Data Science Project Report

1. Principal Investigator:

Aneesha Patan Arifulla patanara1@newpaltz.edu

2. Title of the project:

Disease Outbreak Prediction

3. Introduction:

In recent years, the global community has witnessed the unprecedented impact of infectious diseases on public health, economies, and societies at large. The emergence of outbreaks such as influenza, Swine Flu, COVID-19, and various vector-borne diseases has shown the need for advanced tools and models to predict and mitigate the spread of these threats. In response to this critical need, this project aims to develop a robust predictive model for disease outbreaks. By utilizing the power of historical healthcare data, epidemiological insights, environmental considerations, and patient demographics, the objective is to create a comprehensive framework that enhances our ability to forecast and understand the dynamics of diseases.

3.1 Project Motivation

The motivation behind this project is the increasing challenges posed by infectious diseases to global health security. Recent pandemics have demonstrated the swift and unpredictable nature of disease spread, emphasizing the necessity for proactive measures and predictive models. By delving into the complexities of healthcare data, epidemiology, and environmental factors, this

project contributes to the development of a reliable tool that can assist healthcare professionals, policymakers, and communities in preparing for and responding to potential disease outbreaks effectively. The motivation is to empower individuals and organizations with the knowledge needed to implement timely interventions and mitigate the impact of infectious diseases.

3.2 Aims and Objectives

This project aims to develop a predictive model capable of forecasting the spread of diseases accurately. To achieve this goal, the following specific objectives will be pursued:

- **Data Integration:** Compile and integrate diverse datasets, including historical healthcare records, epidemiological data, environmental variables, and demographic information, to create a comprehensive database for analysis.

- **Model Development:** Design and implement a predictive model that uses machine learning algorithms and statistical techniques to analyze the integrated dataset. The model should be capable of identifying correlations, and trends associated with disease outbreaks.

- **Accuracy Improvement:** Continuously refine and optimize the predictive model to enhance its accuracy and reliability. This involves updating the model based on new data and incorporating feedback from real-world disease events.

- **Validation and Testing:** Validate the predictive model using historical outbreak data and conduct thorough testing to assess its performance under various scenarios and conditions.

4. Background/History of the Study:

Throughout history, disease outbreaks have profoundly impacted public health and economies, necessitating effective prediction and response strategies. Traditionally reliant on expert opinions and past experiences, contemporary approaches integrate diverse data sources, including healthcare records, epidemiology, and environmental factors. This study aims to contribute to this evolution by developing a data-driven predictive model for disease outbreaks.

Departing from conventional methods, this model seeks to enhance accuracy and preparedness, empowering actionable insights. By leveraging advanced analytics, the project aims to surpass historical limitations, providing a comprehensive framework that enables proactive measures and informed decision-making. The goal is to provide a resilient global society capable of mitigating the impact of infectious diseases on public health and well-being.

5. Approach and Implementation:

Approach:

- **Problem Definition:** Predict Covid, and Swine Flu outbreaks based on historical healthcare data, epidemiological information, environmental factors, and patient demographics.

- **Data Collection:** Gather diverse datasets, including historical healthcare records, epidemiological data, environmental variables, and patient demographics.

- **Data Preprocessing:**
  - Handle missing values using imputation techniques.
  - Encode categorical variables using Label Encoding or One-Hot Encoding.
  - Normalize or standardize numerical features.

- **Exploratory Data Analysis:**
  - Visualize data distributions, correlations, and disease prevalence.
  - Identify patterns and insights that can guide feature selection.

- **Feature Selection:**

  Remove less relevant features based on correlation, statistical tests, or domain knowledge.

- **Model Selection:**

  Choose machine learning models suitable for classification tasks, such as Support Vector Machines (SVM), Random Forests, or Neural Networks.

- **Model Training & Refinement:**
  - Divide the dataset into training, validation, and test subsets.
  - Pick models like SVM, Random Forest, and Neural Networks.
  - Train each model using the training data.

- Adjust model parameters to enhance performance using the validation set.

- **Model Assessment:**
  - Assess models using metrics such as accuracy, precision, recall, F1 score, and confusion matrix.
  - Compare model performance on the test set.
  - Choose the model with the most superior overall performance.

- **Development of User-Friendly Interface:**

  This can be implemented as an extension of this project.
  - Construct a straightforward interface for user input and result presentation.
  - Ensure clarity, user-friendliness, and robust error handling.
  - Continuously refine the interface for an enhanced user experience.

Implementation:

- **Data Preprocessing and EDA:**

  Use libraries like pandas, NumPy, and seaborn for data manipulation and visualization.

- **Feature Selection:**

  Implement feature selection based on correlation analysis or feature importance from models.

- **Model Training and Evaluation:**

Use popular machine learning libraries (e.g., sci-kit-learn, TensorFlow, Pytorch) for training and evaluating models.

- **Support Vector Machine:** SVM is a supervised machine learning algorithm that aims to find a hyperplane in N-dimensional space (N is the number of features) that distinctly classifies data points into different categories, making it effective for both linear and non-linear classification tasks.

```
X_train_covid, X_test_covid, Y_train_covid, Y_test_covid = train_test_split(X, Y, test_size=0.2, random_state=42)

svm_model_covid = SVC(kernel='linear', random_state=0)
svm_model_covid.fit(X_train_covid, Y_train_covid.values.ravel())

svm_predictions_covid = svm_model_covid.predict(X_test_covid)


svm_accuracy_covid = accuracy_score(Y_test_covid, svm_predictions_covid)
print(f'SVM Accuracy for Covid: {svm_accuracy_covid}')


x_train_swineflu, x_test_swineflu, Z_train_swineflu, Z_test_swineflu = train_test_split(X, Z, test_size=0.3, random_state=35)


svm_model_swineflu = SVC(kernel='linear', random_state=0)
svm_model_swineflu.fit(x_train_swineflu, Z_train_swineflu.values.ravel())

svm_predictions_swineflu = svm_model_swineflu.predict(x_test_swineflu)

svm_accuracy_swineflu = accuracy_score(Z_test_swineflu, svm_predictions_swineflu)
print(f'SVM Accuracy for SwineFlu: {svm_accuracy_swineflu}')
```

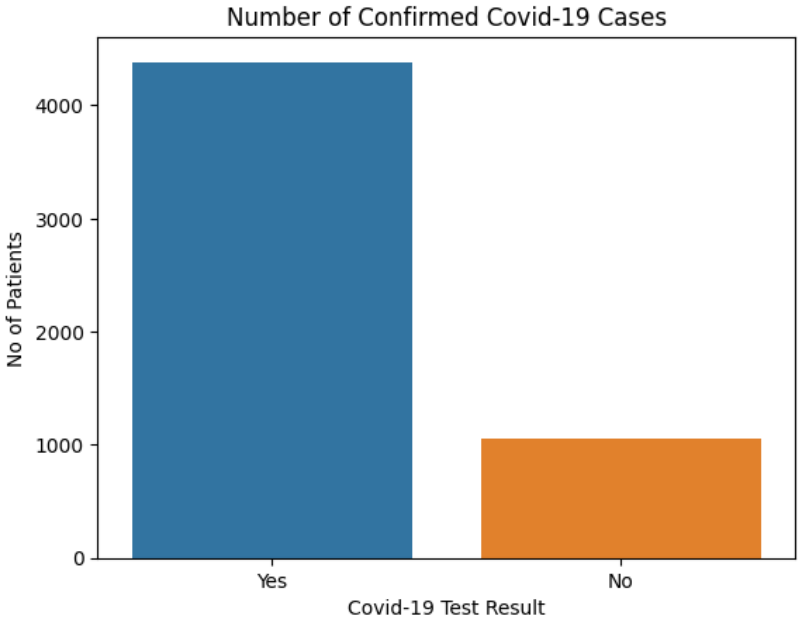
- **Random Forest:** Random Forest is an ensemble learning algorithm that builds multiple decision trees during training and outputs the mode of the classes or the mean prediction of the individual trees, providing robustness and high accuracy by mitigating overfitting.

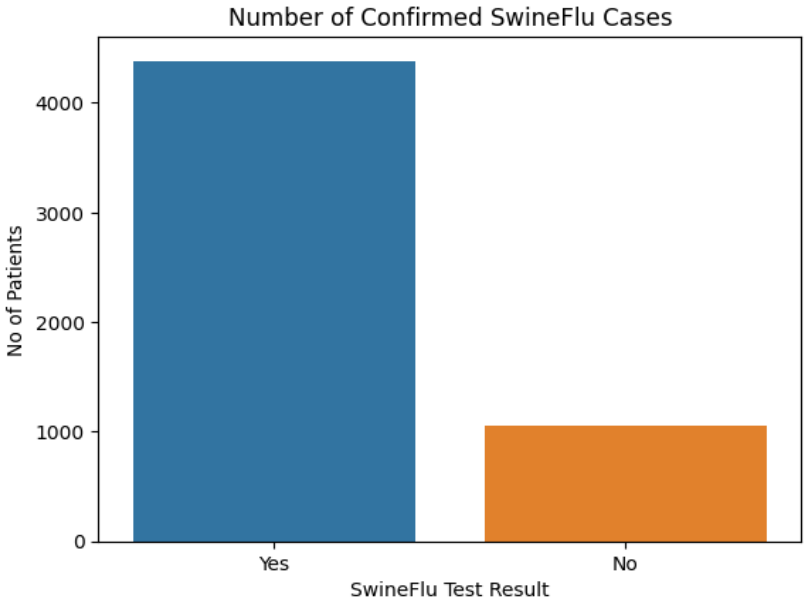
```
X_train_covid, X_test_covid, Y_train_covid, Y_test_covid = train_test_split(X, Y, test_size=0.2, random_state=42)

rf_model_covid = RandomForestClassifier(n_estimators=10, criterion='entropy', random_state=0)
rf_model_covid.fit(X_train_covid, Y_train_covid.values.ravel())

rf_predictions_covid = rf_model_covid.predict(X_test_covid)

rf_accuracy_covid = accuracy_score(Y_test_covid, rf_predictions_covid)
print(f'Random Forest Accuracy for Covid: {rf_accuracy_covid}')

x_train_swineflu, x_test_swineflu, Z_train_swineflu, Z_test_swineflu = train_test_split(X, Z, test_size=0.3, random_state=35)

rf_model_swineflu = RandomForestClassifier(n_estimators=10, criterion='entropy', random_state=0)
rf_model_swineflu.fit(x_train_swineflu, Z_train_swineflu.values.ravel())

rf_predictions_swineflu = rf_model_swineflu.predict(x_test_swineflu)

rf_accuracy_swineflu = accuracy_score(Z_test_swineflu, rf_predictions_swineflu)
print(f'Random Forest Accuracy for SwineFlu: {rf_accuracy_swineflu}')
```

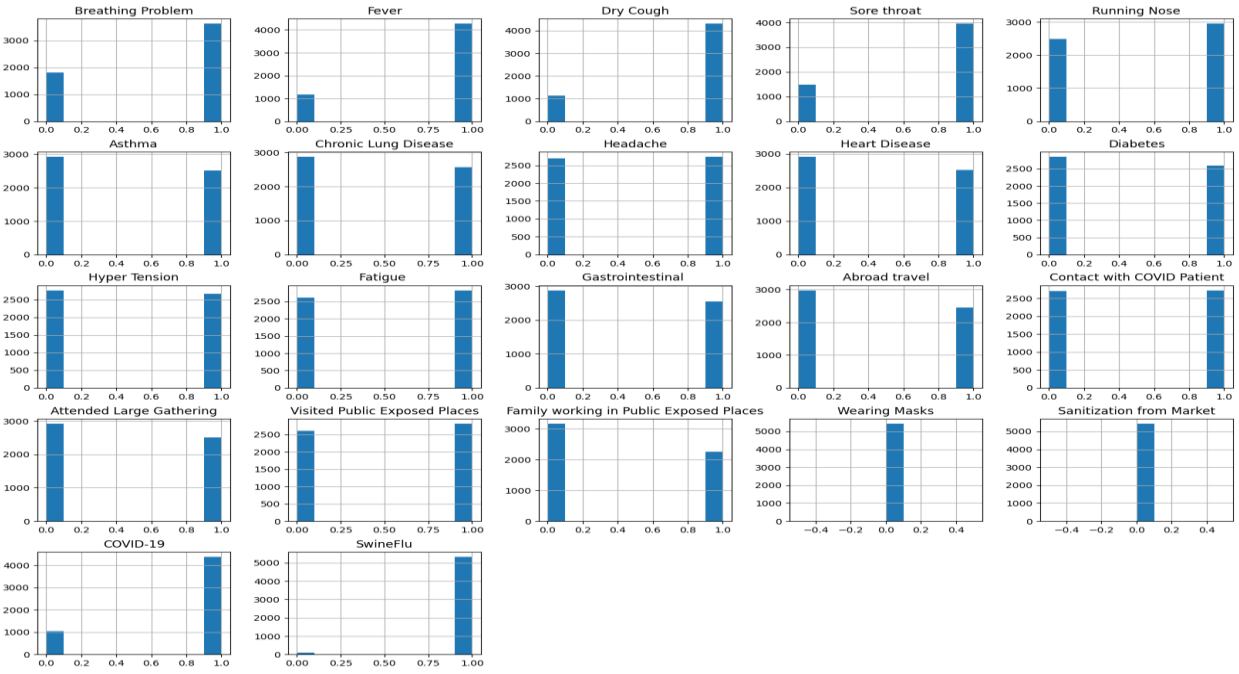
- **Artificial Neural Networks (ANN):** ANN is a deep learning algorithm inspired by the human brain's neural structure, consisting of interconnected nodes organized in layers; it's particularly effective for complex tasks, including image and speech recognition, by learning hierarchical representations from the data.

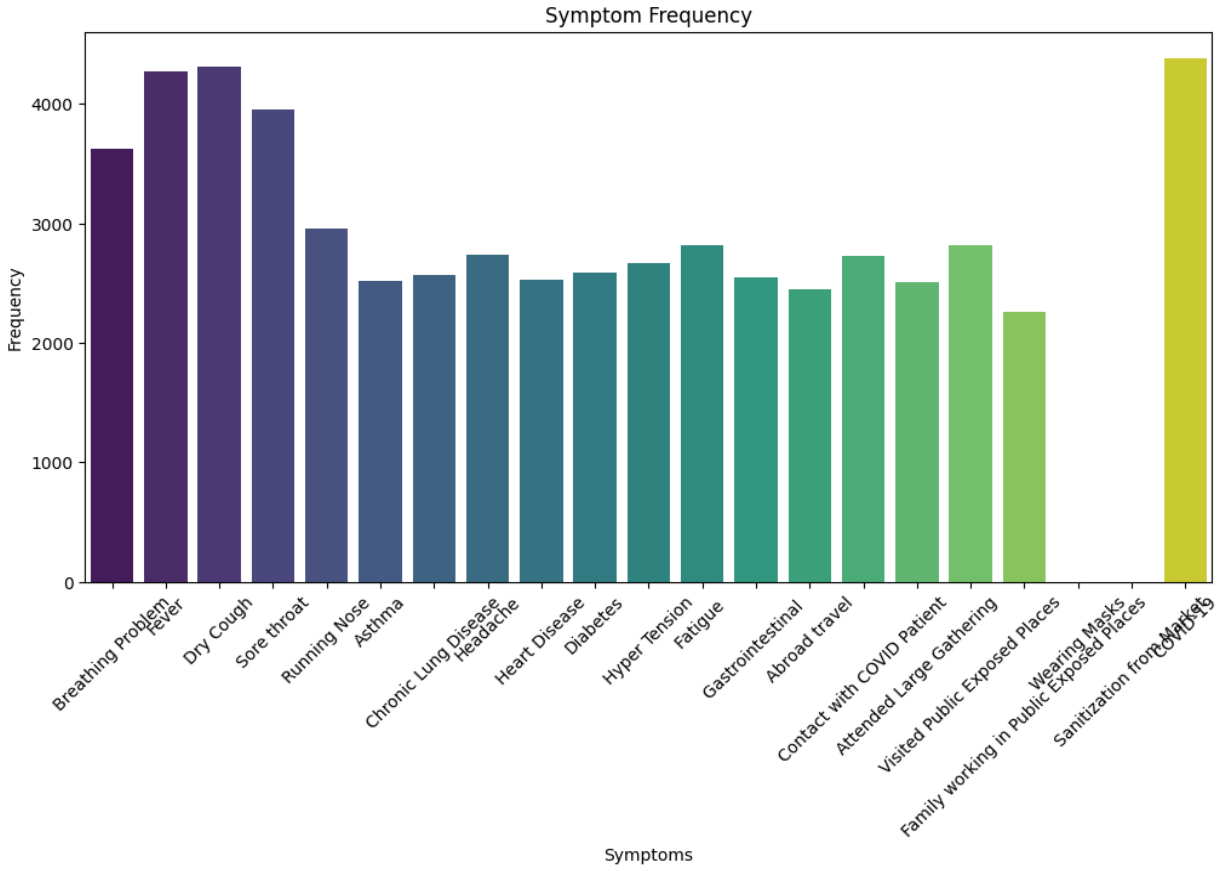
```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
x_train, x_test, Z_train, Z_test = train_test_split(X, Z, test_size=0.3, random_state=35)

ann_model = tf.keras.Sequential([
    tf.keras.layers.Dense(units=6, activation='relu', input_dim=X_train.shape[1]),
    tf.keras.layers.Dense(units=6, activation='relu'),
    tf.keras.layers.Dense(units=1, activation='sigmoid')
])

ann_model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# Train the model
ann_model.fit(X_train, Y_train, epochs=10, batch_size=32, validation_data=(X_test, Y_test))
ann_model.fit(x_train, Z_train, epochs=10, batch_size=32, validation_data=(x_test, Z_test))
# Predict on the test set
ann_predictions = (ann_model.predict(X_test) > 0.5).astype("int32")

# Calculate and print accuracy
ann_accuracy = accuracy_score(Y_test, ann_predictions)
print(f'ANN Accuracy for Covid: {ann_accuracy}')

ann_predictions = (ann_model.predict(x_test) > 0.5).astype("int32")

# Calculate and print accuracy
ann_accuracy = accuracy_score(Z_test, ann_predictions)
```

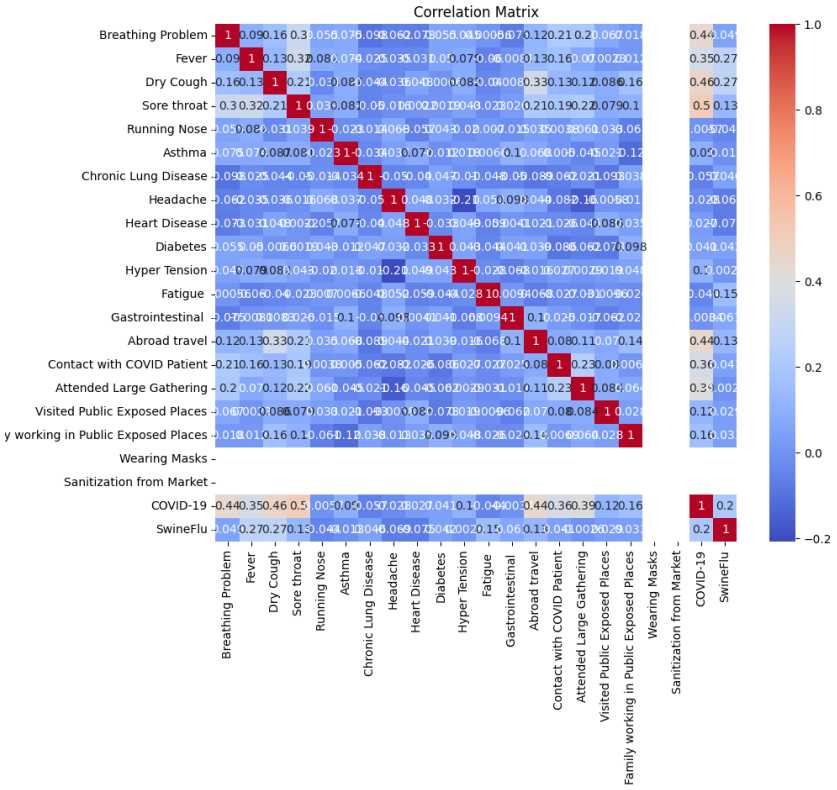
6. Experimental Results and Discussion:

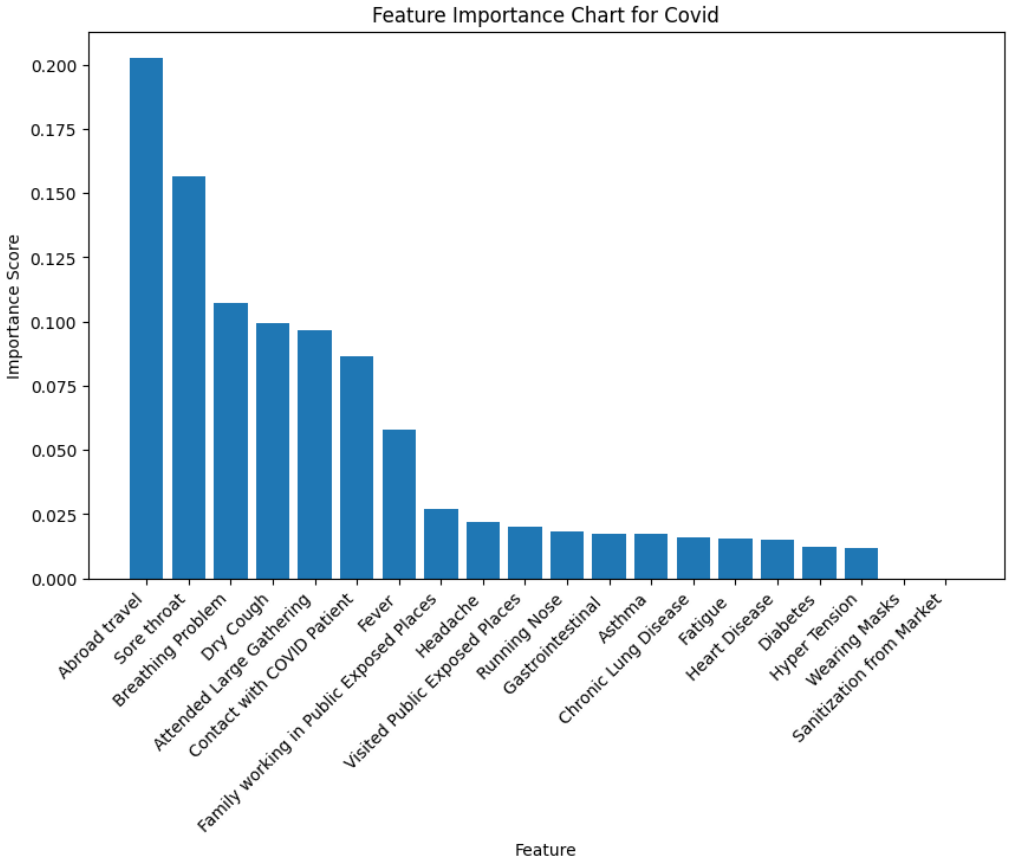


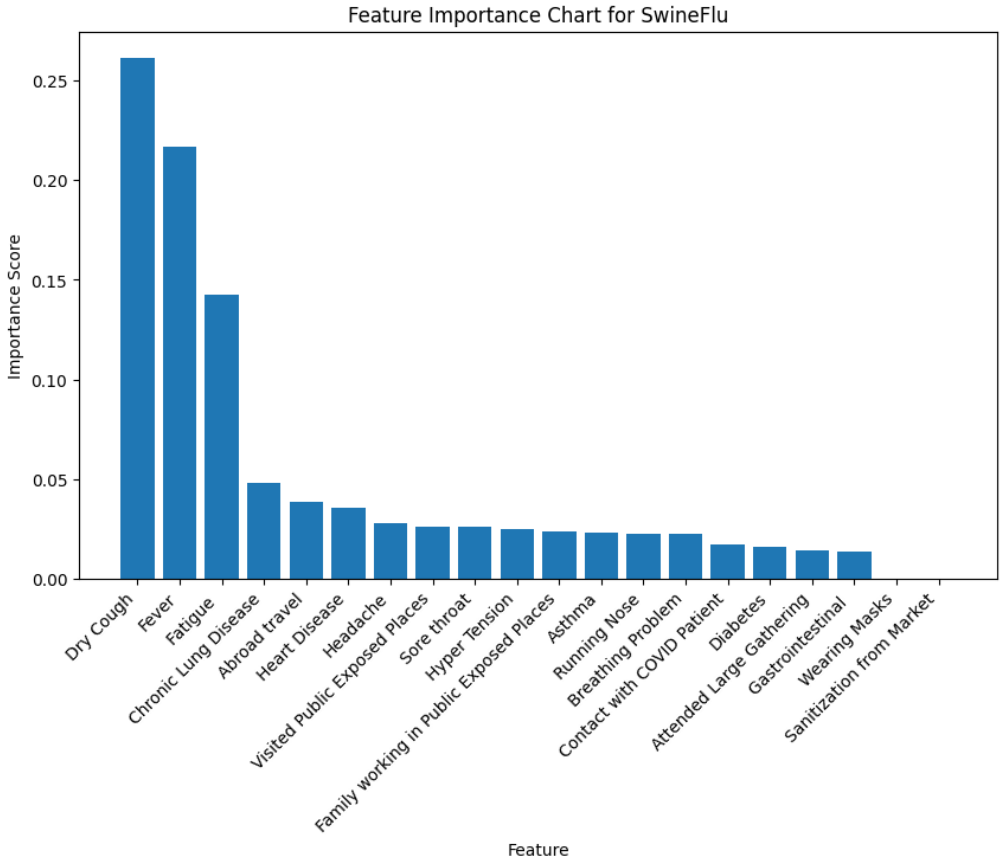



Breathing Problem
Fever
Dry Cough
Sore throat
Running Nose
Asthma
Chronic Lung Disease
Headache
Heart Disease
Diabetes
Hyper Tension
Fatigue
Gastrointestinal
Abroad travel
Contact with COVID Patient
Attended Large Gathering
Visited Public Exposed Places
Family working in Public Exposed Places
Wearing Masks
Sanitization from Market
COVID-19
SwineFlu


Symptom Frequency
Frequency
Symptoms
Breathing Problem
Fever
Dry Cough
Sore throat
Running Nose
Asthma
Chronic Lung Disease
Headache
Heart Disease
Diabetes
Hyper Tension
Fatigue
Gastrointestinal
Abroad travel
Contact with COVID Patient
Attended Large Gathering
Visited Public Exposed Places
Family working in Public Exposed Places
Wearing Masks
Sanitization from Market
COVID-19


Correlation Matrix
Breathing Problem
Fever
Dry Cough
Sore throat
Running Nose
Asthma
Chronic Lung Disease
Headache
Heart Disease
Diabetes
Hyper Tension
Fatigue
Gastrointestinal
Abroad travel
Contact with COVID Patient
Attended Large Gathering
Visited Public Exposed Places
Family working in Public Exposed Places
Wearing Masks
Sanitization from Market
COVID-19
SwineFlu
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Feature Importance Chart for Covid
Importance Score
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Abroad travel
Sore throat
Breathing Problem
Dry Cough
Attended Large Gathering
Contact with COVID Patient
Fever
Family working in Public Exposed Places
Headache
Visited Public Exposed Places
Running Nose
Gastrointestinal
Asthma
Chronic Lung Disease
Fatigue
Heart Disease
Diabetes
Hyper Tension
Wearing Masks
Sanitization from Market
Feature




```
Enter values for each feature:
Breathing Problem: 0
Fever: 1
Dry Cough: 1
Sore throat: 1
Running Nose: 1
Asthma: 0
Chronic Lung Disease: 1
Headache: 1
Heart Disease: 0
Diabetes: 1
Hyper Tension: 1
Fatigue : 1
Gastrointestinal : 0
Abroad travel: 0
Contact with COVID Patient: 1
Attended Large Gathering: 0
Visited Public Exposed Places: 1
Family working in Public Exposed Places: 1
Wearing Masks: 0
Sanitization from Market: 0
The model predicts that you have Covid.
```

In a comparative evaluation of machine learning models for predicting Covid and Swine Flu, the Artificial Neural Network (ANN) achieved an accuracy of 82.8% for Covid and an impressive 99.8% for Swine Flu. Meanwhile, the Random Forest model demonstrated high accuracy with 98.4% for COVID-19 and perfect accuracy (100%) for Swine Flu. The Support Vector Machine (SVM) model exhibited strong predictive performance with 97.3% accuracy for COVID-19 and perfect accuracy for Swine Flu. These results highlight the effectiveness of machine learning in distinguishing between the presence and absence of these diseases, with Random Forest being particularly notable for its robust performance.

7. Conclusion:

In summary, this project aimed to predict disease outbreaks using various models like SVM, Random Forest, and Neural Networks. The results show promise, especially when combined with a user-friendly interface. By creating a website and UI, we can make this tool more accessible and beneficial for healthcare professionals and policymakers. The website's simplicity and real-time updates could aid in making proactive decisions during potential outbreaks. Moving forward, ongoing improvements to the model and website will be crucial to keeping the system effective in addressing evolving public health challenges.