COVID-19 Symptom Analysis and Prediction System (CSAPS)

CSAPS project aims to develop a system to detect and classify symptoms to predict whether a patient has COVID-19.

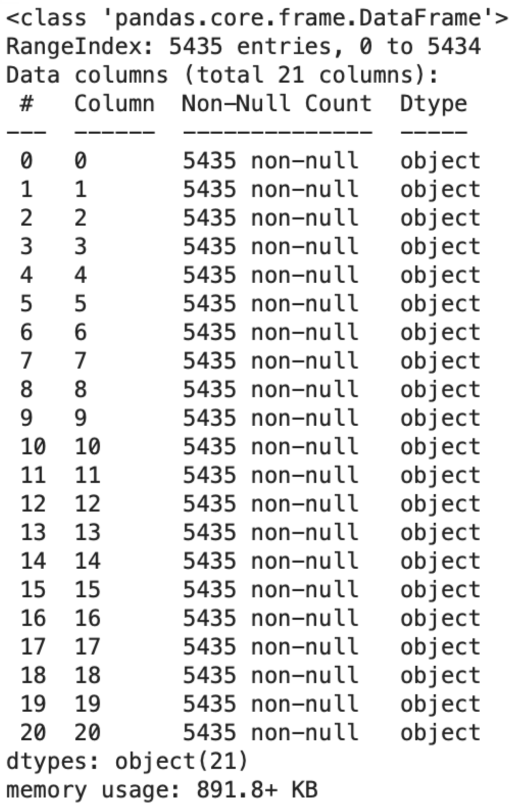
Students: Tao Jin
Yerbol Baizhumanov
Alisher Aliyev

Professor: Dr. Farshid Alizadeh-Shabdiz
TA: Simran Khanna

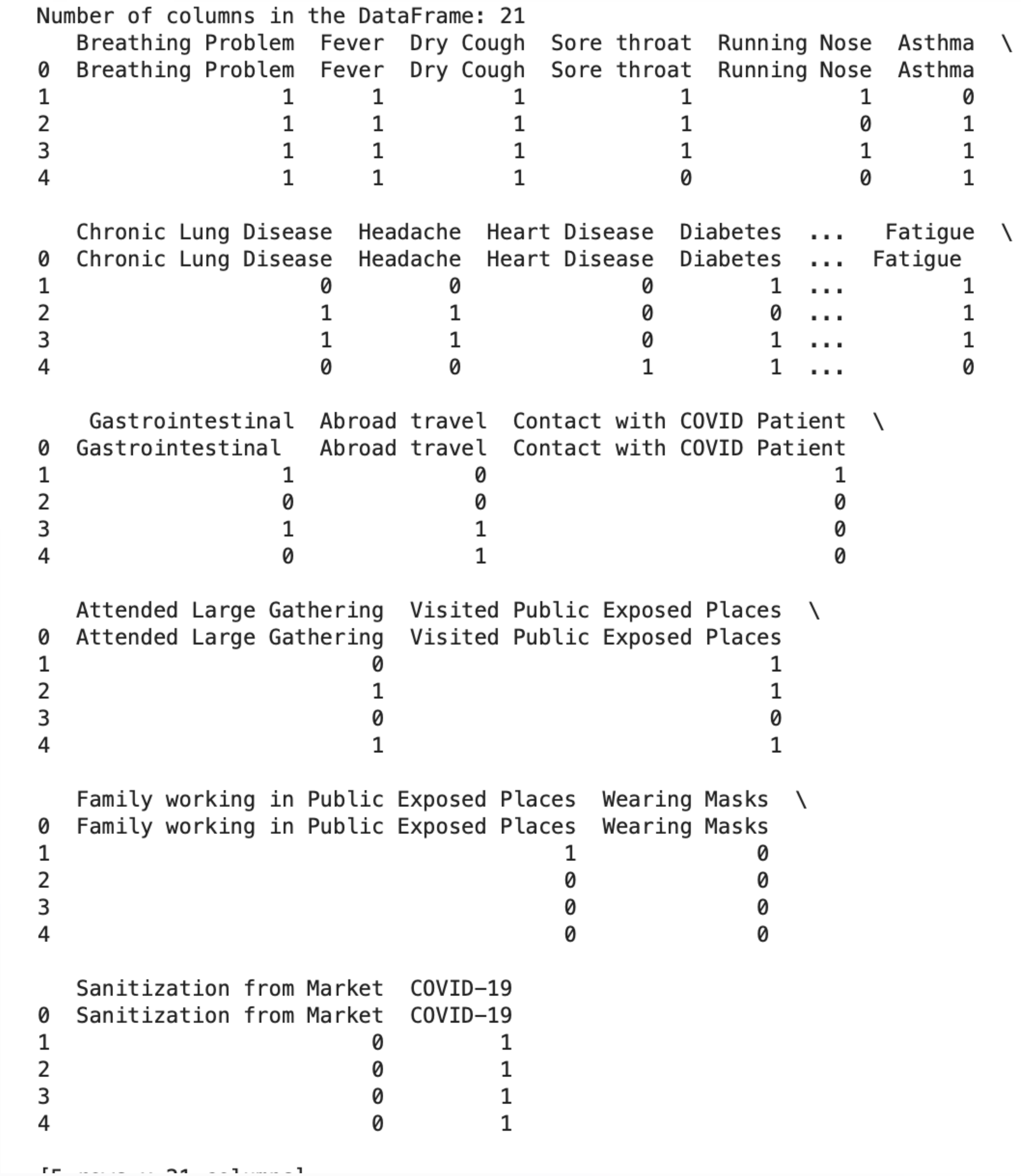
Data Information

Data Preparation

The system reads a CSV dataset containing symptom data and converts categorical columns (Yes/No) into numeric values (1/0). It then extracts symptoms into a feature vector for model training.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5435 entries, 0 to 5434
Data columns (total 21 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   0       5435 non-null   object
 1   1       5435 non-null   object
 2   2       5435 non-null   object
 3   3       5435 non-null   object
 4   4       5435 non-null   object
 5   5       5435 non-null   object
 6   6       5435 non-null   object
 7   7       5435 non-null   object
 8   8       5435 non-null   object
 9   9       5435 non-null   object
 10  10      5435 non-null   object
 11  11      5435 non-null   object
 12  12      5435 non-null   object
 13  13      5435 non-null   object
 14  14      5435 non-null   object
 15  15      5435 non-null   object
 16  16      5435 non-null   object
 17  17      5435 non-null   object
 18  18      5435 non-null   object
 19  19      5435 non-null   object
 20  20      5435 non-null   object
dtypes: object(21)
memory usage: 891.8+ KB
```



```
Number of columns in the DataFrame: 21
   Breathing Problem  Fever  Dry Cough  Sore throat  Running Nose  Asthma  \
0  Breathing Problem  Fever  Dry Cough  Sore throat  Running Nose  Asthma
1                  1      1          1            1             1       0
2                  1      1          1            1             0       1
3                  1      1          1            1             1       1
4                  1      1          1            0             0       1

   Chronic Lung Disease  Headache  Heart Disease  Diabetes  ...   Fatigue  \
0  Chronic Lung Disease  Headache  Heart Disease  Diabetes  ...  Fatigue
1                     0         0              0         1  ...         1
2                     1         1              0         0  ...         1
3                     1         1              0         1  ...         1
4                     0         0              1         1  ...         0

    Gastrointestinal  Abroad travel  Contact with COVID Patient  \
0  Gastrointestinal   Abroad travel  Contact with COVID Patient
1                  1              0                           1
2                  0              0                           0
3                  1              1                           0
4                  0              1                           0

   Attended Large Gathering  Visited Public Exposed Places  \
0  Attended Large Gathering  Visited Public Exposed Places
1                         0                              1
2                         1                              1
3                         0                              0
4                         1                              1

   Family working in Public Exposed Places  Wearing Masks  \
0  Family working in Public Exposed Places  Wearing Masks
1                                        1              0
2                                        0              0
3                                        0              0
4                                        0              0

   Sanitization from Market  COVID-19
0  Sanitization from Market  COVID-19
1                         0         1
2                         0         1
3                         0         1
4                         0         1

[5 rows x 21 columns]
```

Machine Learning Models

1 Random Forest Classifier

This ensemble model combines multiple decision trees to improve accuracy and reduce overfitting.

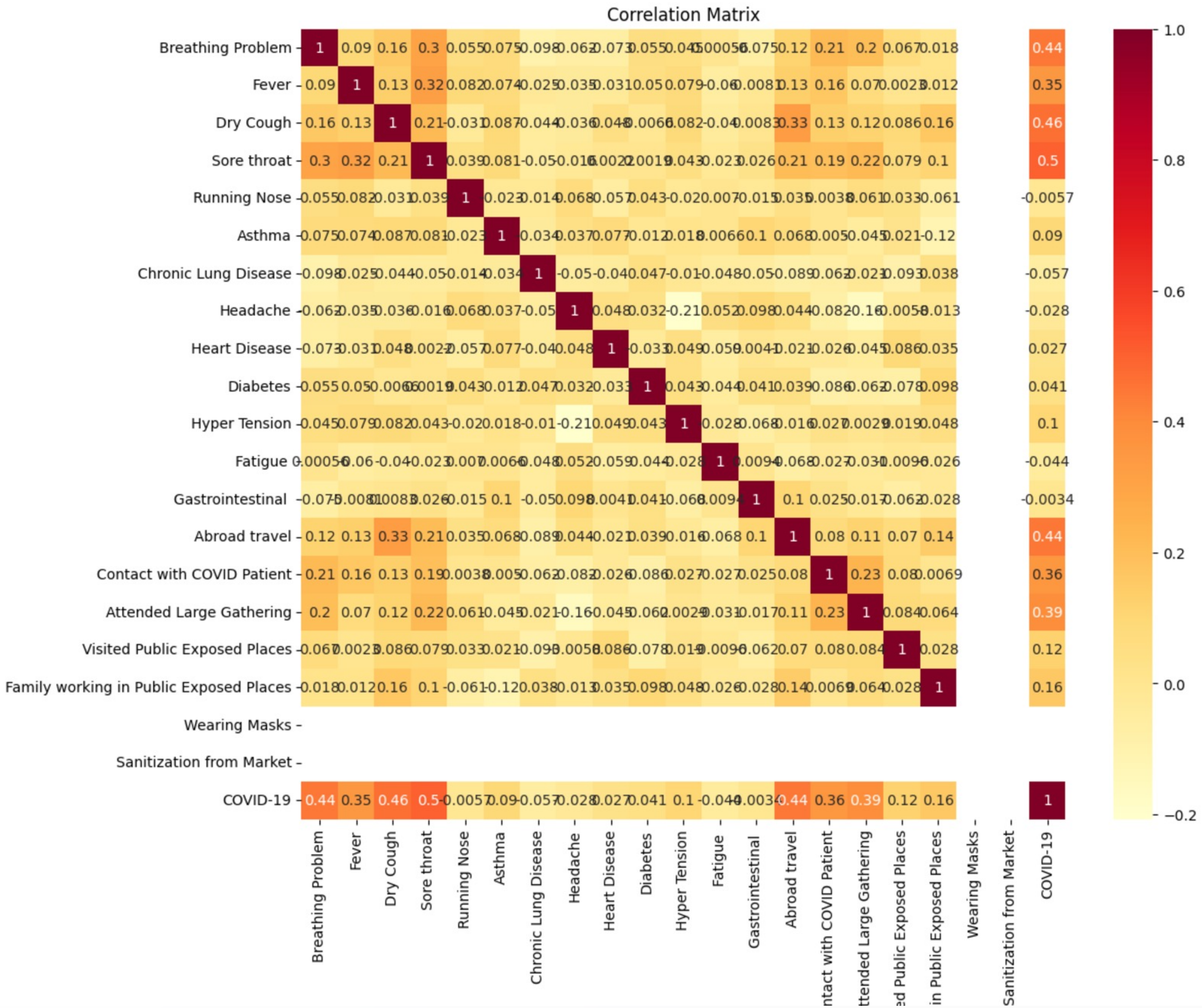
2 Gradient Boosting Classifier

This algorithm sequentially builds decision trees, with each tree correcting errors made by previous trees.

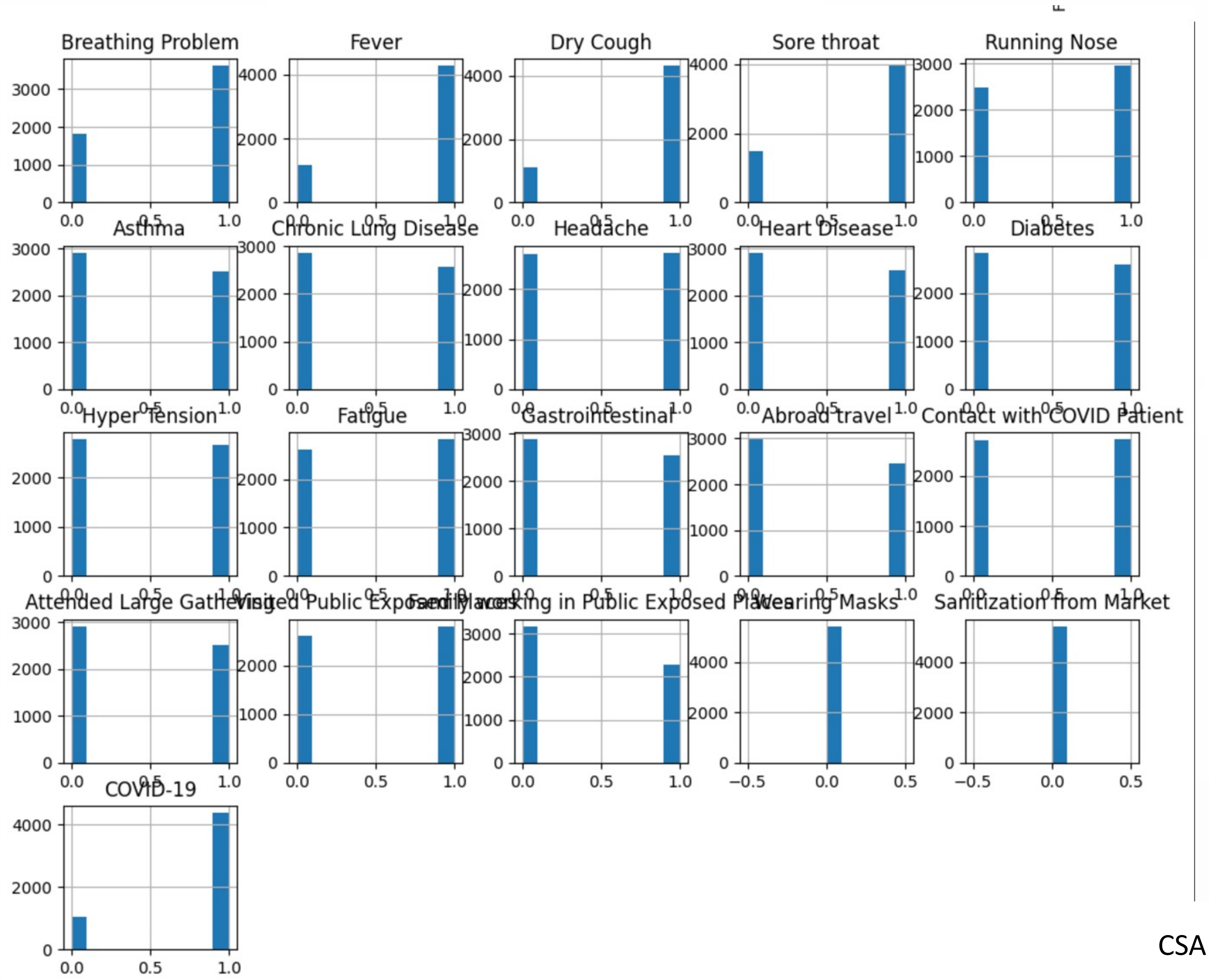
3 Logistic Regression

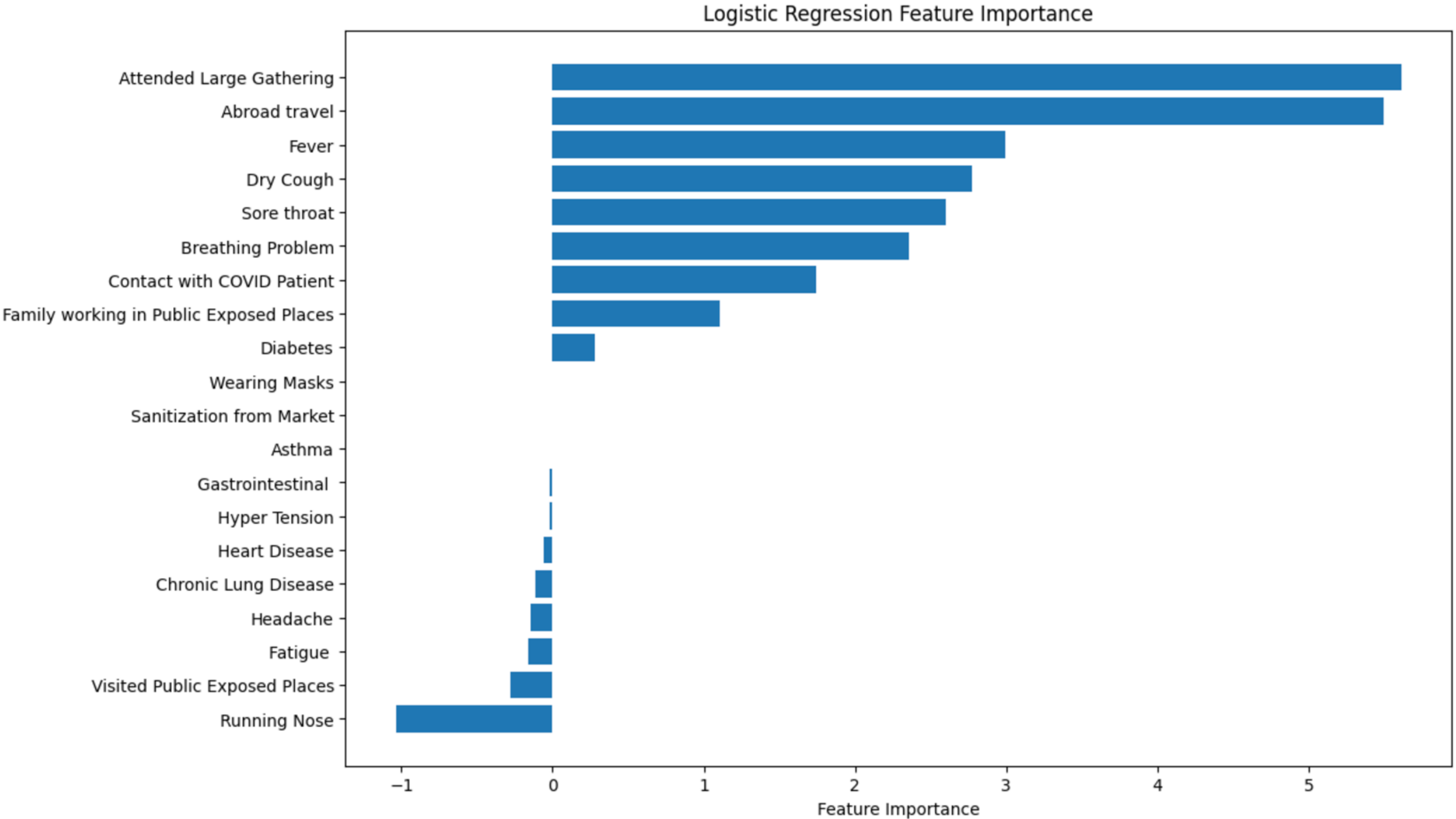
This model predicts the probability of a patient having COVID-19 using a linear combination of features.

Correlation Matrix



Histograms of Features

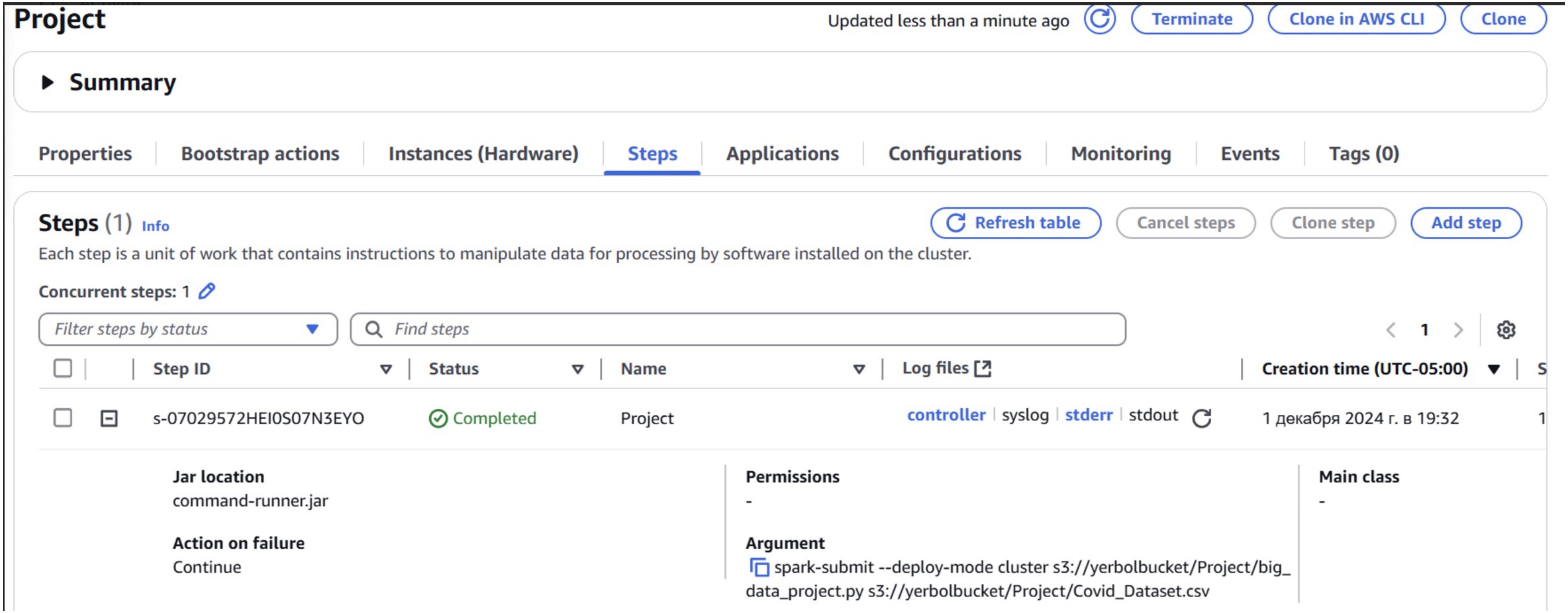



Logistic Regression Feature Importance
Attended Large Gathering
Abroad travel
Fever
Dry Cough
Sore throat
Breathing Problem
Contact with COVID Patient
Family working in Public Exposed Places
Diabetes
Wearing Masks
Sanitization from Market
Asthma
Gastrointestinal
Hyper Tension
Heart Disease
Chronic Lung Disease
Headache
Fatigue
Visited Public Exposed Places
Running Nose
−1
0
1
2
3
4
5
Feature Importance

Scalable Architecture

Apache Spark's distributed computing capabilities enable efficient handling of large volumes of medical data in a scalable and reliable manner.

Cloud Computing



Project

Updated less than a minute ago

Steps (1)

Each step is a unit of work that contains instructions to manipulate data for processing by software installed on the cluster.

Concurrent steps: 1

□		Step ID	Status	Name	Log files	Creation time (UTC-05:00)	S
□	⊟	s-07029572HEI0S07N3EYO	✓ Completed	Project	controller | syslog | stderr | stdout	1 декабря 2024 г. в 19:32	1

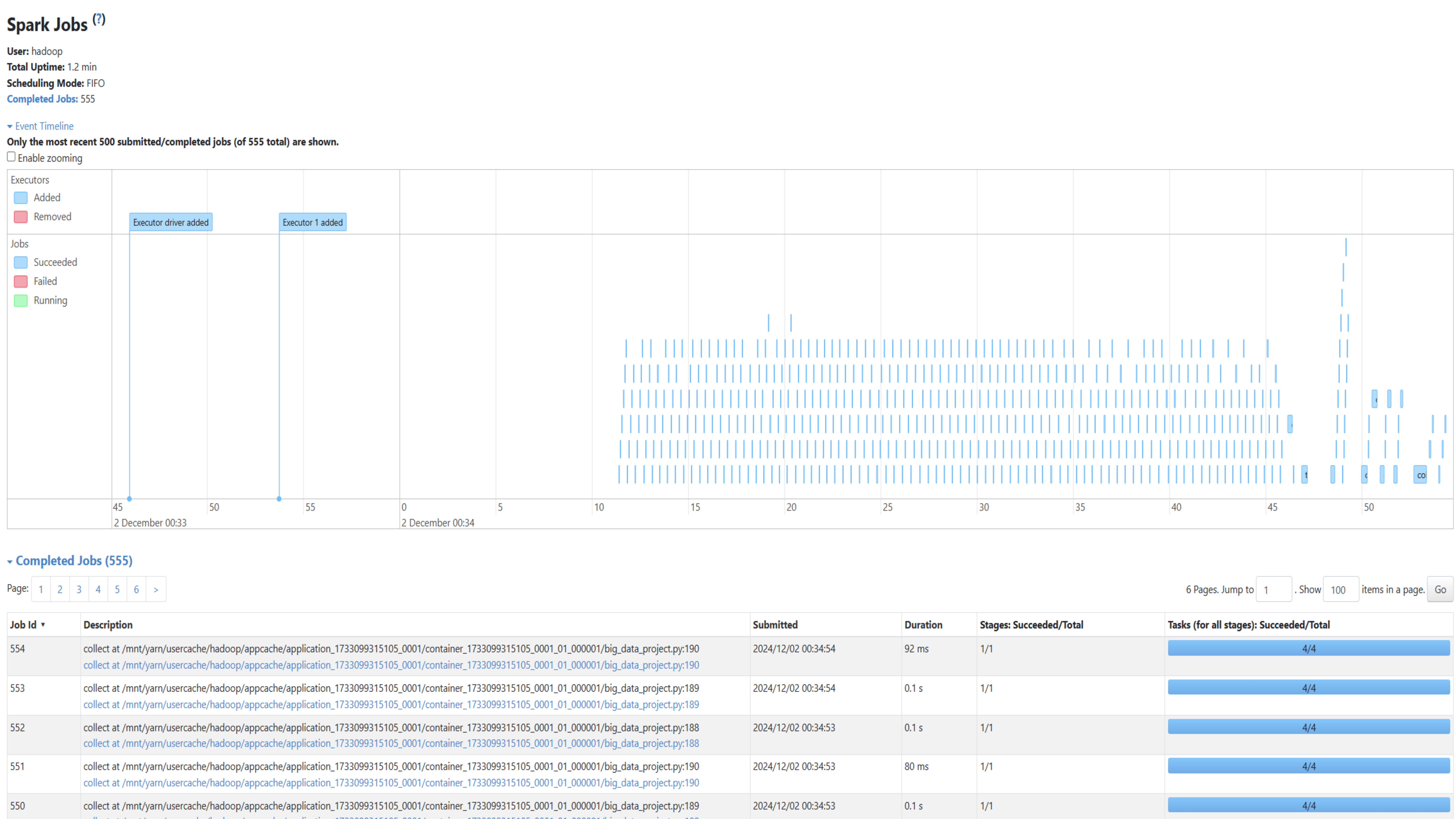
Jar location
command-runner.jar

Action on failure
Continue

Permissions
-

Argument
spark-submit --deploy-mode cluster s3://yerbolbucket/Project/big_data_project.py s3://yerbolbucket/Project/Covid_Dataset.csv

Main class
-



Spark Jobs (?)

User: hadoop
Total Uptime: 1.2 min
Scheduling Mode: FIFO
Completed Jobs: 555

▾ Event Timeline

Only the most recent 500 submitted/completed jobs (of 555 total) are shown.

□ Enable zooming

▾ Completed Jobs (555)

Page: 1 2 3 4 5 6 >

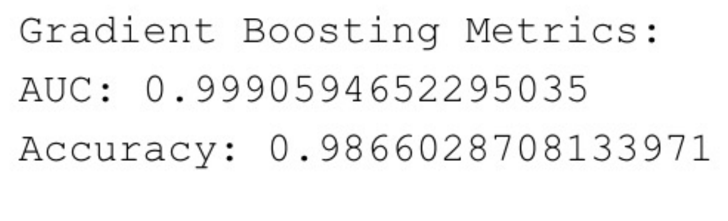
6 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
554	collect at /mnt/yarn/usercache/hadoop/appcache/application_1733099315105_0001/container_1733099315105_0001_01_000001/big_data_project.py:190 collect at /mnt/yarn/usercache/hadoop/appcache/application_1733099315105_0001/container_1733099315105_0001_01_000001/big_data_project.py:190	2024/12/02 00:34:54	92 ms	1/1	4/4
553	collect at /mnt/yarn/usercache/hadoop/appcache/application_1733099315105_0001/container_1733099315105_0001_01_000001/big_data_project.py:189 collect at /mnt/yarn/usercache/hadoop/appcache/application_1733099315105_0001/container_1733099315105_0001_01_000001/big_data_project.py:189	2024/12/02 00:34:54	0.1 s	1/1	4/4
552	collect at /mnt/yarn/usercache/hadoop/appcache/application_1733099315105_0001/container_1733099315105_0001_01_000001/big_data_project.py:188 collect at /mnt/yarn/usercache/hadoop/appcache/application_1733099315105_0001/container_1733099315105_0001_01_000001/big_data_project.py:188	2024/12/02 00:34:53	0.1 s	1/1	4/4
551	collect at /mnt/yarn/usercache/hadoop/appcache/application_1733099315105_0001/container_1733099315105_0001_01_000001/big_data_project.py:190 collect at /mnt/yarn/usercache/hadoop/appcache/application_1733099315105_0001/container_1733099315105_0001_01_000001/big_data_project.py:190	2024/12/02 00:34:53	80 ms	1/1	4/4
550	collect at /mnt/yarn/usercache/hadoop/appcache/application_1733099315105_0001/container_1733099315105_0001_01_000001/big_data_project.py:189	2024/12/02 00:34:53	0.1 s	1/1	4/4

Model Evaluation and Performance Metrics

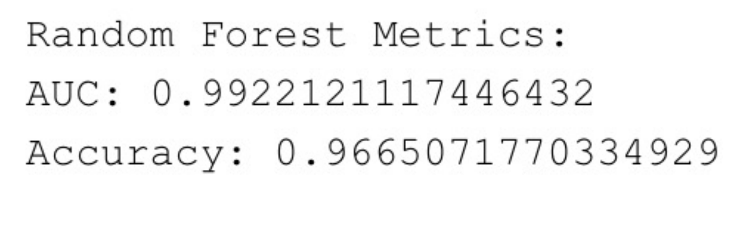
The system evaluates the performance of models using Area Under ROC (AUC), which measures the ability of a model to distinguish between positive and negative classes.

Gradient Boosting



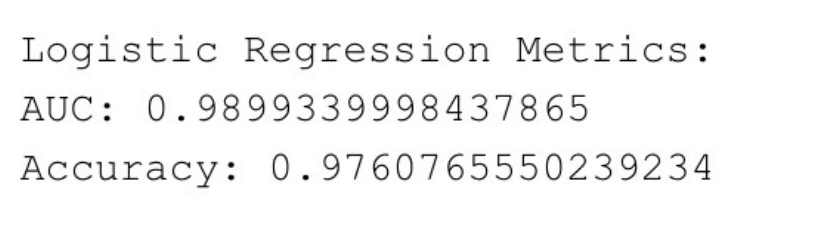
```
Gradient Boosting Metrics:
AUC: 0.9990594652295035
Accuracy: 0.9866028708133971
```

Random Forest



```
Random Forest Metrics:
AUC: 0.9922121117446432
Accuracy: 0.9665071770334929
```

Logistic Regression



```
Logistic Regression Metrics:
AUC: 0.9899339998437865
Accuracy: 0.9760765550239234
```

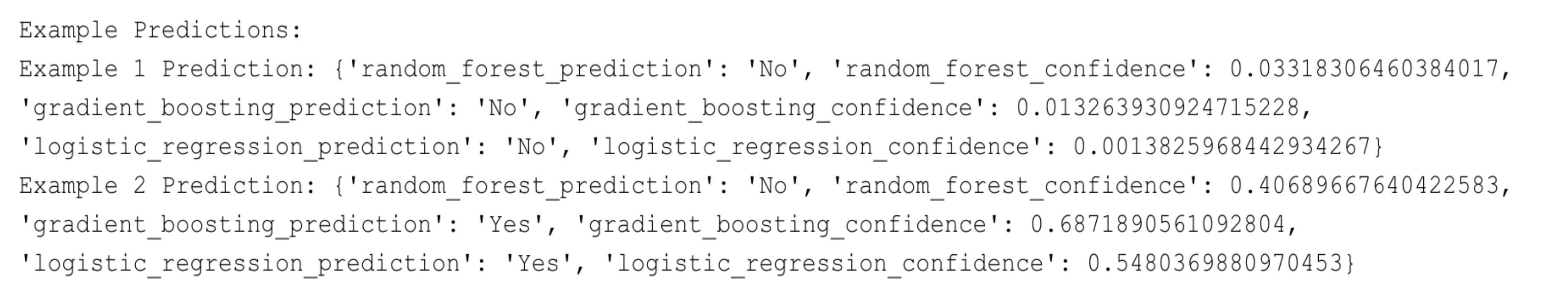
Prediction

Predictions

The CSAPS system provides predictions for COVID-19 status based on user-provided symptom data.

Real-Time Predictions

The system allows for real-time predictions based on user input.



```
Example Predictions:
Example 1 Prediction: {'random_forest_prediction': 'No', 'random_forest_confidence': 0.03318306460384017,
'gradient_boosting_prediction': 'No', 'gradient_boosting_confidence': 0.013263930924715228,
'logistic_regression_prediction': 'No', 'logistic_regression_confidence': 0.0013825968442934267}
Example 2 Prediction: {'random_forest_prediction': 'No', 'random_forest_confidence': 0.40689667640422583,
'gradient_boosting_prediction': 'Yes', 'gradient_boosting_confidence': 0.6871890561092804,
'logistic_regression_prediction': 'Yes', 'logistic_regression_confidence': 0.5480369880970453}
```

Key Considerations and Future Directions

Data Quality

Ensuring data quality is essential for accurate model training and predictions.

Hyperparameter Tuning

Model hyperparameters (e.g., tree depth, regularization) can be adjusted for performance tuning to optimize accuracy and generalization.

Neural Networks

Future development could incorporate neural network implementations to potentially enhance efficiency and improve prediction accuracy.

Conclusion

The CSAPS project is a promising tool for aiding in the diagnosis and prediction of COVID-19. By leveraging machine learning models and scalable computing capabilities, the system can effectively analyze symptom data and provide valuable insights to healthcare professionals.

Github link

https://github.com/ChrisJT47/CS777_Project/