

# Anemia Predictive Analysis

APPLICATION OF MACHINE LEARNING IN BIOLOGICAL SYSTEMS (ES60011)

By

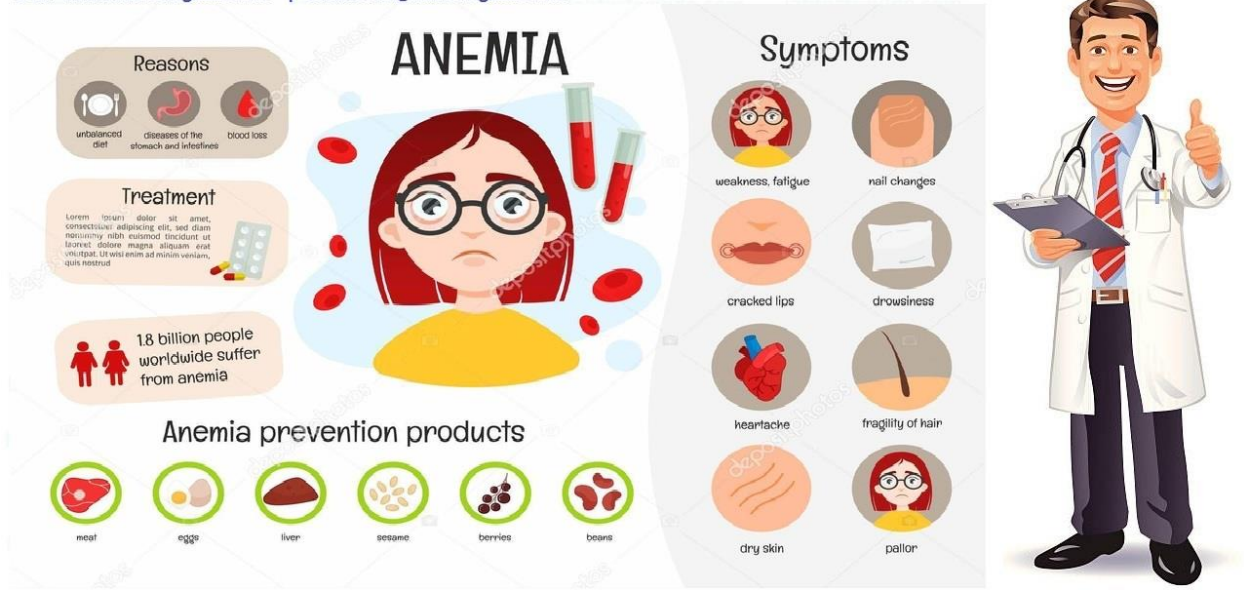
Aswin AK (19BT30006)  
Krishna Dev R (22MM62R01)  
Dr. Rachana (21MM60004)

**Indian Institute of Technology Kharagpur**  
**Autumn Semester, 2022**  
**October 16, 2022**

## About Anemia

# Anemia Symptoms, Causes, Diagnose and more

web: [www.labtestsguide.com](http://www.labtestsguide.com) | Email: [info@labtestsguide.com](mailto:info@labtestsguide.com)



**Anaemia** is a condition in which the number of red blood cells or the haemoglobin concentration within them is lower than normal. Haemoglobin is needed to carry oxygen and if you have too few or abnormal red blood cells, or not enough haemoglobin, there will be a decreased capacity of the blood to carry oxygen to the body's tissues. This results in symptoms such as fatigue, weakness, dizziness and shortness of breath, among others. The optimal haemoglobin concentration needed to meet physiologic needs varies by age, sex, elevation of residence, smoking habits and pregnancy status. The most common causes of anaemia include nutritional deficiencies, particularly iron deficiency, though deficiencies in folate, vitamins B12 and A are also important causes; haemoglobinopathies; and infectious diseases, such as malaria, tuberculosis, HIV and parasitic infections.

## **About data**

Anemia dataset containing attributes Gender, Hemoglobin, MCHC, MCV, MCH, and Results. This dataset predicts if a patient is likely to suffer from anemia. Machine learning binary classifier algorithm to be used.

Gender: 0 - male, 1 - female

Hemoglobin: Hemoglobin is a protein in your red blood cells that carries oxygen to your body's organs and tissues and transports carbon dioxide from your organs and tissues back to your lungs

MCH: MCH is short for "mean corpuscular hemoglobin." It's the average amount in each red blood cell of a protein called hemoglobin, which carries oxygen around your body.

MCHC: MCHC stands for mean corpuscular hemoglobin concentration. It's a measure of the average concentration of hemoglobin inside a single red blood cell.

MCV: MCV stands for mean corpuscular volume. An MCV blood test measures the average size of your red blood cells.

Results: 0- not anemic, 1-anemic.

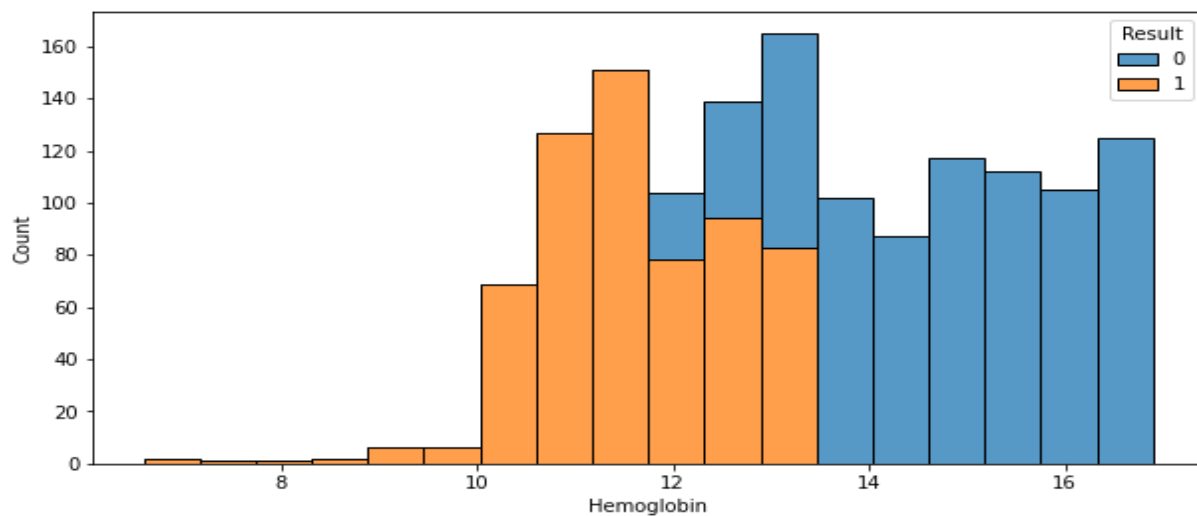
## 2. Imports :

**Pandas**, **Numpy**, and **Matplotlib** libraries were imported for data manipulation.

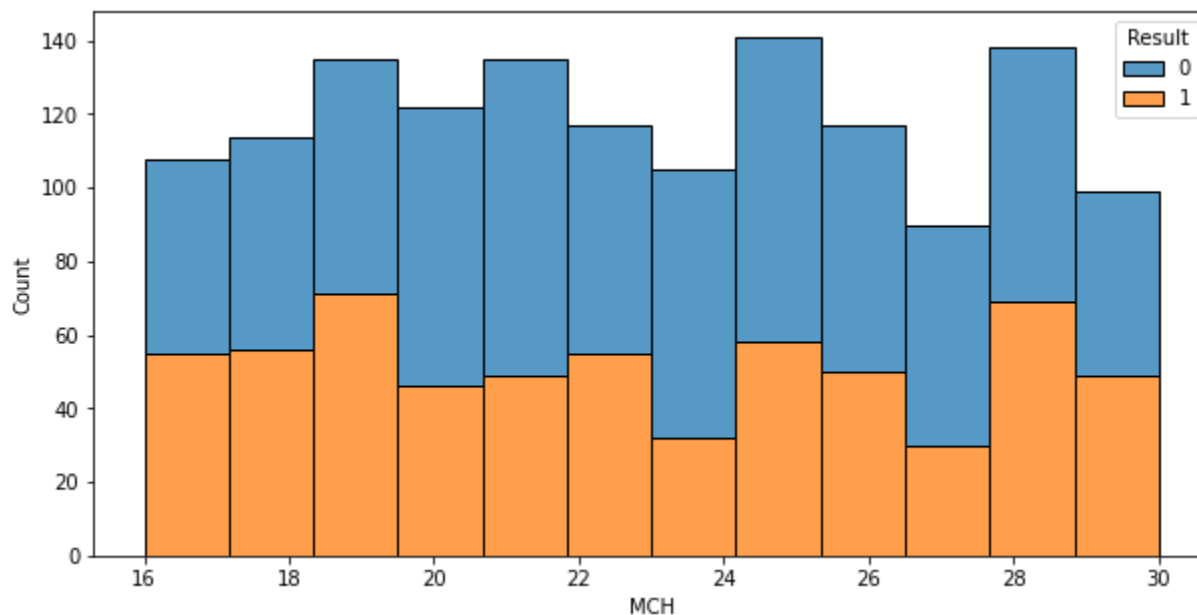
## 3. Analysis :

using the **Seaborn Library** to analyze the data and create a histogram and count plot

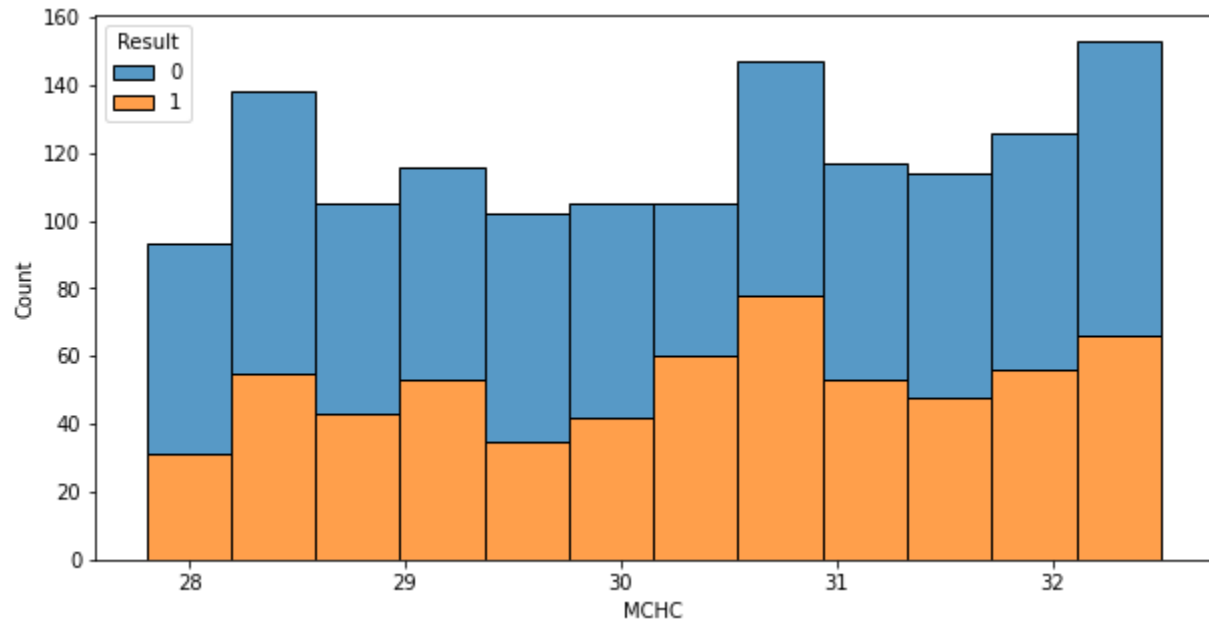
### Hemoglobin Histogram Plot:



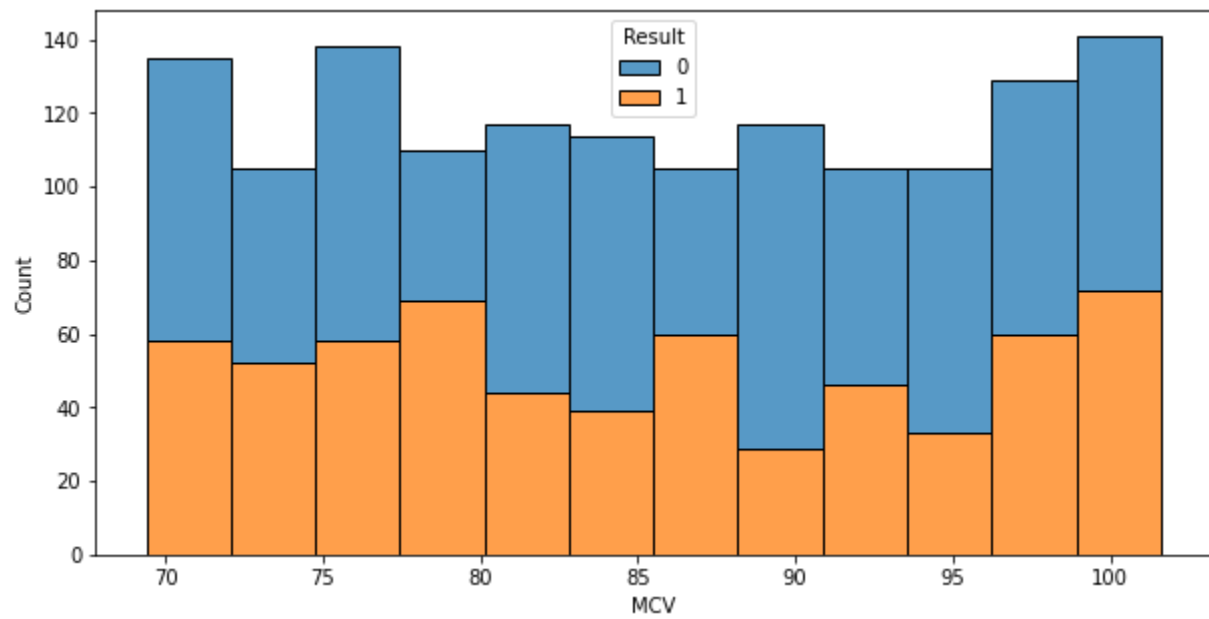
### MCH Histogram Plot:



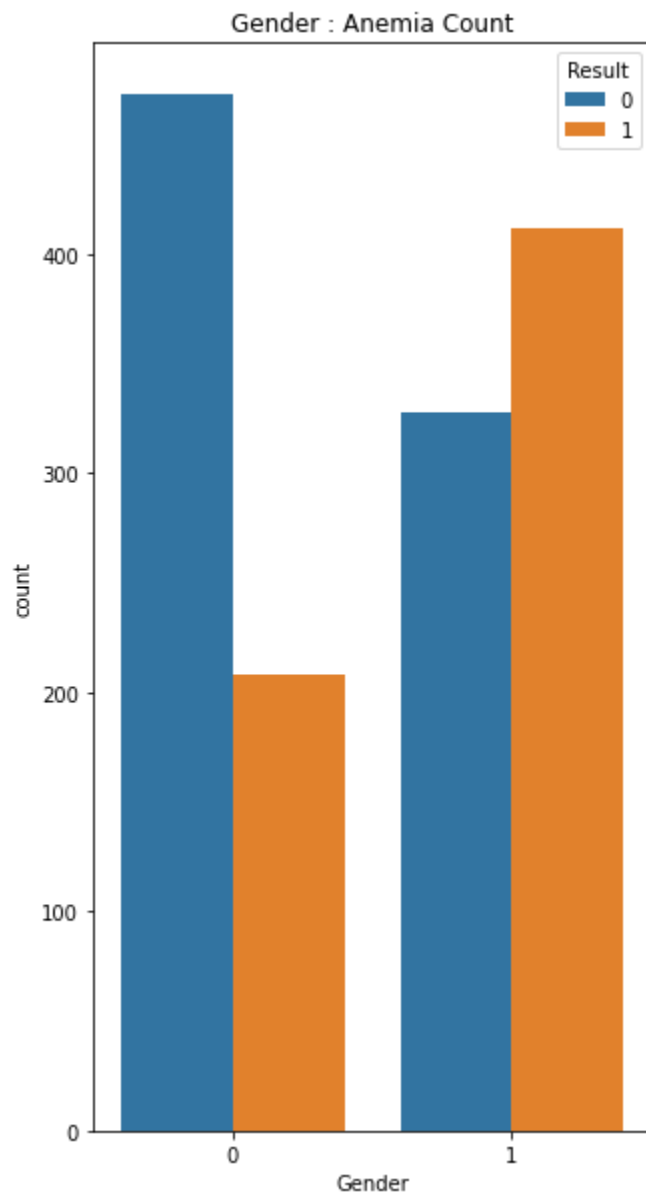
**MCHS Histogram Plot:**



**MCV Histogram Plot:**



## Anemia Count:



### 4. Prediction with Models:

We used **Logistic Regression**, **KNN Classification**, **Random Forest Classification**, and **Support Vector Machine** algorithms for prediction.

Below given is the accuracy of the different models we got:

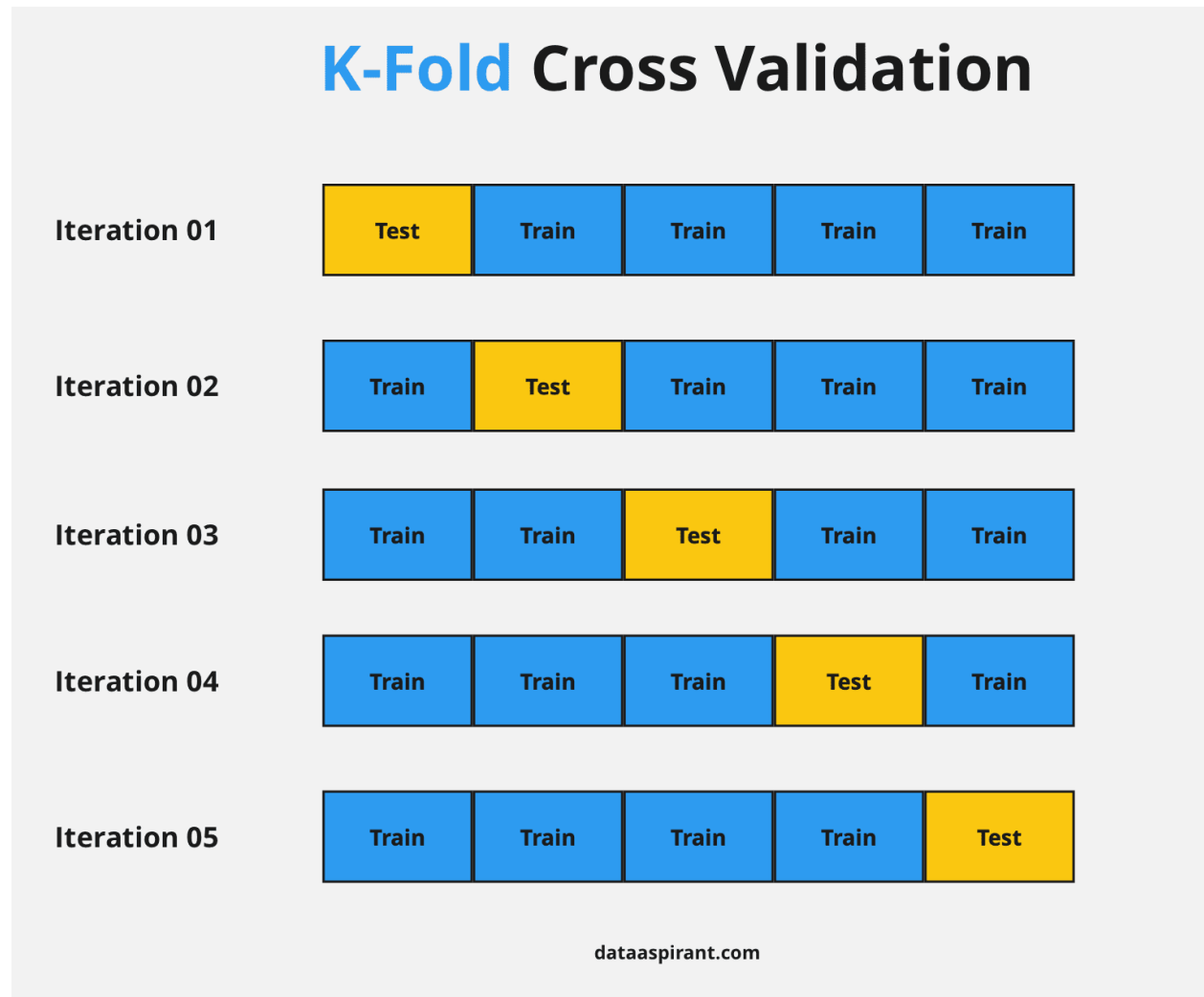
**LogisticRegression()** accuracy is : 0.99

**RandomForestClassifier()** accuracy is : 1.0

**KNeighborsClassifier()** accuracy is : 0.96

**SVC(random\_state=0)** accuracy is : 0.99

## 5. K fold cross Validation score:



various algorithms' **k cross-validation** scores are:

---

```
LogisticRegression() K fold cross validation score-
Accuracy: 98.85 %
Standard Deviation: 0.57 %

RandomForestClassifier(criterion='entropy', n_estimators=10) K fold cross validation score-
Accuracy: 100.00 %
Standard Deviation: 0.00 %

KNeighborsClassifier() K fold cross validation score-
Accuracy: 95.42 %
Standard Deviation: 1.48 %

SVC(random_state=0) K fold cross validation score-
Accuracy: 99.12 %
Standard Deviation: 0.56 %
```

## Algorithm

The **Random Forest Classifier** Algorithm was chosen to predict anemia since it had the greatest results.

### **Random Forest Classifier:-**

**Random forest** is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

#### **Steps involved in random forest algorithm:**

**Step 1:** In Random forest n number of random records are taken from the data set having k number of records.

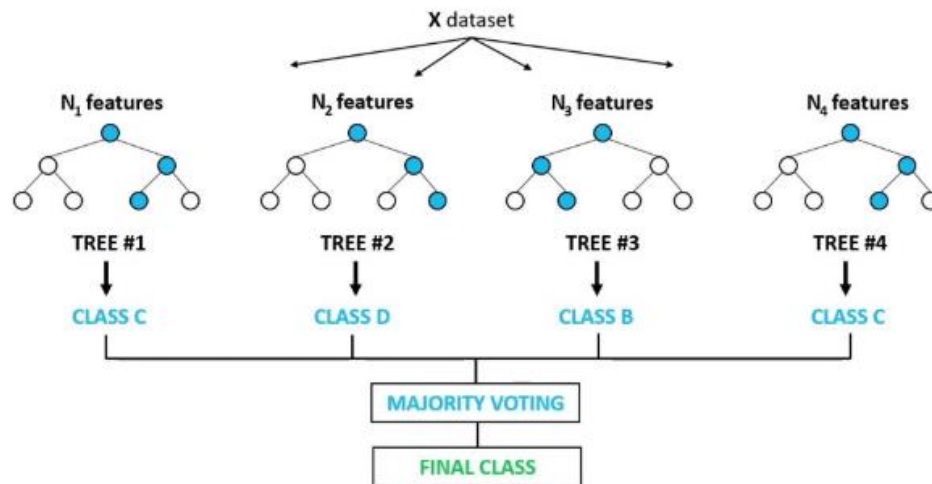
**Step 2:** Individual decision trees are constructed for each sample.

**Step 3:** Each decision tree will generate an output.

**Step 4:** Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.



# Random Forest Classifier



## Result:

Using a random forest classifier, an anemia predictor was created that, when a user enters his MCH, MHCH, and MCV, determines if he has anemia or not.

```
[ ] def anemia_predictor(Gender,Hemoglobin, MCH, MHCH,MCV ):
    a = r.predict(sc.transform([[Gender,Hemoglobin, MCH, MHCH,MCV]]))
    if a == 0:
        print("No anemia.")
    else:
        print("Anemia present.")
```

```
▶ Gender = input("Gender:")
  Hemoglobin = input("Hemoglobin:")
  MCH = input("MCH:")
  MHCH = input("MHCH:")
  MCV = input("MCV:")
  anemia_predictor(Gender,Hemoglobin, MCH, MHCH,MCV )
```

```
↳ Gender:0
  Hemoglobin:2
  MCH:2
  MHCH:2
  MCV:2
  Anemia present
```

