CHIKUNGUNYA PREDICTION

Project Description:

Chikungunya is a disease caused by a virus that spreads through the bites of infected mosquitoes, primarily Aedes aegypti and Aedes albopictus. People infected with chikungunya often experience symptoms such as fever, fatigue, vomiting, arthritis, conjunctivitis, chills, swelling, and rashes. In severe cases, the disease can cause prolonged joint pain that affects a person's mobility and quality of life. These symptoms, however, are not unique to chikungunya and may overlap with other diseases like dengue or malaria, making it difficult to identify chikungunya accurately and quickly.

To address this challenge, we are developing a system that can predict whether a person has chikungunya or not based on their symptoms. The prediction models will include **Linear Regression**, **Logistic Regression**, **K-Nearest Neighbors (KNN)** and**.** This system uses advanced machine learning techniques to analyse symptom inputs such as fever, fatigue, vomiting, arthritis, and other common indicators. The goal is to provide an accurate prediction of chikungunya based on these symptoms without the need for expensive or time-consuming laboratory tests.

Approach:

In this project, we aim to predict the severity of chikungunya symptoms in patients based on features like fever, sex, vomiting, joint pain intensity, rash presence, and other clinical markers. The dataset, sourced from a reliable platform like **Kaggle**, includes comprehensive patient data, such as symptom details, clinical history, and severity levels categorized as mild, moderate, or severe.

We preprocess the data to handle missing values, encode categorical variables like sex and rash presence into numerical values, and scale continuous variables such as fever duration and joint pain intensity to ensure uniformity. The processed data is then used to train three **machine learning models**: **Linear Regression, Logistic Regression, and K-Nearest Neighbors (KNN).**

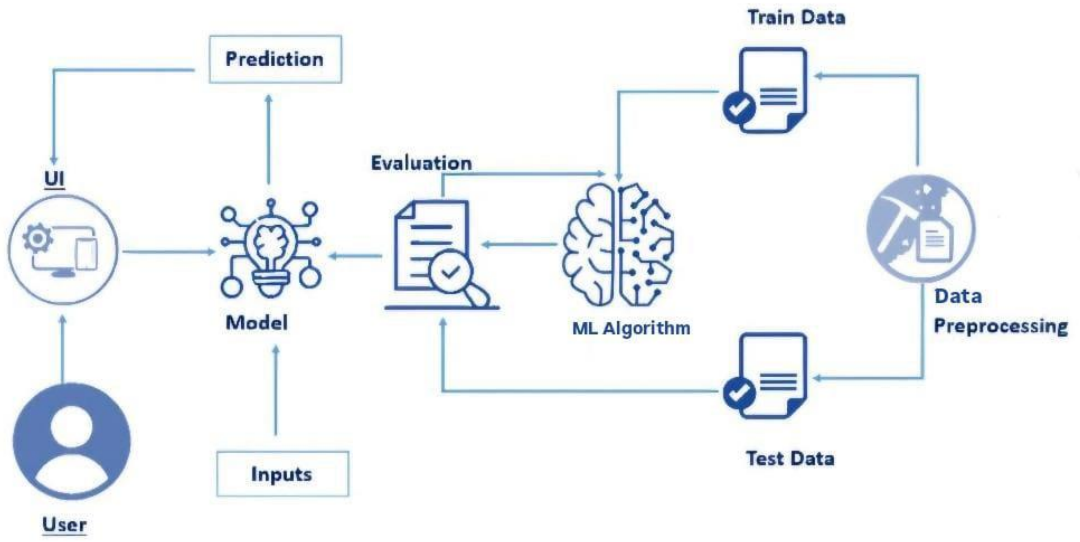
The models are evaluated using metrics like accuracy, confusion matrix, precision, recall, and F1-score to assess their ability to predict symptom severity accurately. To enhance performance, hyperparameter optimization techniques are applied, such as tuning the number of neighbors for KNN or adjusting regularization in Logistic Regression.

By the end of the project, we identify the best-performing model, which can provide reliable predictions for chikungunya symptom severity based on key clinical features. This solution has the potential to support healthcare providers in making timely and informed decisions for patient care.

1. Data Collection and Exploration
2. Data Preprocessing
3. Model Development
4. Model Evaluation

	sex	fever	cold	joint pains	myalgia	headache	fatigue	vomitting	arthritis	Conjuctivit	Nausea	Maculopa	Eye Pain	Chills	Swelling	Severe	Chikungunya	
2	male	yes	yes	no	no	no	no	yes	no	no	yes	yes	yes	yes	yes	yes	yes	male
3	female	yes	yes	yes	no	no	yes	no	yes	no	no	no	no	no	yes	yes	no	female
4	male	yes	no	no	yes	yes	yes	no	yes	no	yes	yes	yes	no	yes	no	yes	male
5	female	no	yes	yes	no	yes	no	yes	yes	yes	yes	yes	no	no	yes	yes	yes	
6	male	yes	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no	no	
7	male	no	yes	yes	yes	yes	no	no	no	no	yes	yes	yes	no	yes	yes		
8	male	no	yes	no	no	yes	no	no	yes	yes	yes	yes	yes	yes	yes	no		
9	female	yes	no	no	no	yes	yes	no	no	yes	yes	yes	no	yes	no	yes		
10	male	yes	no	yes	yes	no	no	no	no	yes	yes	no	no	yes	yes	yes		
11	female	yes	yes	yes	yes	no	no	no	no	yes	yes	no	yes	yes	yes	yes		
12	female	no	yes	yes	yes	no	yes	no	yes	yes	no	no	yes	yes	yes	yes		
13	male	no	yes	no	yes	yes	yes	yes	no	no	yes	yes	yes	yes	yes	no		
14	female	yes	yes	no	yes	yes	yes	yes	yes	no	no	no	yes	yes	yes	no		
15	male	yes	yes	yes	no	yes	no	yes	yes	yes	yes	no	yes	yes	yes	no		

Technical Architecture:



Prerequisites:

To complete this project, you must require the following software's, concepts, and packages

1. **Google Collab:** It provides a cloud-based environment to run Python code and access various resources.

2. **Python Libraries**:

- **NumPy:** For numerical computations and handling arrays.

- **Pandas:** For loading, exploring, and manipulating datasets.
- **Scikit-learn:** For building machine learning models and preprocessing.
- **Matplotlib:** For visualizing the data and model evaluation results.

3. **Machine Learning Models and Classification Algorithms:** Understanding classification algorithms, where the task is to predict a categorical label (alive or dead in this case).

- **Linear Regression:** A simple model used for predicting continuous numerical outcomes.
- **Logistic Regression:** A simple model used for binary classification.
- **KNN**: A simple model used for classification or regression by finding the majority class or average of the k-nearest neighbors.

4. **Dataset:** Dataset Source: The dataset used in this project is from Kaggle. A simple model used for predicting the severity of chikungunya symptoms based on similar cases from the dataset.

Link: https://www.kaggle.com/datasets/richardbernat/chikungunya-symptom-data

Project Objectives:

By the end of this project, you will know:

- How to preprocess and clean the chikungunya-symptom dataset.
- Apply machine learning models like Linear Regression, Logistic Regression, k-nearest neighbors.
- How to evaluate model performance using accuracy, precision, recall, and other metrics.
- How to fine-tune the models for better predictions.
- How to select the best model for predicting chikungunya symptoms.

Project Flow:

1. **Data Collection from Kaggle:**

   - Download the chikungunya symptoms dataset from **Kaggle**. The dataset includes features such as **sex, fever , cold, joint pain , myalgia, headache, fatigue, vomiting, arthritis, Conjuctivitis, Nausea, maculopapular rash, swelling, eye pain, chills ,and the severity level of chikungunya symptoms (mild, moderate, or severe).**
   - Import the dataset into your working environment (Google Collab or Jupyter Notebook) using **Pandas**.

2. **Data Preprocessing:**

   - **Handle Missing Data**: Check for and handle any missing values in the dataset (e.g., by imputing or dropping).
   - **Encode Categorical Data**: Convert categorical variables like sex, fever, cold, joint pain, myalgia, headache, fatigue, vomiting, arthritis, Conjuctivitis, Nausea, maculopapular rash, swelling, eye pain, chill, and the severity level of chikungunya symptoms (mild, moderate, or severe) status (yes/no) into numerical values using **Label Encoding** or **One-Hot Encoding**.
   - **Feature Scaling**: Normalize or standardize continuous features such as tumor size using **Standard Scaler** to bring all features to a similar scale.
   - **Train-Test Split**: Split the dataset into training and test sets (usually an 80-20 split) to ensure the model can be evaluated on unseen data.

3. **Model Building:**

   - Build and train different machine learning models using **Linear Regression, Logistic Regression, k-nearest neighbors.**
   - Train each model using the training data and the relevant features (e.g., like sex, fever, cold, joint pain, myalgia, headache, fatigue, vomiting, arthritis and etc.).

4. **Model Evaluation:**

   Evaluate each model using performance metrics like **accuracy**, **precision**, **recall**, **F1-score**, and the **confusion matrix** to check how well each model performs on the test data.

5. **Predictions:**

- Make predictions on the test set using each trained model.
- Compare the predicted survival status (alive or dead) with the actual values from the test set to assess prediction accuracy.

6. **Results Display:**

   - Display and compare the performance of each model using **classification reports**
   - Analyze which model performs the best in predicting chikungunya symptoms survival, and select the most accurate one based on the evaluation metrics.

Project Structure:

Milestone 1: Data Collection

The success of a chikungunya prediction model heavily depends on the quality, diversity, and representativeness of the data used.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
sex	fever	cold	joint pains	myalgia	headache	fatigue	vomitting	arthritis	Conjuctivit	Nausea	Maculopa|	Eye Pain	Chills	Swelling	Severe Chikungunya		
male	yes	yes	no	no	no	no	yes	no	no	yes	yes	yes	yes	yes	yes	yes	male
female	yes	yes	yes	no	no	yes	no	yes	no	no	no	no	no	yes	yes	no	female
male	yes	no	no	yes	yes	yes	no	yes	no	yes	yes	yes	no	yes	no	yes	male
female	no	yes	yes	no	yes	no	yes	yes	yes	yes	yes	no	no	yes	yes	yes	
male	yes	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no	no	
male	no	yes	yes	yes	yes	no	no	no	no	yes	yes	yes	no	yes	yes		
male	no	yes	no	no	yes	no	no	yes	yes	yes	yes	yes	yes	yes	no		
female	yes	no	no	no	yes	yes	no	no	yes	yes	yes	no	yes	no	yes		
male	yes	no	yes	yes	no	no	no	no	yes	yes	no	no	yes	yes	yes		
female	yes	yes	yes	yes	no	no	no	no	yes	yes	no	yes	yes	yes	yes		
female	no	yes	yes	yes	no	yes	no	yes	yes	no	no	yes	yes	yes	yes		
male	no	yes	no	yes	yes	yes	yes	no	no	yes	yes	yes	yes	yes	no		
female	yes	yes	no	yes	yes	yes	yes	yes	no	no	no	yes	yes	yes	no		
male	yes	yes	yes	no	yes	no	yes	yes	yes	yes	no	yes	yes	yes	no		
male	no	yes	yes	yes	yes	no	yes	no	yes	yes	no	no	yes	yes	yes		
male	yes	no	yes	no	yes	yes	no	yes	yes	yes	no	no	yes	yes	yes		
male	yes	yes	yes	no	no	no	yes	yes	yes	yes	no	yes	yes	no	yes		
female	no	yes	yes	yes	no	no	no	yes	no	yes	yes	yes	yes	yes	no		
male	no	yes	yes	yes	no	no	no	yes	no	yes	no	no	yes	yes	no		
male	yes	yes	no	yes	yes	yes	yes	yes	no	no	no	yes	yes	no	yes		
female	yes	yes	yes	no	yes	yes	no	yes	yes	yes	no	yes	no	yes	no		
female	yes	yes	yes	yes	yes	yes	yes	yes	no	no	yes	yes	no	no	yes		
male	yes	yes	yes	yes	yes	no	yes	yes	no	no	no	yes	yes	no	no		
male	yes	yes	yes	yes	no	yes	no	no	no	no	yes	no	yes	yes	yes		
female	no	no	yes	yes	yes	no	no	no	no	yes	no	yes	yes	yes	no		
male	no	no	yes	yes	yes	yes	no	no	yes	no	no	no	yes	no	yes		

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
976	female	yes	yes	no	no	yes	yes	no	yes	no	yes	no	no	no	no	yes
977	female	yes	no	yes	yes	no	no	no	yes	yes	no	yes	no	yes	no	yes
978	male	no	yes	yes	no	yes	no	yes	yes	no	yes	yes	yes	yes	no	no
979	female	yes	no	yes	yes	yes	no	no	yes	yes	yes	no	no	no	yes	yes
980	female	yes	no	no	no	no	yes	no	no	yes	yes	yes	no	no	yes	yes
981	female	yes	no	yes	yes	yes	yes	yes	yes	no	no	yes	yes	yes	no	no
982	male	yes	no	no	yes	yes	yes	no	no	yes	yes	yes	no	yes	no	yes
983	female	yes	yes	no	yes	yes	yes	no	yes	yes	yes	yes	yes	yes	yes	yes
984	male	no	yes	no	yes	yes	no	yes	yes	no	no	yes	no	yes	yes	yes
985	male	no	no	yes	yes	yes	yes	yes	no	yes	no	no	no	no	yes	yes
986	female	yes	no	no	yes	yes	yes	yes	no	no	yes	yes	no	yes	yes	yes
987	male	no	no	yes	yes	yes	yes	no	yes	yes	no	yes	yes	yes	yes	no
988	male	yes	no	yes	no	yes	no	yes	yes	yes	no	no	yes	no	no	yes
989	female	no	no	no	no	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
990	male	no	yes	yes	yes	yes	yes	yes	no	yes	yes	no	yes	no	yes	yes
991	female	yes	yes	yes	no	no	no	yes	yes	yes	no	no	no	no	yes	yes
992	female	yes	yes	yes	no	no	no	yes	no	no	yes	no	no	yes	no	yes
993	male	no	no	yes	yes	no	no	yes	yes	yes	yes	no	yes	no	no	no
994	female	yes	no	no	yes	no	no	no	no	no	yes	yes	no	yes	yes	no
995	male	yes	no	yes	yes	no	yes	yes	yes	yes	no	yes	no	yes	no	no
996	female	yes	yes	yes	yes	no	yes	no	yes	no	no	yes	no	no	no	no
997	female	no	yes	yes	yes	no	no	yes	yes	yes	no	yes	yes	yes	yes	no
998	female	no	no	yes	yes	yes	yes	no	yes	yes	no	no	no	yes	no	no
999	female	yes	yes	no	yes	no	no	yes	yes	no	no	no	yes	yes	yes	yes
1000	female	no	no	yes	yes	no	no	no	yes	no	no	yes	no	yes	no	yes
1001	female	yes	no	yes	no	no	no	yes	no	yes	no	yes	no	yes	yes	yes
1002	male	yes	yes	yes	yes	yes	no	no	yes	no	yes	yes	no	yes	yes	yes

Milestone2: Data Preprocessing

For a chikungunya prediction project, data preprocessing is a crucial step in preparing the dataset for machine learning. It ensures that the raw data is cleaned, transformed, and formatted in a way that improves the performance of the model. The preprocessing steps are based on the category of data because we use features like symptoms (fever, joint pain, rash sex, fever , cold, joint pain , myalgia, headache, fatigue, vomiting, arthritis, Conjunctivitis, Nausea, maculopapular rash, swelling, eye pain, chills ,and the severity level of chikungunya symptoms (mild, moderate, or severe).

casestudy on chikungunya.ipynb

File Edit View Insert Runtime Tools Help Last edited on December 4

Share

+ Code + Text

Connect Gemini

```
[ ] df.head(5)
```

	sex	fever	cold	joint pains	myalgia	headache	fatigue	vomitting	arthritis	Conjuctivitis	Nausea	Maculopapular rash	Eye Pain	Chills	Swelling	Severe Chikungunya	Unnamed: 16	Unnamed: 17
0	male	yes	yes	no	no	no	no	yes	no	no	yes	yes	yes	yes	yes	yes	yes	male
1	female	yes	yes	yes	no	no	yes	no	yes	no	no	no	no	no	yes	yes	no	female
2	male	yes	no	no	yes	yes	yes	no	yes	no	yes	yes	yes	no	yes	no	yes	male
3	female	no	yes	yes	no	yes	no	yes	yes	yes	yes	yes	no	no	yes	yes	yes	NaN
4	male	yes	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no	no	NaN

```
df.tail(5)
```

	sex	fever	cold	joint pains	myalgia	headache	fatigue	vomitting	arthritis	Conjuctivitis	Nausea	Maculopapular rash	Eye Pain	Chills	Swelling	Severe Chikungunya	Unnamed: 16	Unnamed: 17
996	female	no	no	yes	yes	yes	yes	no	yes	yes	no	no	no	yes	no	no	NaN	NaN
997	female	yes	yes	no	yes	no	no	yes	yes	no	no	no	yes	yes	yes	yes	NaN	NaN
998	female	no	no	yes	yes	no	no	no	yes	no	no	yes	no	yes	no	yes	NaN	NaN
999	female	yes	no	yes	no	no	no	yes	no	yes	no	yes	no	yes	yes	yes	NaN	NaN
1000	male	yes	yes	yes	yes	yes	no	no	yes	no	yes	yes	no	yes	yes	yes	NaN	NaN

casestudy on chikungunya.ipynb

File Edit View Insert Runtime Tools Help Last edited on December 4

Share

+ Code + Text

Connect Gemini

```
[ ] df['Severe Chikungunya'].value_counts()
```

Severe Chikungunya	count
yes	594
no	407

dtype: int64

```
df.tail()
```

	sex	fever	cold	joint pains	myalgia	headache	fatigue	vomitting	arthritis	Conjuctivitis	Nausea	Maculopapular rash	Eye Pain	Chills	Swelling	Severe Chikungunya	Unnamed: 16	Unnamed: 17
996	female	no	no	yes	yes	yes	yes	no	yes	yes	no	no	no	yes	no	no	NaN	NaN
997	female	yes	yes	no	yes	no	no	yes	yes	no	no	no	yes	yes	yes	yes	NaN	NaN
998	female	no	no	yes	yes	no	no	no	yes	no	no	yes	no	yes	no	yes	NaN	NaN
999	female	yes	no	yes	no	no	no	yes	no	yes	no	yes	no	yes	yes	yes	NaN	NaN
1000	male	yes	yes	yes	yes	yes	no	no	yes	no	yes	yes	no	yes	yes	yes	NaN	NaN

```
[ ] df.shape
```

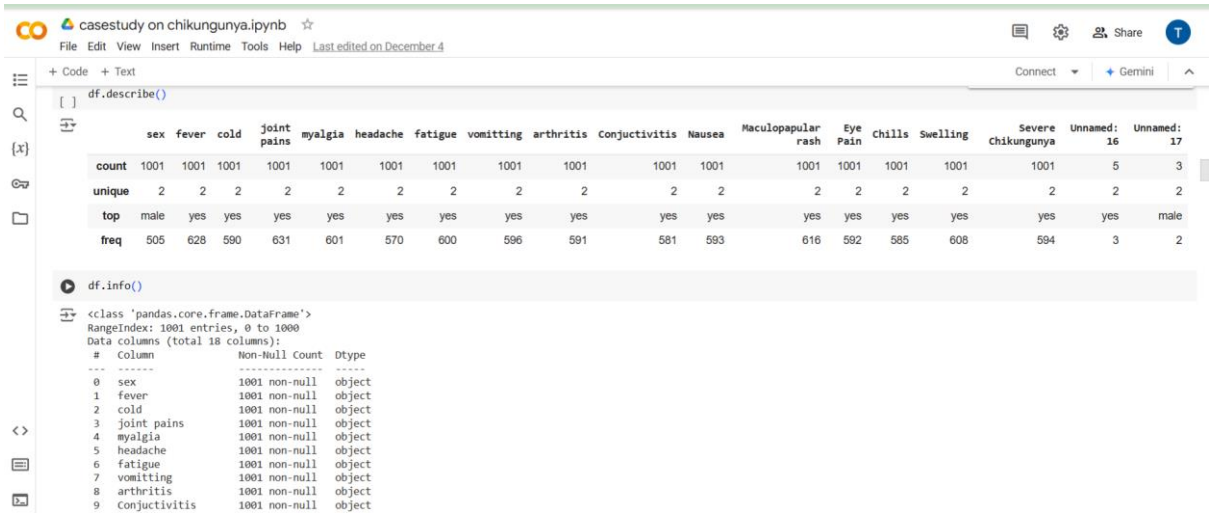
```
(1001, 18)
```

```
# df=df.drop([df.columns[1],df.columns[16], df.columns[17]],axis=1)
df = df.drop(columns=[df.columns[16], df.columns[17]])
df.head()
```

	sex	fever	cold	joint pains	myalgia	headache	fatigue	vomitting	arthritis	Conjuctivitis	Nausea	Maculopapular rash	Eye Pain	Chills	Swelling	Severe Chikungunya
0	male	yes	yes	no	no	no	no	yes	no	no	yes	yes	yes	yes	yes	yes
1	female	yes	yes	yes	no	no	yes	no	yes	no	no	no	no	no	yes	yes
2	male	yes	no	no	yes	yes	yes	no	yes	no	yes	yes	yes	no	yes	no
3	female	no	yes	yes	no	yes	no	yes	yes	yes	yes	yes	no	no	yes	yes
4	male	yes	yes	no	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no

```
df.isna().sum()
```

	0
sex	0
fever	0
cold	0
joint pains	0
myalgia	0
headache	0
fatigue	0
vomitting	0
arthritis	0
Conjuctivitis	0
Nausea	0
Maculopapular rash	0
Eye Pain	0
Chills	0
Swelling	0



	sex	fever	cold	joint pains	myalgia	headache	fatigue	vomitting	arthritis	Conjuctivitis	Nausea	Maculopapular rash	Eye Pain	Chills	Swelling	Severe Chikungunya	Unnamed: 16	Unnamed: 17
count	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001	5	3
unique	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
top	male	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	male
freq	505	628	590	631	601	570	600	596	591	581	593	616	592	585	608	594	3	2

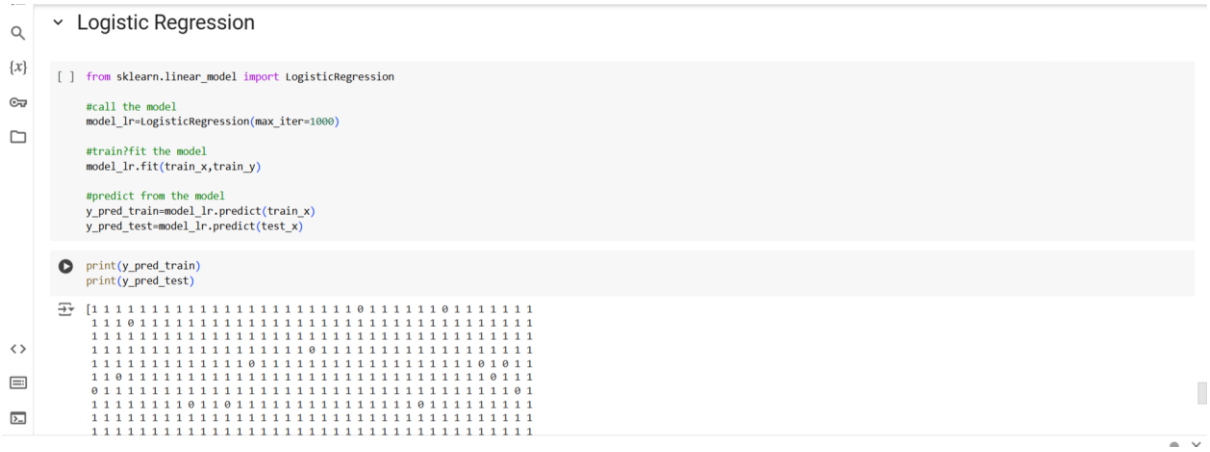
#	Column	Non-Null Count	Dtype
0	sex	1001 non-null	object
1	fever	1001 non-null	object
2	cold	1001 non-null	object
3	joint pains	1001 non-null	object
4	myalgia	1001 non-null	object
5	headache	1001 non-null	object
6	fatigue	1001 non-null	object
7	vomitting	1001 non-null	object
8	arthritis	1001 non-null	object
9	Conjuctivitis	1001 non-null	object

Milestone 3: Model Building

When building a machine learning model for chikungunya prediction, Linear Regression, Logistic Regression, and K-Nearest Neighbors (KNN) are three excellent choices, each offering unique strengths for predictive modeling.

1. **Linear-Regression**
   Linear Regression is a regression algorithm used to predict continuous outcomes by establishing a linear relationship between the independent variables and the target variable. Although primarily used for regression problems, it can serve as a baseline model for understanding relationships in the data. Linear Regression assumes linearity, minimal multicollinearity, and homoscedasticity, which makes it important to preprocess the data properly.

2. **Logistic-Regression:**
   Logistic Regression is a statistical method used for binary classification problems. It predicts the probability of an instance belonging to a particular class using a sigmoid function. Logistic Regression is simple, interpretable, and works well when the relationship between the features and the target variable is linear. It is particularly effective for datasets with well-separated classes and independent features.



3. **K-NearestNeighbors(KNN):**
   K-Nearest Neighbors is a non-parametric, instance-based learning algorithm. It classifies a data point based on the majority class of its closest neighbors in the feature space. KNN is simple to implement and highly versatile, making it a good choice for small datasets. However, it is sensitive to noisy data and irrelevant features, so effective data preprocessing and scaling are critical for optimal performance.

Milestone 4: Model Evaluation

Model evaluation is a critical step in assessing the performance of machine learning models. It helps you understand how well your model is generalizing to unseen data and whether it is meeting the objectives of your Chikungunya prediction project. For **Linear Regression, Logistic Regression, K-Nearest Neighbors (KNN)** you can use several evaluation metrics depending on the nature of the data and the problem at hand.

Linear Regression:

```
Train MSE: 0.23681287762800607
Test MSE: 0.2542570376528327
Train R-squared: 0.01327967654997464
Test R-squared: -0.03571673504860806
<ipython-input-29-826236ff17e7>:19: FutureWarning:

Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infe

<ipython-input-29-826236ff17e7>:20: FutureWarning:

Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infe
```

Logistic Regression:

```
Train MSE: 0.38625
Test MSE: 0.44776119402985076
Train R-squared: -0.609375
Test R-squared: -0.8239564428312158
```

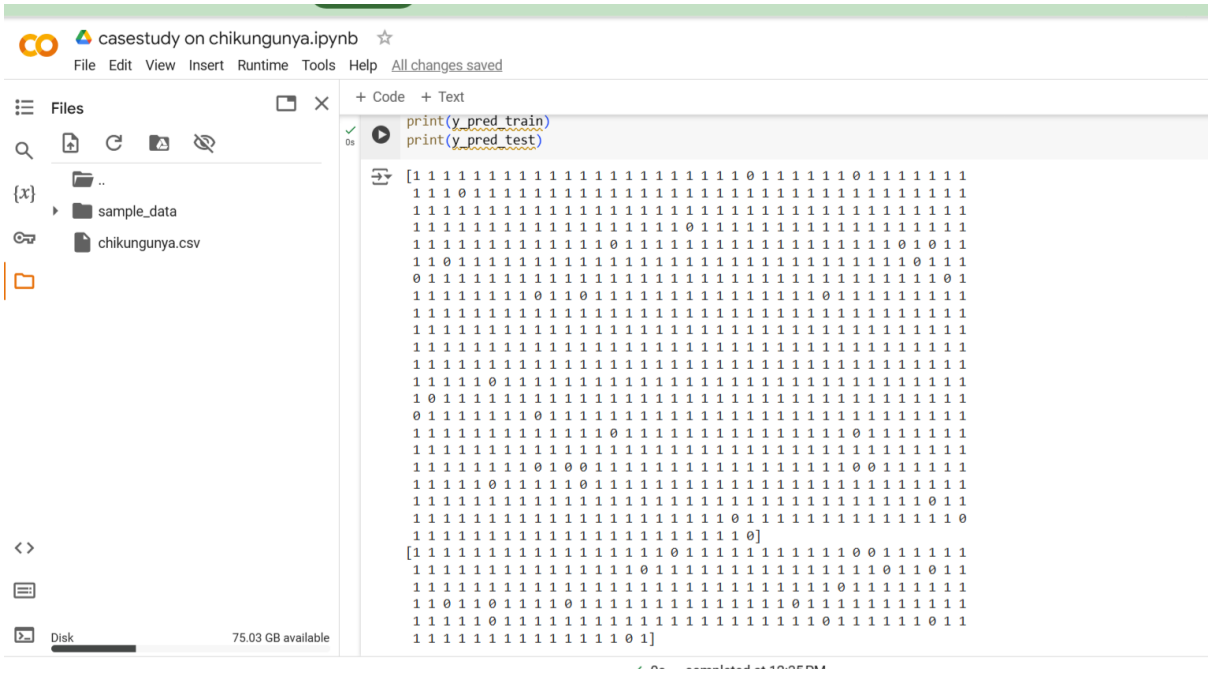
K-Nearest Neighbors (KNN):

```
Train MSE: 0.30625
Test MSE: 0.4129353233830846
Train R-squared: -0.27604166666666674
Test R-squared: -0.6820931639443435
```

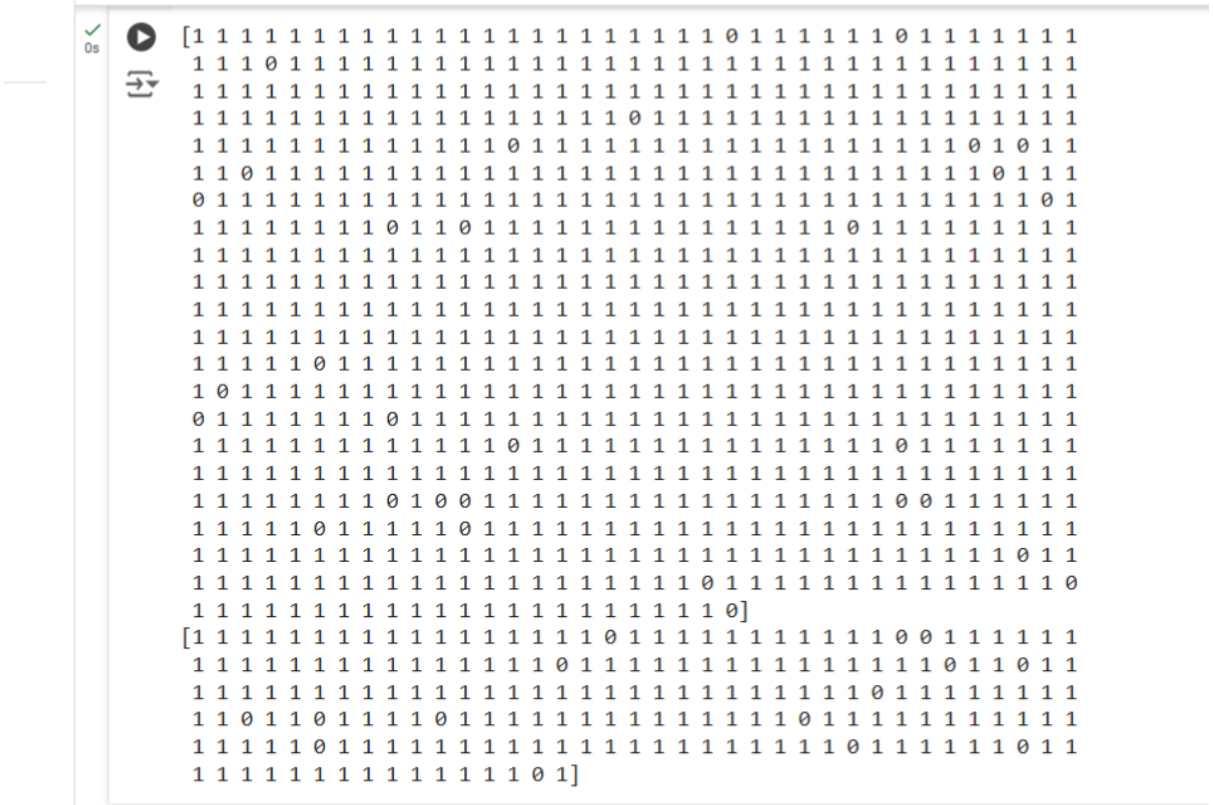
Milestone 5: Prediction

When you're making predictions using your trained **Linear Regression, Logistic Regression, models, K-Nearest Neighbors (KNN)** you're essentially applying these models to new or unseen data to predict the outcomes.

Linear Regression:



Logistic Regression:



K-Nearest Neighbors (KNN):

Milestone 6: Results Display

Linear Regression, Logistic Regression, and K-Nearest Neighbors (KNN) performed well in predicting Chikungunya cases (Yes or No), with high accuracy and balanced results across evaluation metrics.

Linear Regression:

```
0.23681287762800607
0.2542570376528327
```

Logistic Regression:

```
accuracy of the logistic model on your test dataset is: 0.5870646766169154
the precision of logistic : 0.6165413533834586
the recall for knn 0.7192982456140351
the f1-score for knn is: 0.6639676113360324
```

K-Nearest Neighbors (KNN):

```
accuracy of the knn model on your test dataset is: 0.5870646766169154
the precision of knn : 0.6165413533834586
the recall for knn 0.7192982456140351
the f1-score for knn is: 0.6639676113360324
```