**Final Term Project**

**Introduction to Data Science**

**Topic: Cervical Dataset**

**Tonoy Chandra Sarker**

**ID: 20-43804-2**

**Section: C**

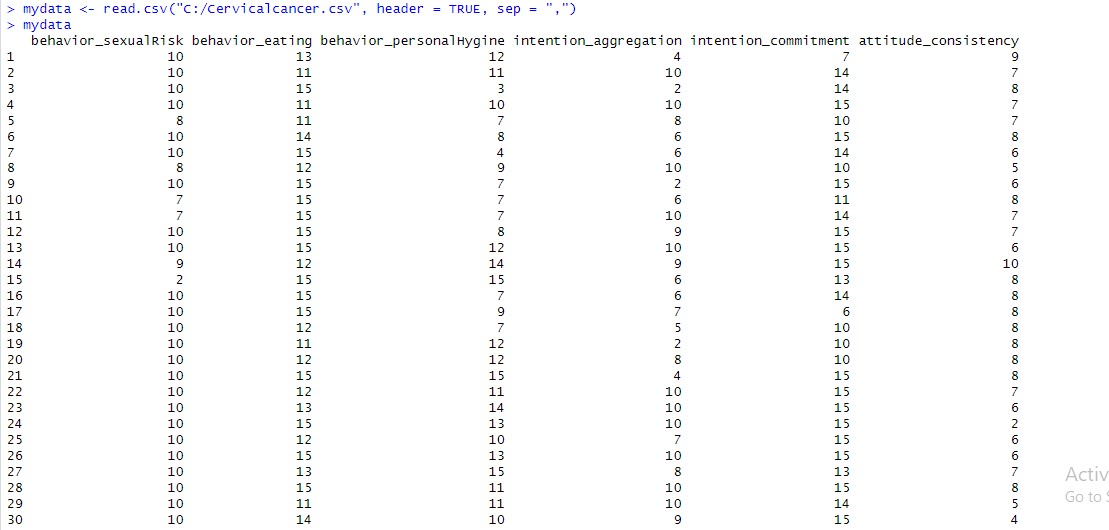
**Dataset Description:**

Cancer is a condition when the body's cells grow out of control. Even if cancer later spreads to other bodily areas, it is always given the name of the portion of the body where it first manifests. Cervical cancer is the name given to cancer that first appears there. The cells of the cervix are where cervical cancer begins. The cervix is the uterus' (the womb's) lower, thin end. The uterus and vagina, or birth canal, are joined by the cervix. In most cases, cervical cancer grows gradually over time. Before cancer develops in the cervix, the cervical tissue undergoes a process called dysplasia, during which aberrant cells start to emerge in the tissue. Over time, if not destroyed or removed, the abnormal cells may become cancer cells and start to grow and spread more deeply into the cervix and to surrounding areas.Anyone with a cervix is at risk for cervical cancer. It occurs most often in people over age 30. Long-lasting infection with certain types of human papillomavirus (HPV) is the main cause of cervical cancer. HPV is a common virus that is passed from one person to another during sex. At least half of sexually active people will have HPV at some point in their lives, but few women will get cervical cancer. Screening tests and the HPV vaccine can help prevent cervical cancer. When cervical cancer is found early, it is highly treatable and associated with long survival and good quality of life.

**Import the data set as csv and print the data set:**



**Output:**

****

**Description:**

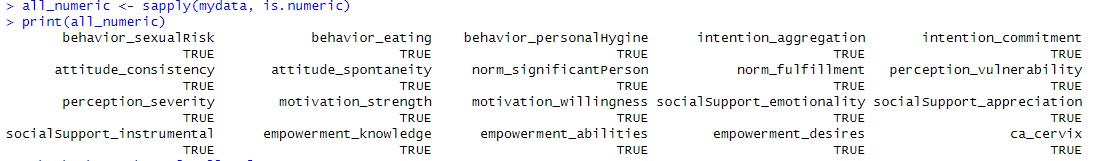
Here is the code of import the cervical dataset as csv file. It is the output of the dataset which is imported in RStudio.

# **Check all data is Numeric or Not:**

**Code :**

****

**Output:**



**Description :**

By using this code we can check if all the data is numeric or not. We can see in that all the data is numeric.

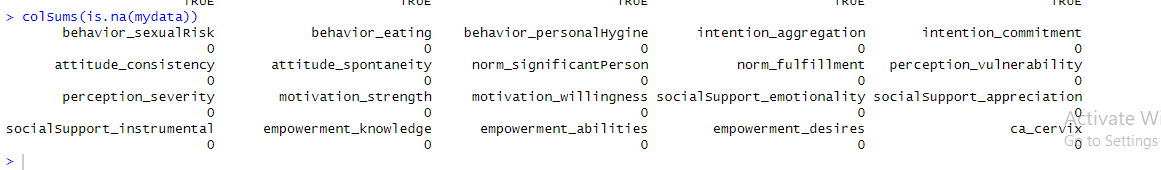
**Check Missing Values:**

**Code:**

****

.

**Output:**



**Description:**

In this code we can check the missing value in every column of dataset. No missing value is available in this dataset.

**Normalization**

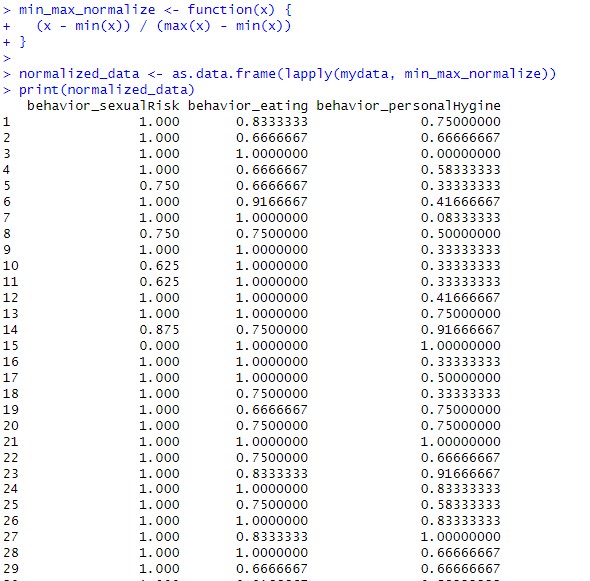
**Code:**

**A computer code with black text

Description automatically generated**

.

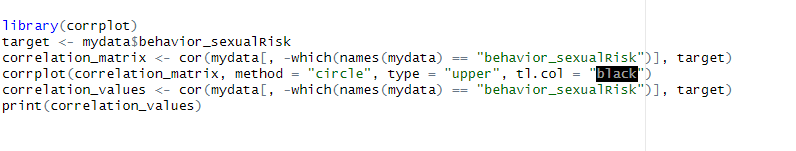
**Output:**



**Description :** Defines a function for min-max normalization, applies it to the entire dataset, and prints the normalized data. This method scales the features to a specific range, often between 0 and 1 we can see in the output.

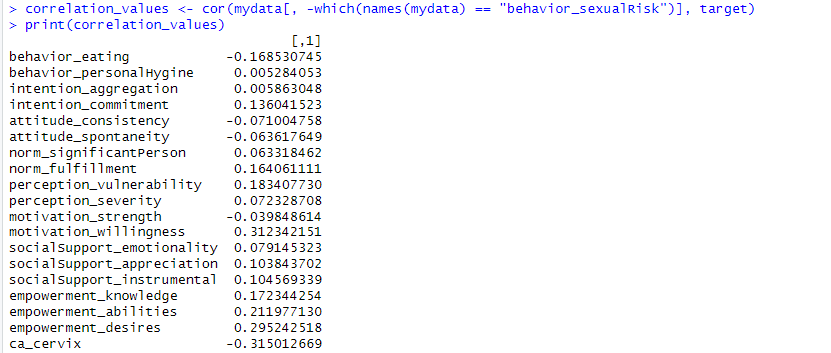
**Correlation for 2 var:**

**Code:**

****

.

**Output:**



**Description :** Correlation between 2 variable. Here target attribute is behavior\_sexualRisk. We can also see the correlation value of all variables against the target attribute and a graph of correlation behavior\_eating against the target attribute.

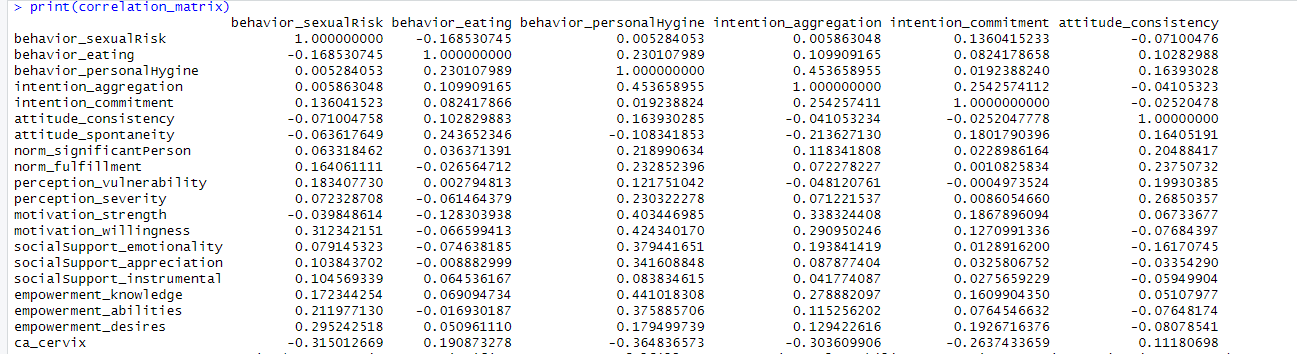
**Correlation for all var:**

**Code:**

. A black and white text

Description automatically generated

**Output:**



**Description :** Correlation matrix for all variable.

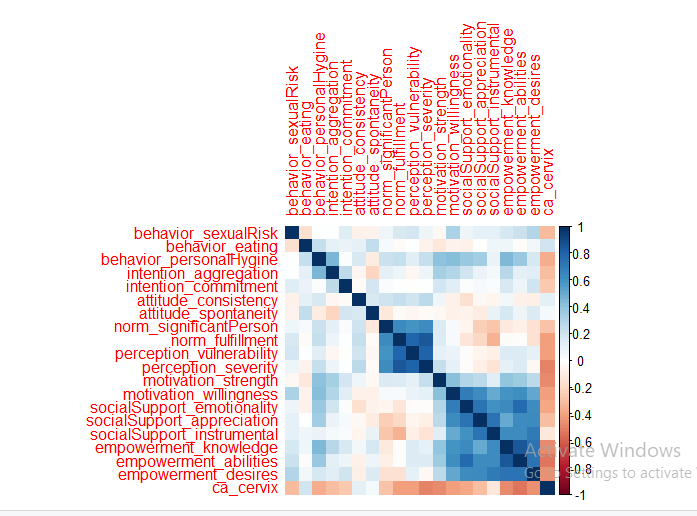
**Correlation using a library:**

**Code:**

**A black text on a white background

Description automatically generated**

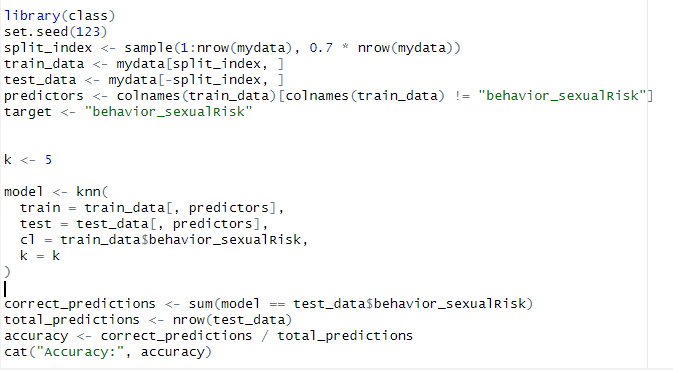
**Output:**

****

**Description :** Uses corrplot library to visualize the correlation matrix with color-coded cells.

**Built the KNN model**

**Code:**

****

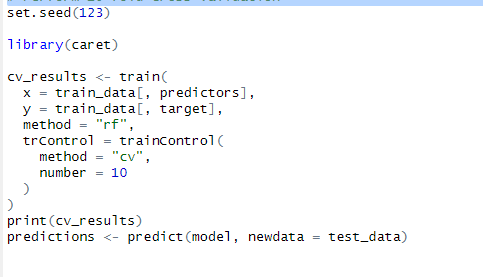
**Output:**



**Description :** Apply knn algorithm to build a model.Here 70% is train data and 30% is test data.Calculates and prints the accuracy of a trained model's predictions on a test dataset. The model's predictions are accurate for about 90.46% of the total cases in the dataset, or an accuracy of 0.9047619. In other words, only a very tiny subset of the data is subject to correct predictions from the model.

**10-fold cross validation**

**Code:**

****

**Output:**

**A screenshot of a computer program

Description automatically generated**

**Description:** Displays the results of cross-validation for the random forest model. The performance of a Random Forest model on your dataset is demonstrated by this output, which uses various values of mtry (the number of features taken into account at each split). You may get insight into the performance of the model using the metrics (RMSE, Rsquared, MAE), with lower RMSE and MAE values and higher Rsquared values indicating greater performance. You may assess and pick the ideal hyperparameter configuration for your Random Forest model using the values of these metrics for each model setup.

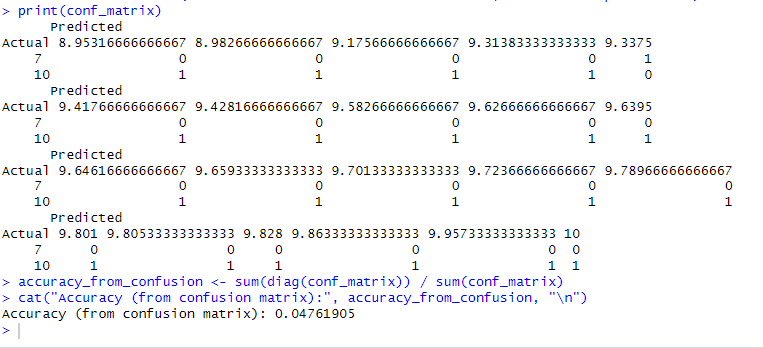
**Confusion Matrix**

**Code:**

**A screenshot of a computer code

Description automatically generated**

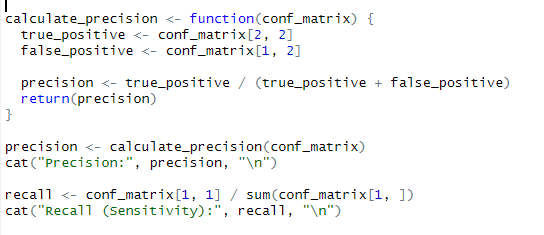
**Output:**



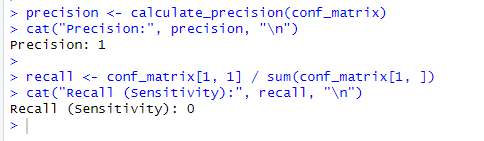
**Description:** Makes predictions on the test set, constructs a confusion matrix, and calculates and prints accuracy. With an accuracy of 0.04761905, the model correctly predicted 4.762% of the dataset's cases.

**Calculate Recall and Precision of Confusion Matrix**

**Code:**

****

**Output:**



**Description:** This code is for finding Recall and Precision for confusion matrix.