Section 1

Name: Samardeep Singh Gurudatta
Thrust: Microservice track
The microservice thrust employs the analysis of the Luddy student climate survey dataset in the form of a trained Machine Learning model that responds to request for predictions. The dataset consists of responses from several students in various courses at Luddy School to questions that analyze their sense of belonging to and involvement in their department.

Section 2

There are a series of steps that were followed to make the dataset useful for building a machine learning model. The preprocessing stage focused on the following things:

a. Identifying null values and creating measures to handle them
b. Removing undesired data points from the dataset
c. Encoding categorical values to numbers

The target variable for prediction is 'sense of belonging _14'. The question that is being answered using this feature is: If a student is proud to be part of Luddy School. The steps involved in preprocessing the data are as follows:

a. Create a mapping (using dictionary) of survey questions to feature names in the dataset and store it in a JSON file.

b. Removing the first two rows as they did not contain survey responses.

c. Removing ambiguous responses, such as 'White or Caucasian, Black or African American' from Q11(Chose one or more races that you consider yourself to be)

d. Removing the metadata features from the dataset as these features are irrelevant for the prediction process

e. Dealing with null values:
    i. If a row contains null values for the target variable, then the entire row is dropped from the dataset
    ii. If a feature contains more than 100 null values (e.g. nationality), the feature is dropped

iii. If a feature contains less than 100 null values, the null values are imputed with the most common category in the column (as the features are categorical).

f. Encoding categorical features:
   i. If a feature contains ordinal categorical responses, such as 'Strongly Disagree', the feature is encoded using the following mapping:
      a. Response Mapping: {'Strongly Disagree': 1,
         ' Somewhat Disagree': 2,
         'Neither agree nor disagree': 3,
         'Somewhat agree': 4,
         'Strongly agree': 5}
   ii. If a feature contains 'Yes/No' responses, the feature is encoded using the following mapping:
      a. Binary Mapping: {'Yes': 1, 'No': 0}
   iii. If a feature contains nominal categorical responses, such as 'luddy_department', the feature is One Hot Encoded. One Hot encoding is a representation of categorical variables as binary vectors.

g. The target variable, sense of belonging _14, is a feature with multiple classes. It is transformed into a binary variable for classification purposes. All values of the target variable greater than 3 are replaced by 1 which suggest that a student is proud to be a part of Luddy School. All values less than or equal to 3 are replaced with 0 which suggest that a student is not proud to be a part of Luddy School.

h. The processed dataset is stored as a CSV file.

Section 3

The virtual machine allocation on Jetstream was pivotal for development and deployment of the microservice. Below is a list of tasks that were done using the Jetstream VM:

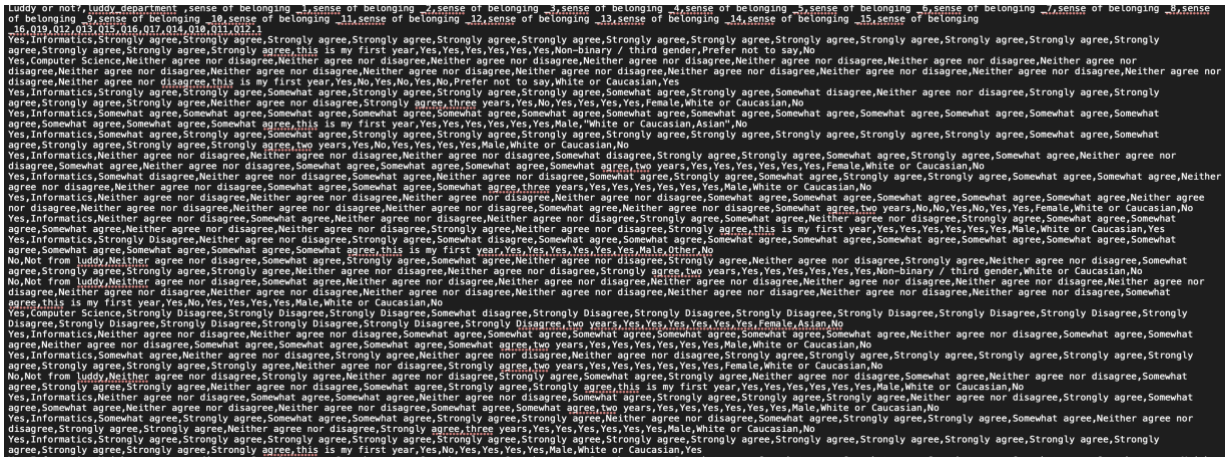
a. Data Preprocessing: All the steps mentioned in the preprocessing phase were performed by creating a Jupyter notebook on the Jetstream VM. Model training and storage: The process of testing different models such as Logistic Regression, Decision Tree, and Random Forest were performed by creating a Jupyter notebook on the VM. Moreover, optimizing the hyper-parameters for the most optimal algorithm (Logistic Regression) using GridSearchCV was completed in the same notebook. Finally, the model is stored as a pickle file on the VM.

b. Endpoint Creation: The implementation of different endpoints using the trained model and Flask framework was also performed on the VM.

c. Dockerization and HTML app hosting: The two most important tasks performed on Jetstream allocation were to create a Docker image of the microservice and deploy the image as a container on the virtual machine. Since, the Jetstream instance provides good uptime, it was used to deploy the application.

Section 4

a. Storage: The storage for the dataset is a flat file system in the format of Comma Separated Values (CSV). CSV format is preferred because it is convenient to create pandas dataframe that are used for analysis from CSV files. Here is a snippet of the CSV file as well as the pandas dataframe created from the CSV file.



	Luddy or not?	luddy_department	sense of belonging _1	sense of belonging _2	sense of belonging _3	sense of belonging _4	sense of belonging _5	sense of belonging _6	sense of belonging _7	sense of belonging _8	...	Q19	Q12	Q13	Q15	Q16	Q17	Q14	Q10	Q11	Q12.1
0	Yes	Informatics	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree	...	this is my first year	Yes	Yes	Yes	Yes	Yes	Yes	Non-binary / third gender	Prefer not to say	No
1	Yes	Computer Science	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	...	this is my first year	Yes	No	Yes	No	Yes	No	Prefer not to say	White or Caucasian	Yes
2	Yes	Informatics	Strongly agree	Strongly agree	Somewhat agree	Strongly agree	Strongly agree	Strongly agree	Somewhat agree	Strongly agree	...	three years	Yes	No	Yes	Yes	Yes	Yes	Female	White or Caucasian	No
3	Yes	Informatics	Somewhat agree	Somewhat agree	Somewhat agree	Somewhat agree	Somewhat agree	Somewhat agree	Somewhat agree	Somewhat agree	...	this is my first year	Yes	Yes	Yes	Yes	Yes	Yes	Male	White or Caucasian,Asian	No
4	Yes	Informatics	Somewhat agree	Strongly agree	Somewhat agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree	...	two years	Yes	No	Yes	Yes	Yes	Yes	Male	White or Caucasian	No

5 rows × 28 columns

The schema of the dataframe is as follows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 221 entries, 0 to 220
Data columns (total 28 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Luddy or not?          221 non-null    object
 1   luddy_department       221 non-null    object
 2   sense of belonging _1  221 non-null    object
 3   sense of belonging _2  221 non-null    object
 4   sense of belonging _3  221 non-null    object
 5   sense of belonging _4  221 non-null    object
 6   sense of belonging _5  221 non-null    object
 7   sense of belonging _6  221 non-null    object
 8   sense of belonging _7  221 non-null    object
 9   sense of belonging _8  221 non-null    object
 10  sense of belonging _9  221 non-null    object
 11  sense of belonging _10 221 non-null    object
 12  sense of belonging _11 221 non-null    object
 13  sense of belonging _12 221 non-null    object
 14  sense of belonging _13 221 non-null    object
 15  sense of belonging _14 221 non-null    object
 16  sense of belonging _15 221 non-null    object
 17  sense of belonging _16 221 non-null    object
 18  Q19                    221 non-null    object
 19  Q12                    221 non-null    object
 20  Q13                    221 non-null    object
 21  Q15                    221 non-null    object
 22  Q16                    221 non-null    object
 23  Q17                    221 non-null    object
 24  Q14                    221 non-null    object
 25  Q10                    221 non-null    object
 26  Q11                    221 non-null    object
 27  Q12.1                  221 non-null    object
dtypes: object(28)
```

b. Microservice: The microservice is a Machine Learning based prediction service to classify if a student is proud of the Luddy School. It is developed using Python programming language and the Flask framework. The platform used for development and deployment is a Jetstream Virtual Machine based on Ubuntu operating system.

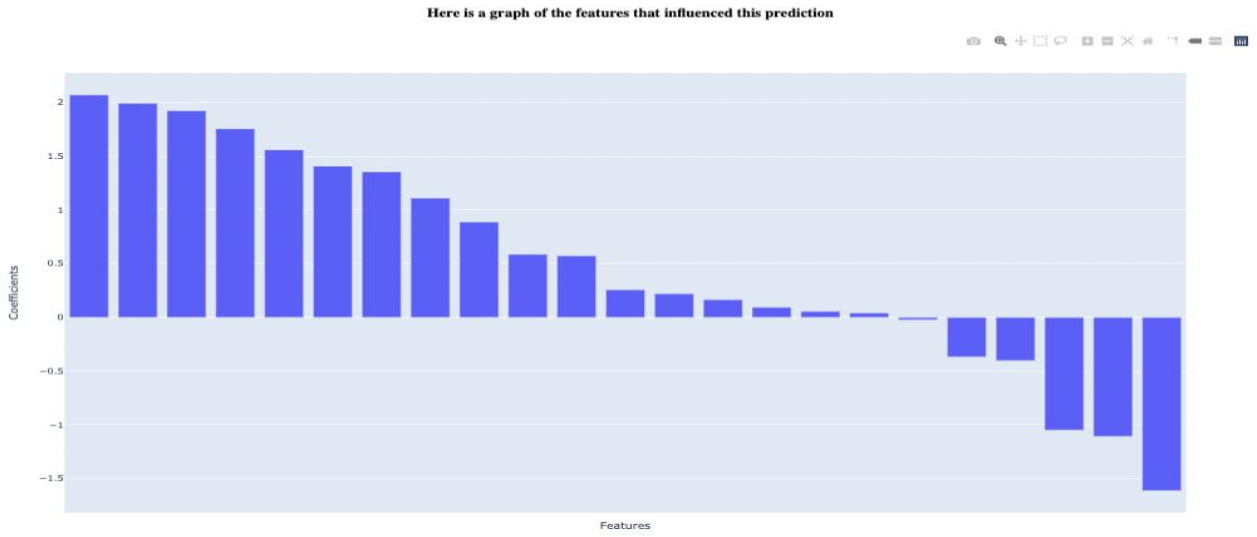
c. API Endpoints: Below is the list of API endpoints that were developed for this project:

   i. Predict:
      1. URI: http://149.165.168.244:5001/predict
      2. Method: GET
      3. Content-type : text/html, charset=utf-8
      4. Response(Server-side rendering): An HTML markup containing questions that are required for predicting the target variable value

ii. Predict

1. URI: http://149.165.168.244:5001/predict
2. Method: POST
3. Payload:
   'q1=Yes&q2=Computer+Science&q3=Strongly+Disagree&q4=Strongly+Disagree&q5=Strongly+Disagree&q6=Strongly+Disagree&q7=Strongly+Disagree&q8=Strongly+Disagree&q9=Strongly+Disagree&q10=Strongly+Disagree&q11=Strongly+Disagree&q12=Strongly+Disagree&q13=Strongly+Disagree&q14=Strongly+Disagree&q15=Strongly+Disagree&q16=Strongly+Disagree&q17=Strongly+Disagree&q18=this+is+my+first+year&q19=Yes&q20=Yes&q21=Yes&q22=Yes&q23=Yes&q24=Yes&q25=Female&q26=White+or+Caucasian&q27=Yes'
4. Content-Type: 'application/x-www-form-urlencoded'
5. Response : The response is a HTML markup which contains the prediction as well as the graph of features that is used to make the prediction. Here is the snapshot of the POST response

Based on the responses, the student is not proud to be a part of Luddy School

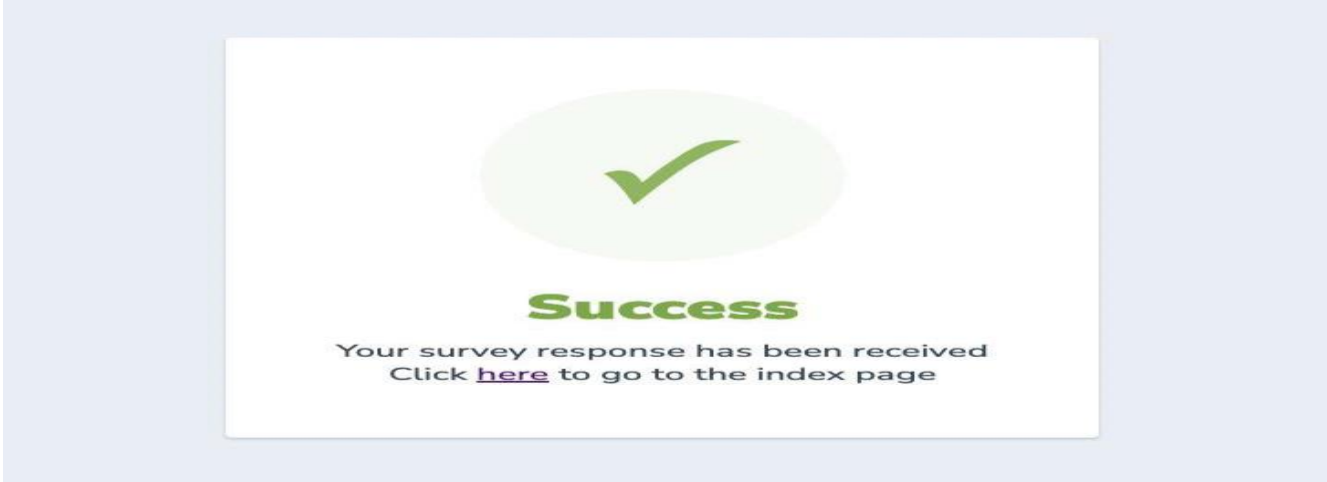


iii. Add Data:

1. URI: http://149.165.168.244:5001/add_data
2. Method: GET
3. Content-type : text/html, charset=utf-8
4. Response(Server-side rendering): An HTML markup containing questions that correspond to the features in the data set.

iv. Add Data :

1. URI: http://149.165.168.244:5001/add_data
2. Method: POST
3. Content-type: 'application/x-www-form-urlencoded'
4. Payload:
   'q1=Yes&q2=Computer+Science&q3=Strongly+Disagree&q4=Strongly+Disagree&q5=Strongly+Disagree&q6=Strongly+Disagree&q7=Strongly+Disagree&q8=Strongly+Disagree&q9=Strongly+Disagree&q10=Strongly+Disagree&q11=Strongly+Disagree&q12=Strongly+Disagree&q13=Strongly+Disagree&q14=Strongly+Disagree&q15=Strongly+Disagree&q16=Strongly+Disagree&q17=Strongly+Disagree&q18=Strongly+Disagree&q19=this+is+my+first+year&q20=Yes&q21=Yes&q22=Yes&q23=Yes&q24=Yes&q25=Yes&q26=Female&q27=White+or+Caucasian&q28=Yes'
5. Response: The response is a HTML markup which tells the user if the data has been added to the flat file storage. Here is a snapshot of the POST response



v. Get Data :
    1. URI: http://149.165.168.244:5001/form_page
    2. Method: GET
    3. Content-type: text/html, charset=utf-8
    4. Response(Server-side rendering): An HTML markup that asks the user about the data that needs to be retrieved

vi. Get Data:
    1. URI: http://149.165.168.244:5001/form_page
    2. Method: POST
    3. Payload: 'q1=1&q2=1&q3=Luddy+or+not%3F&q3=luddy_department+&q3=sense+of+belonging+_1&q3=sense+of+belonging+_2'
    4. Content-type: 'application/x-www-form-urlencoded'
    5. Response: The response is a HTML markup containing a table which contains the data that the user requested. Here is a snapshot of the POST response.

index	luddy_department	sense of belonging _1	sense of belonging _2	sense of belonging _3	sense of belonging _9	sense of belonging _10	sense of belonging _11
0	Informatics	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree
1	Computer Science	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree
2	Informatics	Strongly agree	Strongly agree	Somewhat agree	Somewhat disagree	Neither agree nor disagree	Strongly agree
3	Informatics	Somewhat agree	Somewhat agree	Somewhat agree	Somewhat agree	Somewhat agree	Somewhat agree
4	Informatics	Somewhat agree	Strongly agree	Somewhat agree	Strongly agree	Strongly agree	Strongly agree
5	Informatics	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Somewhat agree	Neither agree nor disagree	Somewhat agree
6	Informatics	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	Somewhat agree	Somewhat agree
7	Informatics	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Somewhat agree	Neither agree nor disagree	Neither agree nor disagree
8	Informatics	Neither agree nor disagree	Somewhat agree	Neither agree nor disagree	Somewhat agree	Somewhat agree	Somewhat agree
9	Informatics	Strongly Disagree	Neither agree nor disagree	Strongly agree	Somewhat agree	Somewhat agree	Somewhat agree
10	Not from luddy	Neither agree nor disagree	Somewhat agree	Strongly agree	Neither agree nor disagree	Somewhat agree	Strongly agree
11	Not from luddy	Neither agree nor disagree	Somewhat agree	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree
12	Computer Science	Strongly Disagree	Strongly Disagree	Strongly Disagree	Strongly Disagree	Strongly Disagree	Strongly Disagree
13	Informatics	Neither agree nor disagree	Neither agree nor disagree	Somewhat agree	Neither agree nor disagree	Somewhat agree	Somewhat agree
14	Informatics	Somewhat agree	Neither agree nor disagree	Strongly agree	Strongly agree	Strongly agree	Strongly agree
15	Not from luddy	Neither agree nor disagree	Strongly agree	Neither agree nor disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
16	Informatics	Neither agree nor disagree	Somewhat agree	Somewhat agree	Neither agree nor disagree	Strongly agree	Somewhat agree
17	Informatics	Somewhat agree	Strongly agree	Somewhat agree	Strongly agree	Strongly agree	Somewhat agree
18	Informatics	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree	Strongly agree
19	Informatics	Neither agree nor disagree	Somewhat agree	Somewhat agree	Somewhat agree	Somewhat agree	Somewhat agree
20	Not from luddy	Somewhat agree	Somewhat agree	Somewhat agree	Neither agree nor disagree	Neither agree nor disagree	Neither agree nor disagree
21	Informatics	Neither agree nor disagree	Strongly agree	Neither agree nor disagree	Somewhat agree	Somewhat disagree	Somewhat agree
22	Informatics	Somewhat disagree	Strongly agree	Somewhat agree	Somewhat agree	Somewhat agree	Strongly agree
23	Informatics	Somewhat disagree	Neither agree nor disagree	Neither agree nor disagree	Somewhat agree	Somewhat agree	Somewhat agree
24	Not from luddy	Neither agree nor disagree	Strongly agree	Somewhat agree	Strongly agree	Strongly agree	Strongly agree

References

https://www.w3schools.com/w3css/4/w3.css
https://www.w3schools.com/w3css/tryit.asp?filename=tryw3css_templates_start_page&stacked=h
https://www.geeksforgeeks.org/build-a-survey-form-using-html-and-css/
https://stackoverflow.com/questions/4406501/change-the-name-of-a-key-in-dictionary
https://towardsdatascience.com/web-visualization-with-plotly-and-flask-3660abf9c946
https://www.kite.com/blog/python/flask-restful-api-tutorial/
https://medium.com/swlh/bringing-your-ml-models-to-life-with-flask-620c21461c8
https://github.com/sbmthakur/ingestor/blob/master/Dockerfile
https://github.com/sbmthakur/ingestor/blob/master/server.py