**Drexel University**

**College of Computing and Informatics**

**INFO 442 – Data Science Project**

**Weekly Progress Report**

**Due Date: Sunday of Week 1-8**

**Each team is expected to submit a weekly progress report during week 1-8. The goal of these reports is not just to track your progress, but to facilitate learning. Be honest about your challenges and don't be afraid to ask questions**.

**Team Name**:

**Team Members**:

**Week Number**:

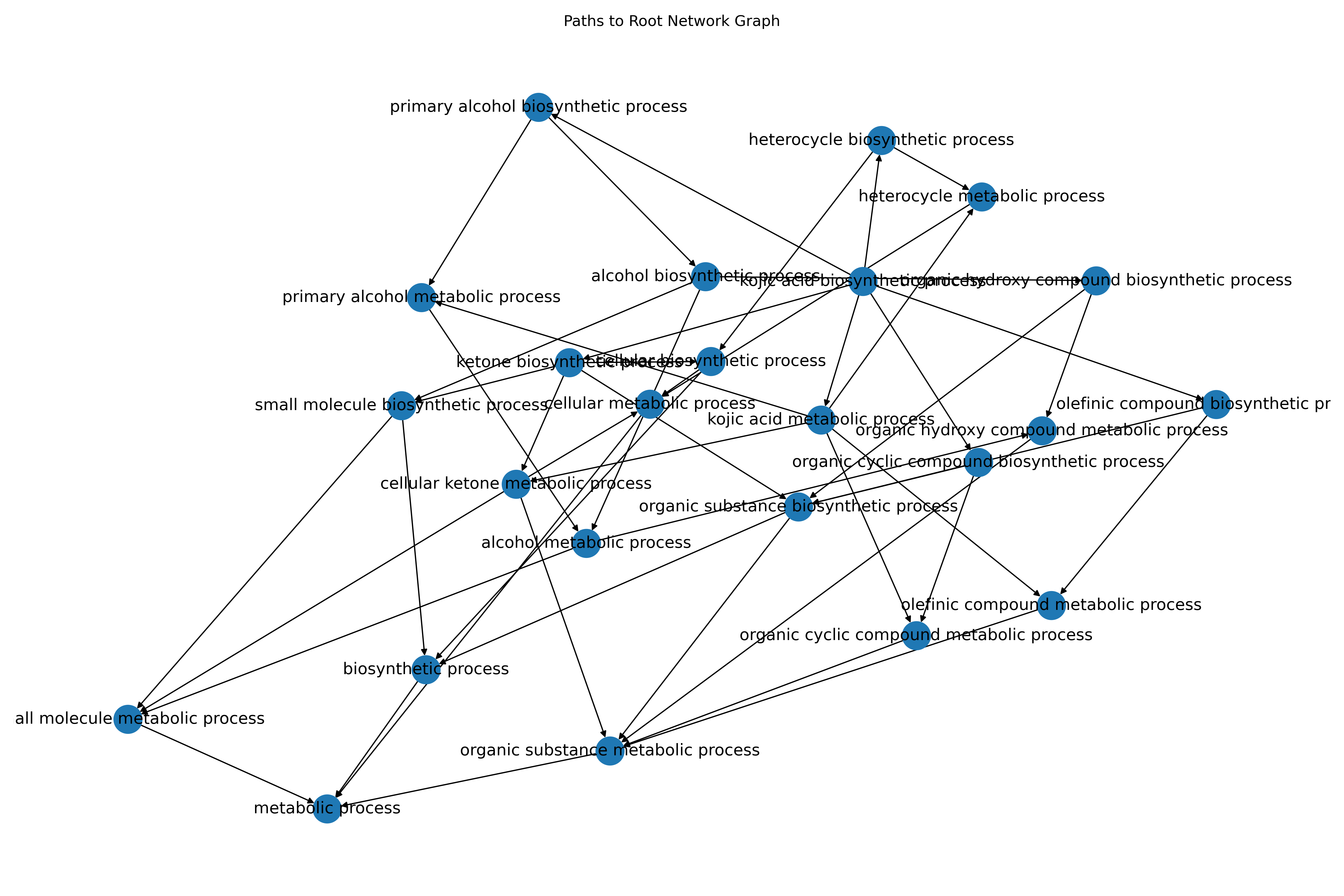
1. **Accomplishments**: Detail what your team has achieved this week. This could include any new models you've tried, any new techniques you've learned, or any significant steps forward in your understanding of the problem.

* This week, team made progress in terms of understanding protein sequences and their structure.
* We dived into details of protein ontology and explored ‘obonet’ package in python whose main purpose is to analyze and interpret network graphs.
* We familiarized ourselves with the concept of super terms and sub terms in network graphs, this helped us to grasp the hierarchical relationships.
* We learned how to create Word Clouds and comprehended the distributions of super terms and sub terms.

1. **Challenges and Solutions**: Describe any challenges or roadblocks you encountered this week. This could be technical (issues with the code or the tools), conceptual (struggling to understand certain aspects of protein function), or related to your modeling approach. Also, detail the solutions you found for these challenges.

The challenging part was understanding protein structures which are composed of dozens or hundreds of amino acids linked sequentially. There is a file called ‘go-basic.obo’ that represents the ontology data. So we decided upon using this file to understand protein structure web network. We selected obonet package to visualize protein structure web network. Creating visualizations and analyzing connections between parent terms, child terms and edges that connect them, helped us to gain better understanding of how protein sequences/functions are interconnected. We also fixed a node as source, super term as target and observed the paths from source to the root. We believe that this will also help us in future to model and feature engineer the available data

1. **Data Understanding**: Discuss any new insights or understanding about the data you've gained this week. Include any relevant data visualization or statistics.



This visualization helped us to understand how protein sequences are linked.

1. **Feature Engineering**: Describe any steps you've taken towards feature engineering. What changes did you make to the input data, and why? What impact have these changes had on your model performance?

During this week, we haven’t undertaken any feature engineering steps.

1. **Modeling**: Detail any new models you've tried this week. Include the type of model, the specific parameters you used, and why you chose this model and these parameters.

This week, we haven’t experimented with any new models, and we have been using a starter DNN TensorFlow model.

1. **Model Evaluation**: Describe how you evaluated your models. What metrics did you use? What are your scores? How do they compare to the starter code score of 0.39? Provide an explanation if your score is higher or lower.
2. **Leaderboard Position**: Mention your current position on the Kaggle leaderboard and how it has changed over the week.
3. **Next Steps**: What are your plans for next week? What strategies, models, or techniques will you try? What are your specific goals?

We have a plan for feature engineering and feature selection for this data. Moreover, we will also look over other available files to uncover any potentially valuable insights. Our primary goal will be to perform feature engineering and assess its impact on model’s performance.