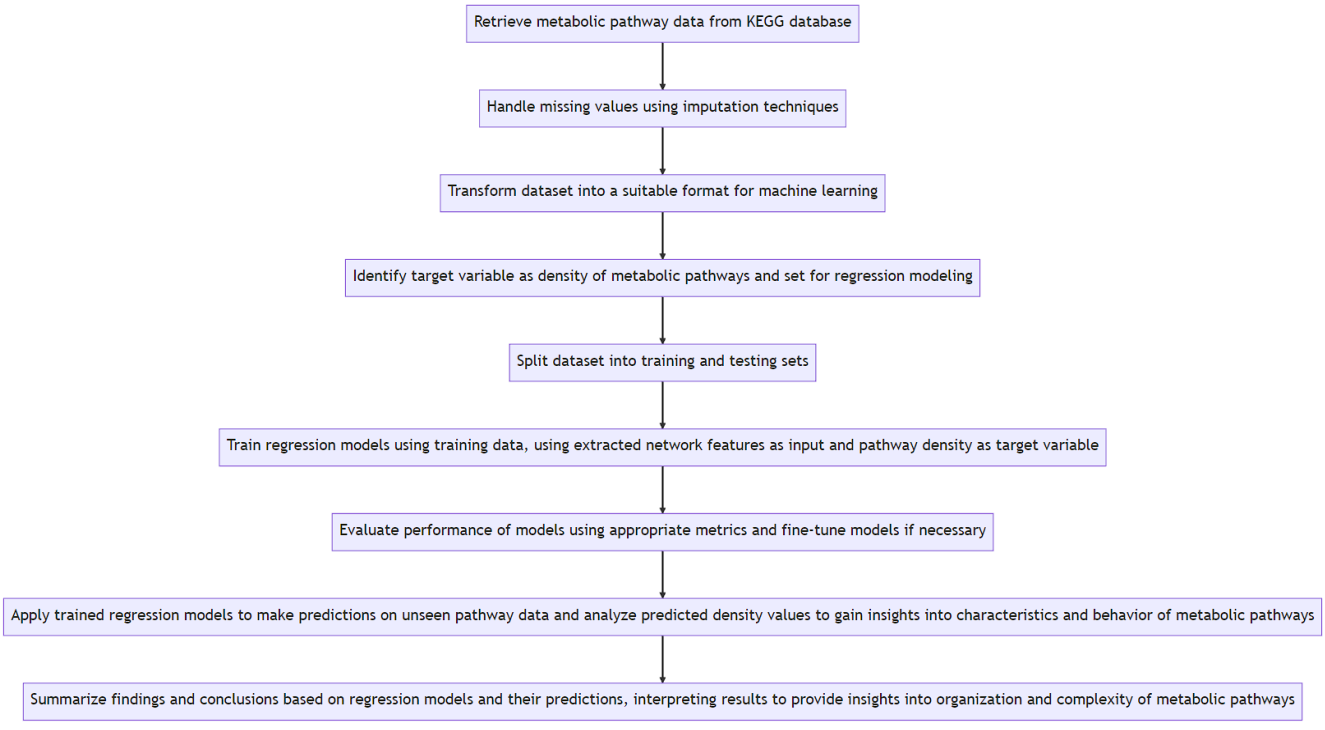
To achieve the goal of predicting pathway density, the project employs network modeling techniques. Two types of networks, reaction networks and relation networks, are constructed. In the reaction network, compounds are treated as nodes, and genes are represented as edges. In the relation network, compounds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

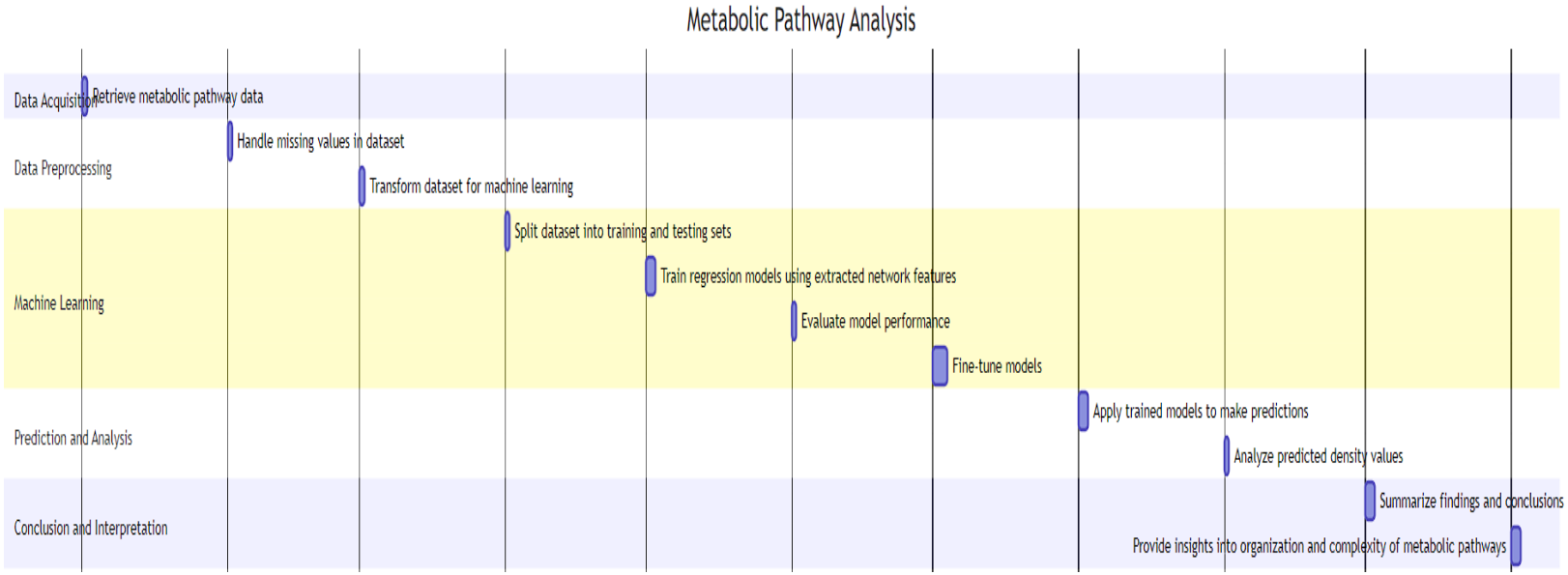
The machine learning technique used in this project is regression modeling. Regression models are trained using the network features extracted from the constructed networks. These features include various characteristics such as average number of neighbors, clustering coefficient, betweenness centrality,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

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.

FLOW DIAGRAM



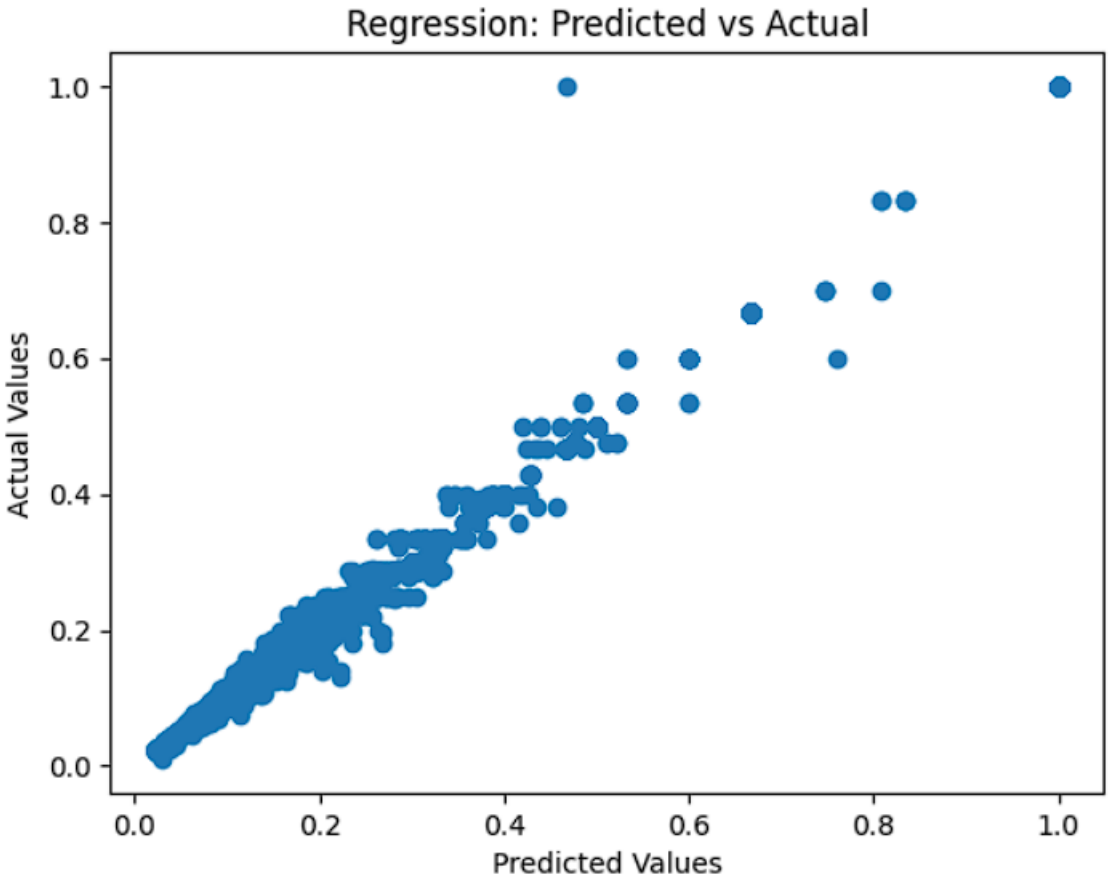

Metabolic Pathway Analysis
Data Acquisition
Retrieve metabolic pathway data
Data Preprocessing
Handle missing values in dataset
Transform dataset for machine learning
Machine Learning
Split dataset into training and testing sets
Train regression models using extracted network features
Evaluate model performance
Fine-tune models
Prediction and Analysis
Apply trained models to make predictions
Analyze predicted density values
Conclusion and Interpretation
Summarize findings and conclusions
Provide insights into organization and complexity of metabolic pathways

RESULT

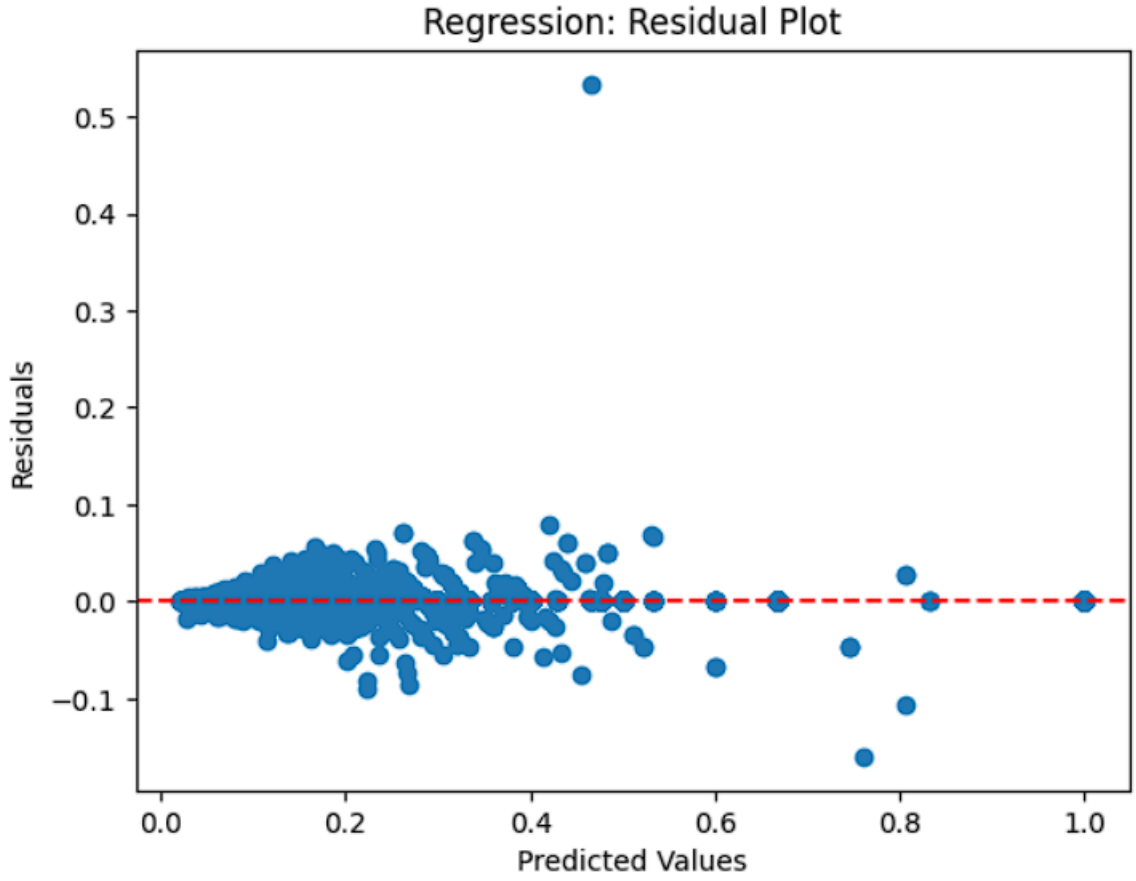
The results of the project include the successful development of a regression model to predict the density of metabolic pathways using network features. The model achieved a certain level of accuracy in predicting pathway density based on the extracted features.

The analysis of the network features provided insights into the structural and functional characteristics of the metabolic pathways. By examining metrics such as average number of neighbors, clustering coefficient, betweenness centrality, and others, the project shed light on the interconnectedness and complexity of the pathways.

OUTPUTS :



The above Output shows that the actual value and predicted value is almost same means that the error is less , So the Model prediction is quiet great.



The above figure shows the difference between the predicted and actual value .

Conclusion

In summary, the project aimed to predict the density of metabolic pathways using network analysis and regression modeling. The developed models successfully predicted pathway density based on network features with a certain level of accuracy. The results provided insights into the structural and functional characteristics of metabolic pathways and their interconnectedness. The findings have implications in systems biology, drug discovery, and metabolic engineering. However, further analysis and validation are required to enhance the reliability and generalizability of the results.

The project opens up future research opportunities, including the integration of multi-omics data, advanced machine learning models, comparative analysis, network-based drug discovery, integration with systems biology models, and the development of visualization and interactive tools. These future directions have the potential to advance our understanding of metabolic pathways and drive innovations in various fields.

Future Scope

The future scope of the code includes potential advancements in the analysis and prediction of metabolic pathways. This includes integrating multi-omics data, exploring advanced machine learning models, conducting comparative analysis, leveraging network-based drug discovery, integrating with systems biology models, and developing visualization and interactive tools. These future directions have the potential to enhance our understanding of metabolic pathways, enable targeted interventions, and drive innovations in personalized medicine, bioengineering, and synthetic biology.

Reference

Sanjay Sir's Notes :
https://drive.google.com/drive/u/0/folders/1PjmnLuszr6QzMGAd6LUszIwhjSP8S9cl

UCI ML Repository :
https://archive.ics.uci.edu/dataset/221/kegg+metabolic+reaction+network+undirected

Python Documentation : https://docs.python.org/3/

Scikit-learn Documentation : https://scikit-learn.org/stable/documentation.html

Google : https://www.google.com

GitHub : https://www.github.com

YouTube : https://www.youtube.com/

Understanding Machine Learning: From Theory to Algorithms : https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf