

Project Report

To perform predictive analyses on a dataset and develop model using the Python programming language

## Module Title: Data Analytics and Visualisation – PART II

## Module Code: B9IS107

## GROUP MEMBERS:

## SINGUPURAM NAVIN KUMAR | 10612366

## Module Leader: Dr Shazia A Afzal

TABLE OF CONTENTS

[1. CHOOSING AND EXPLORING DATA 4](#_Toc133523123)

[1.1 DATA SOURCE 4](#_Toc133523124)

[1.2 DATA EXPLORATION 5](#_Toc133523125)

[a. Understanding the data as part of the data-set of the mental health. 5](#_Toc133523126)

[b. On reading of the csv file the columns formed and are displayed as below 5](#_Toc133523127)

[c. On reading of the csv file the tail data displayed as below 5](#_Toc133523128)

[2. ANALYSING, CLEANING, AND PREPARING DATA 6](#_Toc133523129)

[2.1 ANALYSING 6](#_Toc133523130)

[2.2 CLEANING DATA 10](#_Toc133523131)

[2.3 PREPARING DATA 12](#_Toc133523132)

[2.4 WHAT IS FEATURE SCALING? 13](#_Toc133523133)

[2.5 SELECTING THE DEPENDENT AND INDEPENDENT VARIABLES 13](#_Toc133523134)

[3. MODEL BUILDING, TESTING, AND EVALUATION 15](#_Toc133523135)

[3.1 FEATURE SELECTION 15](#_Toc133523136)

[3.2 CORRELATION MATRIX WITH HEATMAP 15](#_Toc133523137)

[3.2.1 Treatment correlation matrix – Heap Map 17](#_Toc133523138)

[3.3 PAR PLOT 18](#_Toc133523139)

[3.4 RELATIONAL PLOTS 19](#_Toc133523140)

[3.5 DISTRIBUTION PLOT 20](#_Toc133523141)

[3.6 NESTED BARPLOT 21](#_Toc133523142)

[3.7 SPLITTING AND TUNING THE DATASET FOR MODEL 22](#_Toc133523143)

[3.7.1 REGRESSION TYPES 22](#_Toc133523144)

[LOGISTIC REGRESSION 22](#_Toc133523145)

[LINEAR REGRESSION 23](#_Toc133523146)

[4. MODELS EVALUATION AND PREDICTIONS 24](#_Toc133523147)

[4.1 CONFUSION MATRIX WITH LOGISTIC CLASSIFIER 24](#_Toc133523148)

[4.2 KNEIGHBORS CLASSIFIER WITH CROSS VALIDATION SCORE 27](#_Toc133523149)

[TUNING WITH CROSS VALIDATION SCORE 27](#_Toc133523150)

[4.3 DECISION TREE CLASSIFIER 29](#_Toc133523151)

[TRAINING THE DECISION TREE CLASSIFIER 29](#_Toc133523152)

[5. MODEL COMPARISON 31](#_Toc133523153)

[5.1 STOCHASTIC GRADIENT DESCENT (SGD) 31](#_Toc133523154)

[6. CONCLUSION 32](#_Toc133523155)

[7. REFERENCES 33](#_Toc133523156)

[8. APPENDIX 34](#_Toc133523157)

[9. COLAB LINK 45](#_Toc133523158)

TABLE OF FIGURES

[Figure 1: Reading of the Data 5](#_Toc133523413)

[Figure 2: Data Set Columns 5](#_Toc133523414)

[Figure 3: Tail Data 5](#_Toc133523415)

[Figure 4: Shape and Info of the Dataset 8](#_Toc133523416)

[Figure 5: Describe of the data 8](#_Toc133523417)

[Figure 6: nunique of the Dataset 9](#_Toc133523418)

[Figure 7: Missing Data 9](#_Toc133523419)

[Figure 8: Dealing with missing data 10](#_Toc133523420)

[Figure 9: Assigning default values 11](#_Toc133523421)

[Figure 10: Seggregating the gender groups 11](#_Toc133523422)

[Figure 11: Updating the default string for self, work and age with mean 12](#_Toc133523423)

[Figure 12: Encoding of the data 14](#_Toc133523424)

[Figure 13: Testing of the data after preparation 15](#_Toc133523425)

[Figure 14: Correlation matrix with heatmap 16](#_Toc133523426)

[Figure 15: Treatment Correlation Heat Matrix 17](#_Toc133523427)

[Figure 16: Pair Plot Matrix 1 18](#_Toc133523428)

[Figure 17: Pair Plot Matrix 2 18](#_Toc133523429)

[Figure 18: Relational Plot 19](#_Toc133523430)

[Figure 19: Distribution plot 20](#_Toc133523431)

[Figure 20: Distribution Plot on seperate treatments 21](#_Toc133523432)

[Figure 21: Nested Barplot 21](#_Toc133523433)

[Figure 22: Logistic Regression predicted values 23](#_Toc133523434)

[Figure 23: Linear regression of the xtest 24](#_Toc133523435)

[Figure 24: Confusion Accuracy Matrix 25](#_Toc133523436)

[Figure 25: Histogram of the prediction 26](#_Toc133523437)

[Figure 26: ROC Curve for the classifier 26](#_Toc133523438)

[Figure 27: Tuning the KNN 27](#_Toc133523439)

[Figure 28: KNN prediction with confusion matrix 27](#_Toc133523440)

[Figure 29: Histogram of the KNN classifier with predicted values 28](#_Toc133523441)

[Figure 30: ROC Curve for the KNN classifier 28](#_Toc133523442)

[Figure 31: Training of the Tree Classifier 29](#_Toc133523443)

[Figure 32: Decision Tree prediction with confusion matrix 30](#_Toc133523444)

[Figure 33: Histogram Predictions of the Tree Classifier 30](#_Toc133523445)

[Figure 34: ROC Curve for the Tree Classifier 31](#_Toc133523446)

[Figure 35: SGDC Classifier score 32](#_Toc133523447)

[Figure 36: SGDC prediction with Grid Search CV 32](#_Toc133523448)

# CHOOSING AND EXPLORING DATA

The survey on the mental health was selected to know whether a patient needs treatment or not. Also, to understand the predictive analysis which is a series of surveys conducted in more than 20 countries that assesses the prevalence and treatment of mental disorders.

## DATA SOURCE

**LINK FOR THE DATASET**

It is available on the [Kaggle Dataset Repository](https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey) website.

**DATASET INFORMATION**

This information comes from a 2014 poll that assesses attitudes about mental health as well as the prevalence of mental health issues in the IT workplace.

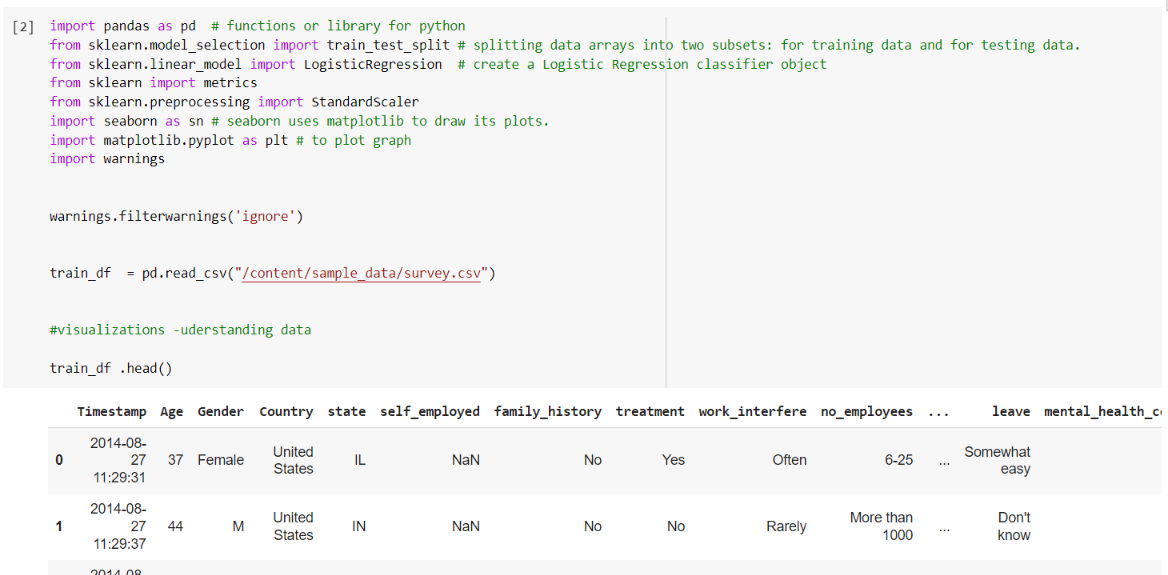
**REASONS FOR SELECTING DATBASE**

**TO UNDERSTAND:**

* In what ways does the prevalence of mental health disorders and societal attitudes towards mental health differ across various geographical regions?
* What are the most significant factors that can predict the occurrence of mental health disorders or specific attitudes towards mental health within a work environment?

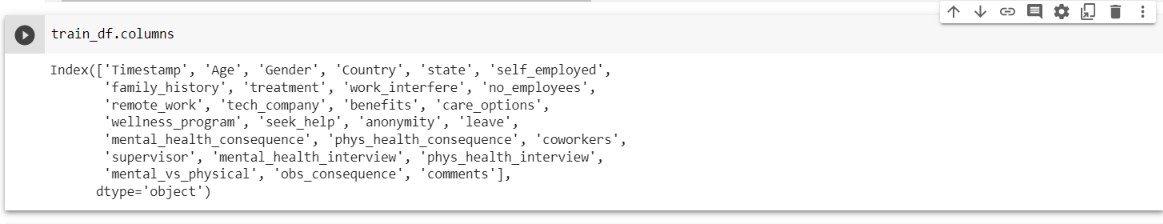
## DATA EXPLORATION

### Understanding the data as part of the data-set of the mental health.



**Figure 1: Reading of the Data**

### On reading of the csv file the columns formed and are displayed as below



**Figure 2: Data Set Columns**

### On reading of the csv file the tail data displayed as below

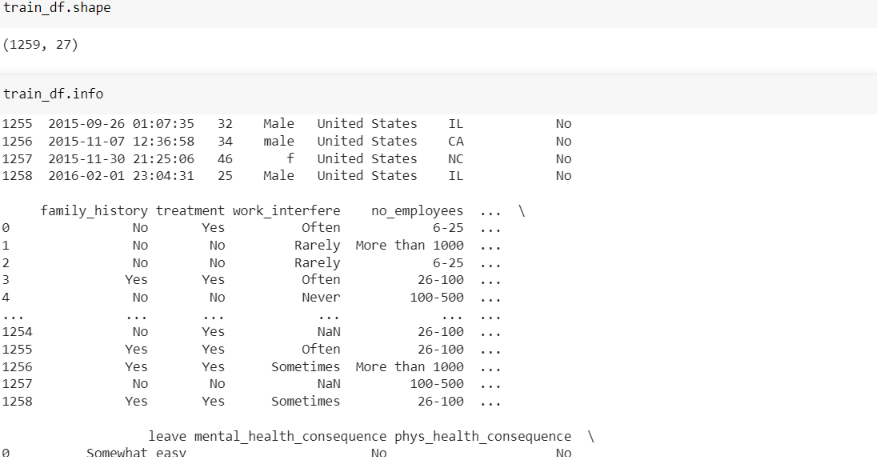


**Figure 3: Tail Data**

# ANALYSING, CLEANING, AND PREPARING DATA

## ANALYSING

* **Timestamp**
* **Age**
* **Gender**
* **Country**
* **state**: What is your current state or territory of residence, if you live in the United States?
* **self**\_**employed**: Are you self-employed?
* **family**\_**history**: Is there a history of mental illness in your family?
* **treatment**: Have you sought treatment for a mental health condition?
* **work**\_interfere: If you have a mental health condition, do you feel that it interferes with your work?
* **no**\_**employees**: How many employees does your company or organization have?
* **remote**\_work: Do you work remotely (outside of an office) at least 50% of the time?
* **tech\_company**: Is your employer primarily a tech company/organization?
* **benefits**: Does your employer provide mental health benefits?
* **care**\_options: Do you know the options for mental health care your employer provides?
* **wellness**\_**program**: Has your employer ever discussed mental health as part of an employee wellness program?
* **seek\_help**: Does your employer provide resources to learn more about mental health issues and how to seek help?
* **anonymity**: If you decide to use mental health or substance abuse treatment resources, is your anonymity protected?
* **leave**: How easy is it for you to take medical leave for a mental health condition?
* **mental\_health\_consequence**: Do you think that discussing a mental health issue with your employer would have negative consequences?
* **phys\_health\_consequence**: Do you think that discussing a physical health issue with your employer would have negative consequences?
* **coworkers**: Would you be willing to discuss a mental health issue with your coworkers?
* **supervisor**: Would you be willing to discuss a mental health issue with your direct supervisor(s)?
* **mental\_health\_interview**: Would you bring up a mental health issue during a job interview with a potential employer?
* **phys\_health\_interview**: Would you bring up a physical health issue during a job interview with a potential employer?
* **mental\_vs\_physical**: Do you believe that your employer takes mental health as seriously as physical health?
* **obs\_consequence**: Have you personally observed or heard of negative consequences for coworkers with mental health conditions in your workplace?
* **comments**: Are there any additional notes or comments you would like to provide?
  1. Analysing the data in a form of different ways like displayed below as shap and the info of the data- set chosen.



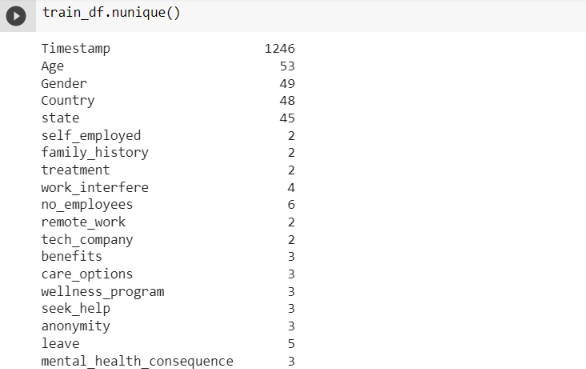
**Figure 4: Shape and Info of the Dataset**

* 1. Describing the data of the data- set chosen with the age and the count, mean and different parameters.



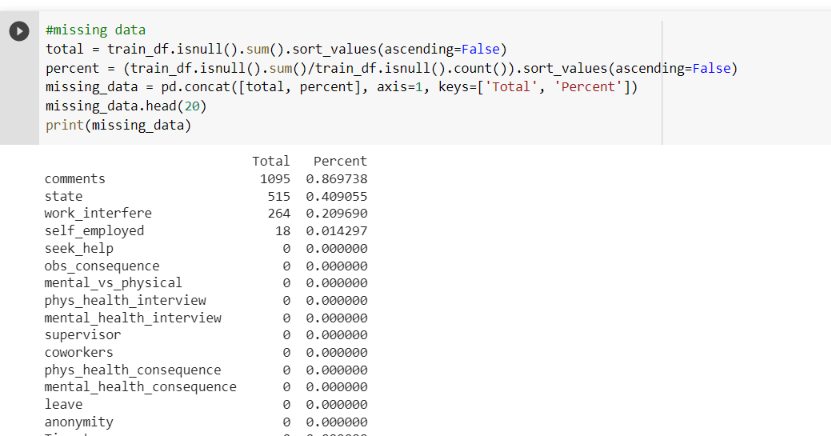
**Figure 5: Describe of the data**

* 1. NUnique the data of the data- set chosen with the age and the count, mean and different parameters.



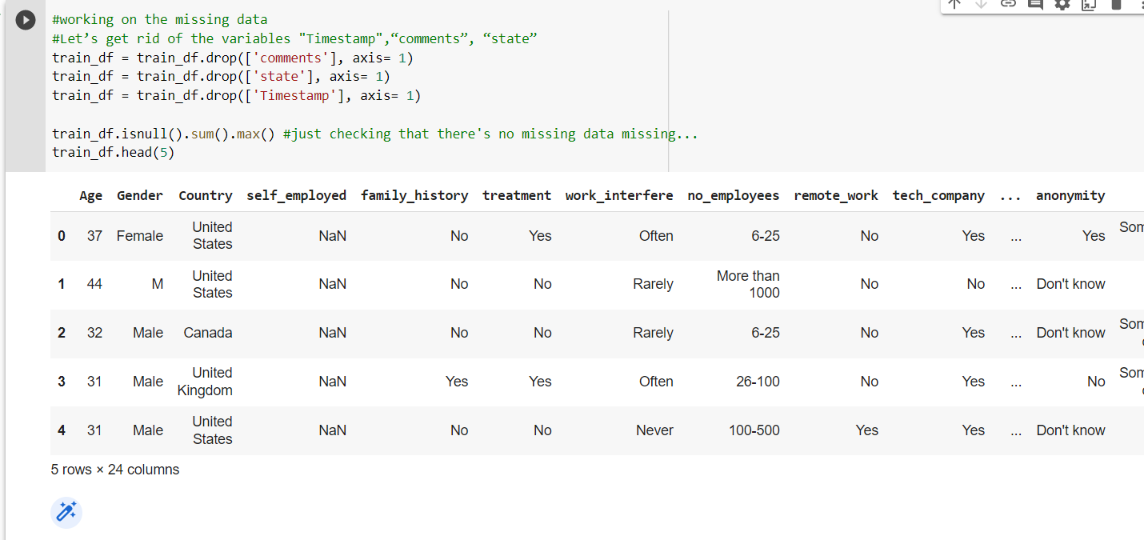
**Figure 6: nunique of the Dataset**

* 1. Analysing the missing data and checking the head by sorting it



**Figure 7: Missing Data**

* 1. Working on the missing data by dropping the three columns and checking further if there is any missing data and printing the head.



**Figure 8: Dealing with missing data**

## CLEANING DATA

* Eliminating unnecessary columns from a dataframe.

- Analyzing the dataset, computing the total of null values for each column, and eliminating any rows that have null values.

a. Cleaning the NaN values and assigning the default values for each column.



**Figure 9: Assigning default values**

b. Cleaning the gender values by lowering the case of the column elements, selecting the gender elements and segregating into gender groups and also removing the unwanted string from the values in the dataset.



**Figure 10: Seggregating the gender groups**

c. Updated the missing age with mean and check the age ranges and for the self employed and the self work interfere which updated with No and Don’t know respectively.



**Figure 11: Updating the default string for self, work and age with mean**

## PREPARING DATA

To double-check the column datatypes

We need to have input as numerical values in machine learning, thus we must ensure that all values are in digits format. We made certain that categorical data was transformed to numbers in accordance with the categories.

1. Data cleaning: This involves identifying and correcting errors in the dataset, such as missing values, outliers, or inconsistent formatting. It is important to ensure that the data is complete and accurate before using it for machine learning.

2. Feature selection: This involves choosing which variables to include in the machine learning model. It is important to select variables that are relevant to the research question and have been shown to be associated with mental health outcomes in previous research.

3. Feature engineering: This involves transforming the variables into a format that can be used in the machine learning model. This may include scaling the variables, encoding categorical variables, or creating new variables based on existing ones.

4. Data splitting: This involves dividing the dataset into training, validation, and testing sets. The training set is used to train the machine learning model, the validation set is used to tune the hyperparameters of the model, and the testing set is used to evaluate the performance of the final model.

## WHAT IS FEATURE SCALING?

The technique of standardizing or normalizing the range of features or variables in a dataset in order to make them comparable and understandable for machine learning algorithms is known as feature scaling.

When the characteristics in a dataset have multiple sizes or units, some algorithms may struggle to appropriately understand the data. Certain algorithms, such as K-nearest neighbors or Support Vector Machines, may be sensitive to feature scale, but others, such as gradient descent, may converge quicker when the features are scaled. As a result, it is frequently advised to scale the features before employing them in machine learning models..

## SELECTING THE DEPENDENT AND INDEPENDENT VARIABLES

**DEPENDENT VARIABLE:**

Mental health status: This could be measured using a standardized assessment tool such as the General Health Questionnaire (GHQ), the Depression Anxiety and Stress Scale (DASS), or the Patient Health Questionnaire (PHQ-9).

**Perception of mental health support**: This could be measured using a Likert scale or qualitative data.

**INDEPENDENT VARIABLES:**

**Demographic variables:** This could include age, gender, ethnicity, and education level.

**Employment-related variables:** This could include job role, length of time with current employer, number of working hours, and type of contract.

**Technology-related variables:** This could include frequency of computer use, type of technology used, and social media use.

**Workplace-related variables:** This could include workplace culture, work-life balance, and job satisfaction.

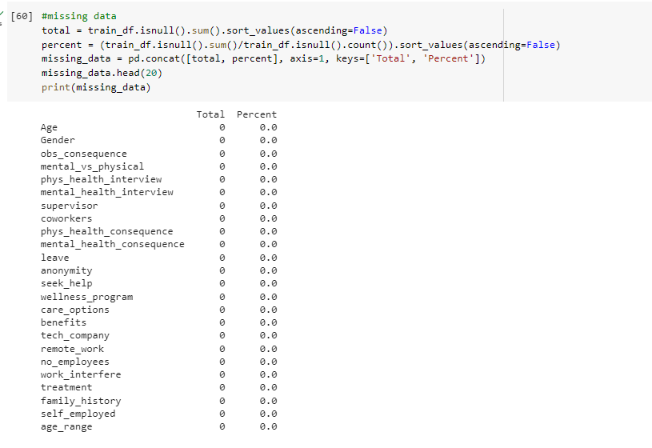
* + 1. Below shows the encoding of the data and getting rid of the country attribute





**Figure 12: Encoding of the data**

* + 1. Testing that there aren’t any missing data after the cleaning and updating the values



**Figure 13: Testing of the data after preparation**

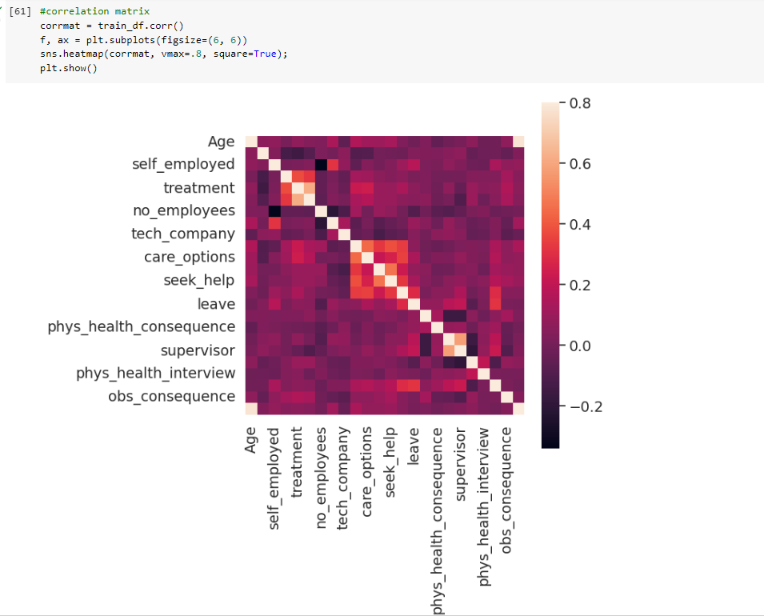
# MODEL BUILDING, TESTING, AND EVALUATION

## FEATURE SELECTION

While developing a predictive model, the feature selection procedure entails reducing the number of input variables. In some cases, limiting the amount of input variables may improve model performance while simultaneously cutting modeling processing costs.

## CORRELATION MATRIX WITH HEATMAP

A correlation heatmap is a type of graph that displays the strength of correlations between numerical variables. Correlation graphs are used to discover which variables are connected to one another and how strongly this relationship exists. Positive values represent a strong association, whilst negative values represent a weak relationship. The cell values show the strength of the link. Correlation heatmaps may be used to find potential correlations between variables and to assess the strength of these associations.



**Figure 14: Correlation matrix with heatmap**

## 3.2.1 Treatment correlation matrix – Heap Map



**Figure 15: Treatment Correlation Heat Matrix**

Correlation is a well-known statistic for measuring how similar two features are. If two characteristics are linearly related, their correlation coefficient is 1. There is no connection between two traits if they are uncorrelated. If the value is greater than the threshold, the feature is enabled.

- A correlation heatmap is an image of a correlation matrix that demonstrates the link between several variables.

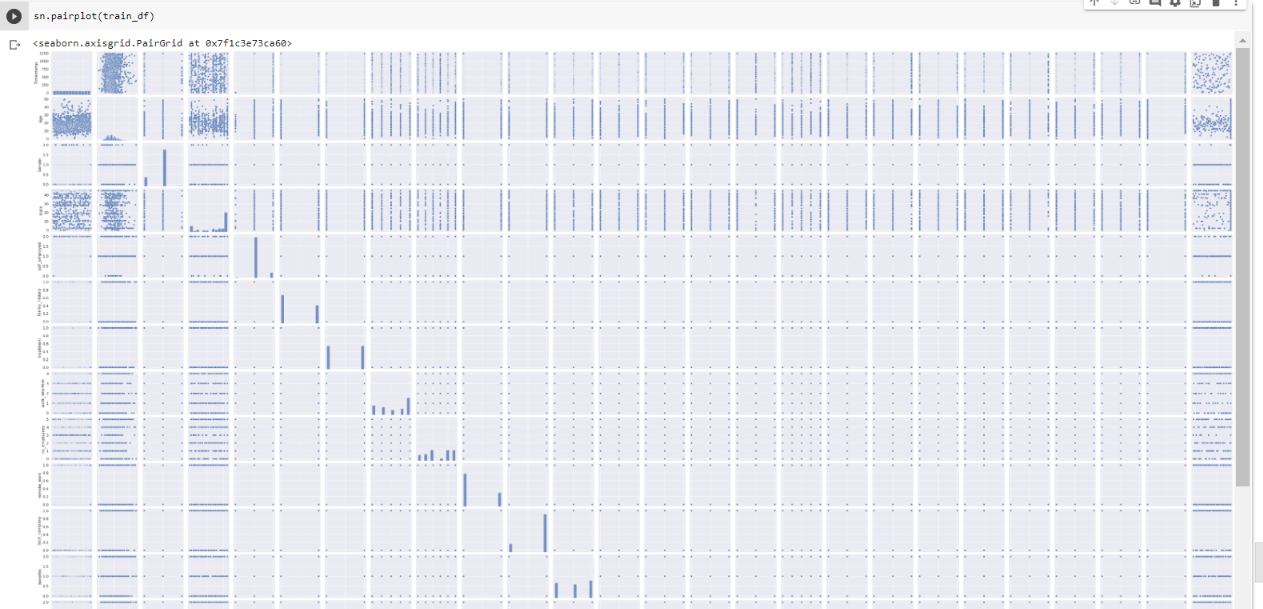
- Correlation can range between -1 and 1.

- A correlation between two random variables or bivariate data does not always suggest a causal relationship.

- A scatter plot of these two variables may also be used to determine the correlation between them.

## PAR PLOT

A "pairs plot" scatterplot matrix aids in understanding the pairwise relationship between distinct variables in a dataset.



**Figure 16: Pair Plot Matrix 1**



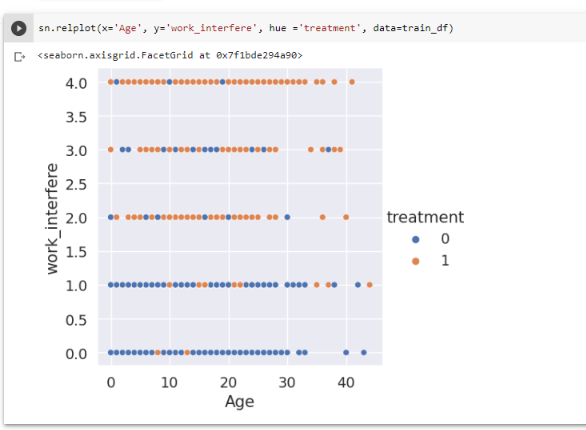
**Figure 17: Pair Plot Matrix 2**

Users can use the Seaborn Pairplot function to create an axis grid that distributes each numerical variable in the data over the X- and Y-axes in the form of columns and rows. We may use scatter plots to demonstrate pairwise correlations in addition to the distribution plot, which depicts the data distribution in the column diagonally.

We may plot several variable types on rows and columns, or we can use the pairplot() method to display a subset of variables.

## RELATIONAL PLOTS

Relational graphs are used to represent the statistical relationship between the data elements. Visualization is critical because it allows humans to detect trends and patterns in data. The process of discovering the correlations between and among variables in a dataset is known as statistical analysis.



**Figure 18: Relational Plot**

Here the relational plot uses weight parameter as X and Y parameter as height and we can see that BMI is plotted on the graph with different hues.

## DISTRIBUTION PLOT

A distribution plot, or Distplot, depicts the change in the data distribution. A seaborn plot depicts the overall distribution of continuous data variables.



**Figure 19: Distribution plot**

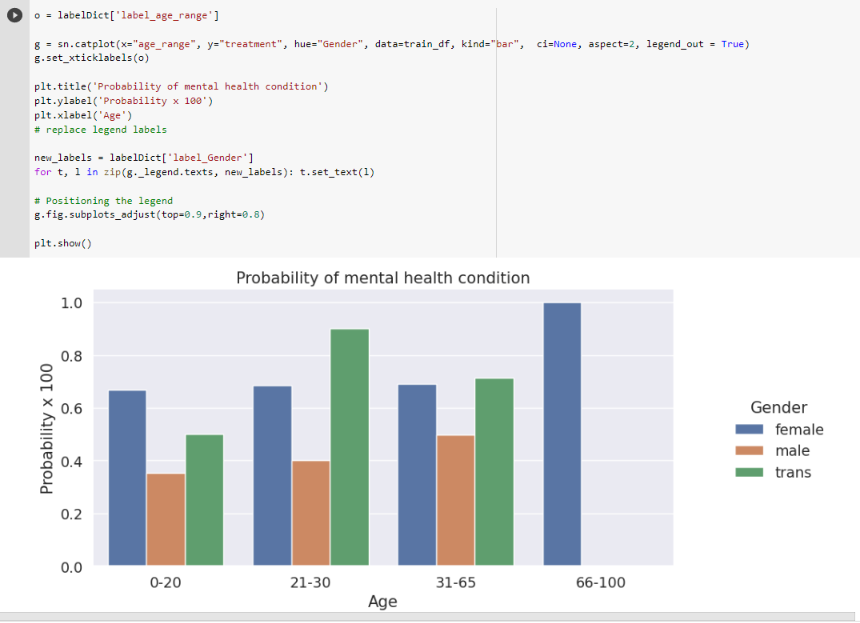
This scatter plot depicts the treatment values that are dispersed throughout the dataset. We may also see a trend in the distribution of treatment values between ages and treatment group.



**Figure 20: Distribution Plot on seperate treatments**

## NESTED BARPLOT

Below is the bar plot which depicts the age on the x axis and the probability on the y axis through the mental health condition probability



**Figure 21: Nested Barplot**

## SPLITTING AND TUNING THE DATASET FOR MODEL

Evaluating a Classification Model

This function will evalue:

* **Classification accuracy:**percentage of correct predictions
* **Null accuracy:** accuracy that could be achieved by always predicting the most frequent class
* **Percentage of ones**
* **Percentage of zeros**

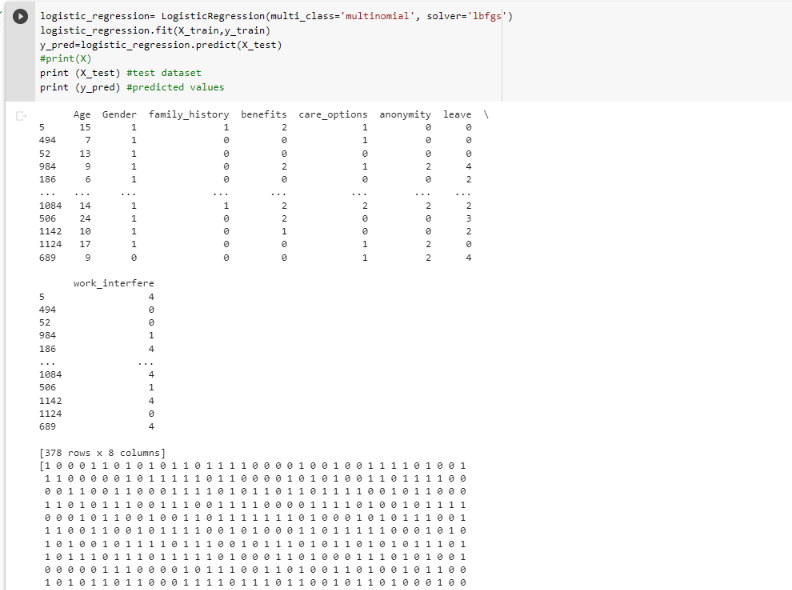
## 3.7.1 REGRESSION TYPES

Regression can be used to determine the statistical relationship between a dependent variable and one or more independent variables. The change in the independent variable is related to the change in the independent variable. This may be divided into two broad categories.

## LOGISTIC REGRESSION

When the dependent variable is discrete, logistic regression is employed as a regression analysis approach. For example, true or false, 0 or 1, and so on. As a result, the target variable can only have two values, and the relationship between the target variable and the independent variable is illustrated by a sigmoid curve.

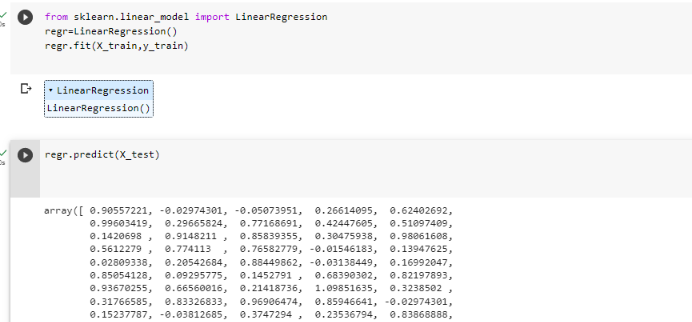
The logit function is used in logistic regression to quantify the relationship between the dependent and independent variables. The logistic regression equation is presented in the figure below.

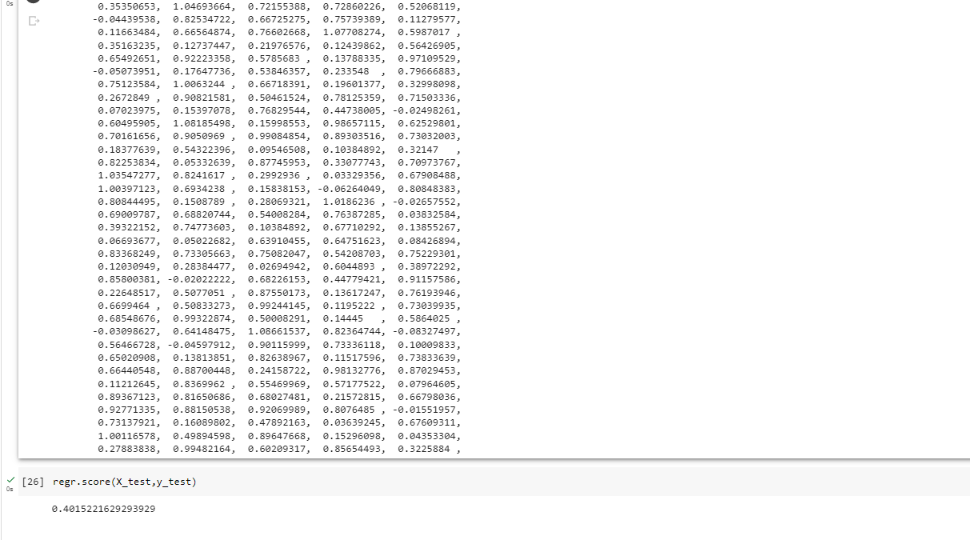


**Figure 22: Logistic Regression predicted values**

## LINEAR REGRESSION

Linear regression is one of the most fundamental types of regression in machine learning. The linear regression model is made up of a predictor variable and a dependant variable that are linearly connected to one another. When there are several independent variables in the data, linear regression is known as numerous linear regression models.





**Figure 23: Linear regression of the xtest**

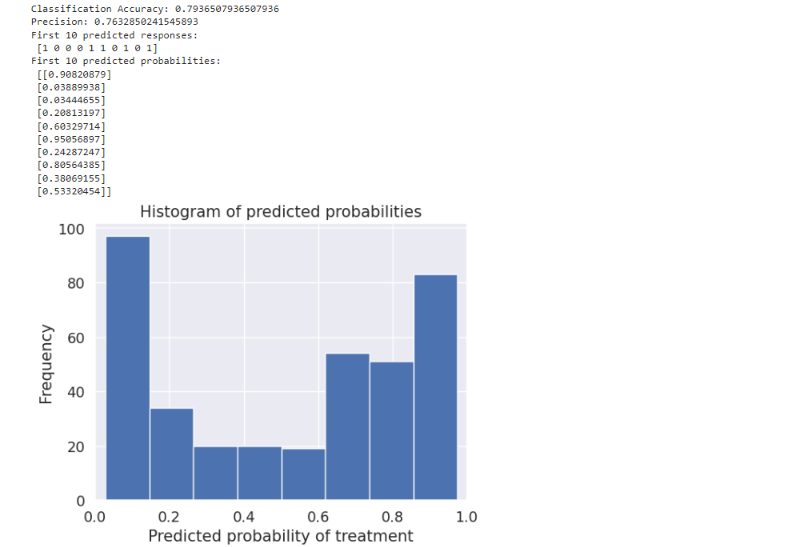
# MODELS EVALUATION AND PREDICTIONS

## CONFUSION MATRIX WITH LOGISTIC CLASSIFIER

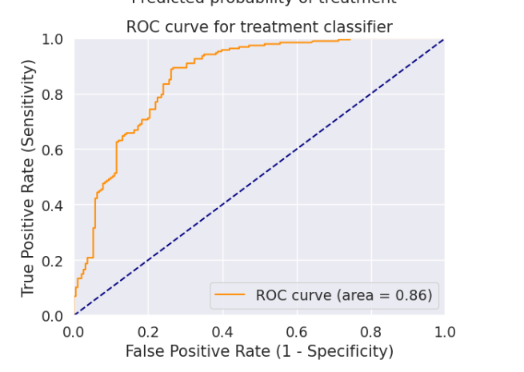
Confusion matrices are a popular tool for attempting to solve classification problems. It can perform both binary classification and multiclass classification problems.



**Figure 24: Confusion Accuracy Matrix**



**Figure 25: Histogram of the prediction**



**Figure 26: ROC Curve for the classifier**

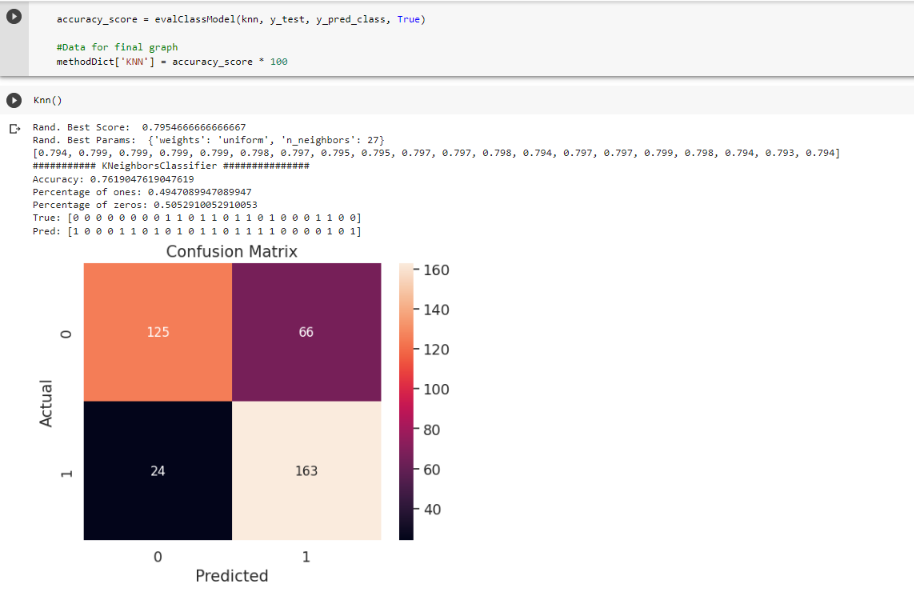
## KNEIGHBORS CLASSIFIER WITH CROSS VALIDATION SCORE

### TUNING WITH CROSS VALIDATION SCORE

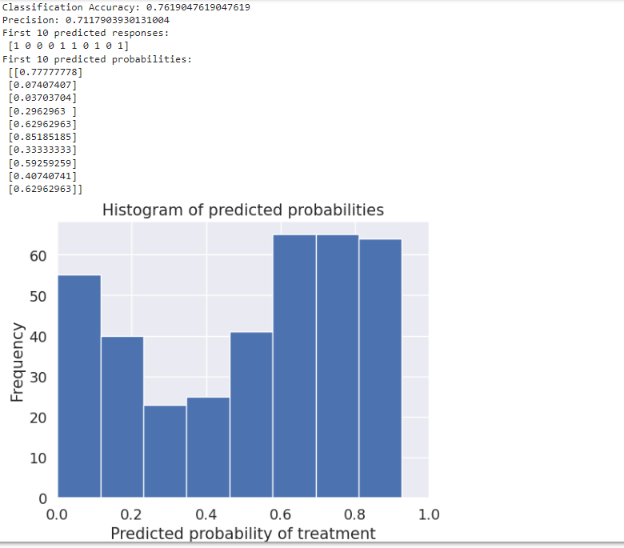
Tuning a model with cross-validation is a common practice in machine learning to improve the model's performance and avoid overfitting. Cross-validation is a technique where the data is split into multiple folds, and the model is trained and validated on each fold. This procedure aids in assessing the effectiveness of the model across various data subsets and delivers a more reliable estimate of its performance



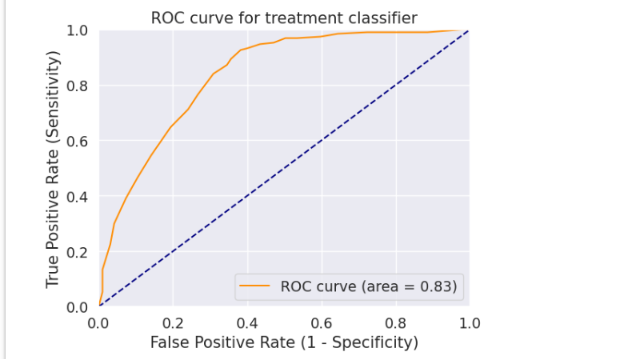
**Figure 27: Tuning the KNN**



**Figure 28: KNN prediction with confusion matrix**



**Figure 29: Histogram of the KNN classifier with predicted values**



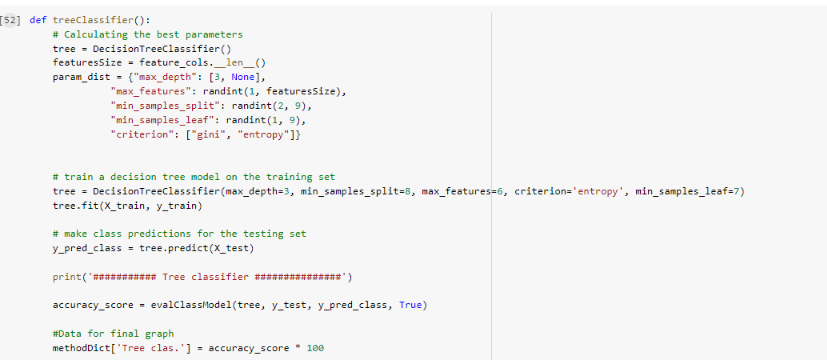
**Figure 30: ROC Curve for the KNN classifier**

## DECISION TREE CLASSIFIER

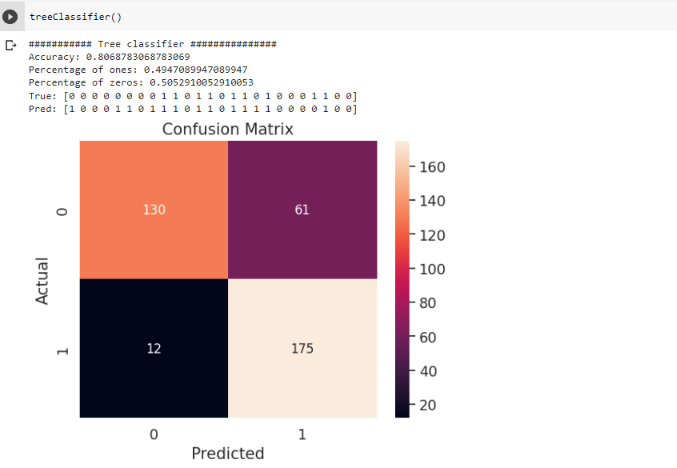
### TRAINING THE DECISION TREE CLASSIFIER

Tuning a Decision Tree Classifier is an essential step in optimizing the model's performance. The hyperparameters of the decision tree model that can be tuned include:

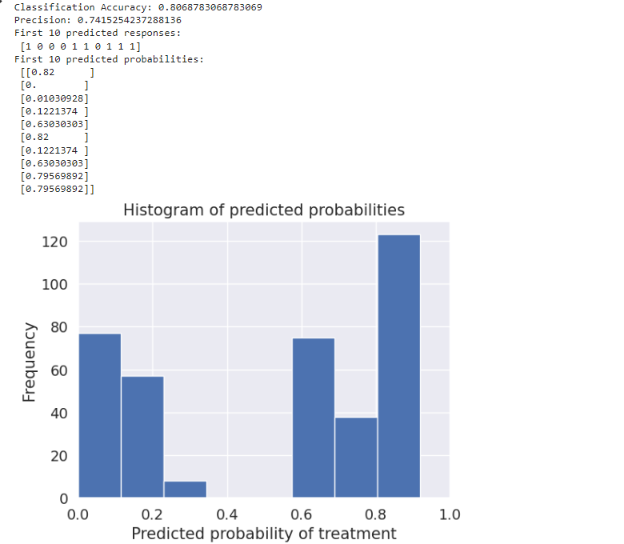
* Maximum depth of the tree (max\_depth): The maximum depth of the decision tree, which limits the number of levels of nodes in the tree.
* Minimum number of samples required to split an internal node (min\_samples\_split): This refers to the smallest quantity of samples that must be present for a node to be divided in the decision tree.
* Minimum number of samples required to be at a leaf node (min\_samples\_leaf): The minimum number of samples required to be at a leaf node in the decision tree.
* Maximum number of leaf nodes in the tree (max\_leaf\_nodes): The maximum number of leaf nodes allowed in the decision tree.



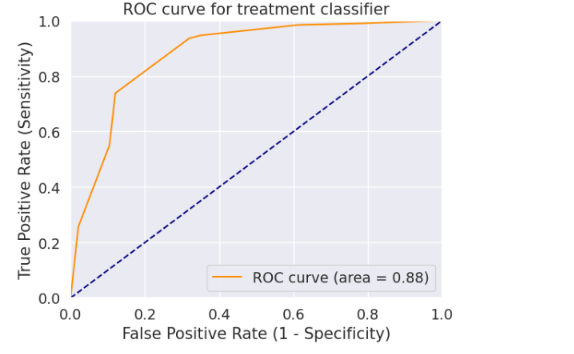
**Figure 31: Training of the Tree Classifier**



**Figure 32: Decision Tree prediction with confusion matrix**



**Figure 33: Histogram Predictions of the Tree Classifier**



**Figure 34: ROC Curve for the Tree Classifier**

# MODEL COMPARISON

## STOCHASTIC GRADIENT DESCENT (SGD)

SGDC (Stochastic Gradient Descent Classifier) is a linear classifier that is part of the scikit-learn library. It is a modification of the classic gradient descent algorithm and uses stochastic gradient descent to optimize a linear model with a hinge loss (linear SVM) or log loss (logistic regression). SGDC Classifier has several hyperparameters that can be tuned to optimize its performance. Some of these hyperparameters include:

* **penalty**: The regularization term used to prevent overfitting. It can be set to 'l1', 'l2', 'elasticnet', or 'none'.
* **alpha**: The regularization strength, which controls the amount of shrinkage applied to the weights. Smaller values of alpha lead to less regularization and a more complex model, while larger values lead to more regularization and a simpler model.
* **loss**: The loss function used to optimize the model. It can be set to 'hinge' for linear SVM or 'log' for logistic regression.
* **max**\_**iter**: The maximum number of iterations allowed for the optimization algorithm.



**Figure 35: SGDC Classifier score**



**Figure 36: SGDC prediction FIT with Grid Search CV**

# CONCLUSION

In conclusion, data visualization has numerous possible applications in various fields, but we must also consider the practical and ethical challenges it poses. We have discussed essential theoretical and practical guidelines for creating data visualizations and evaluated several visualization examples to identify common mistakes and helpful tips. Developing a trustworthy and ethical data visualization can be a complex process, and in many cases, automating data mining using Python can be a useful and efficient solution. Popular Python data mining techniques include association rules, clustering, regression, and classification.

Data visualization is a critical aspect of machine learning, helping to explore and analyze datasets, communicate insights and findings, and evaluate model performance. Effective data visualization can enhance the accuracy and interpretability of machine learning models, allowing stakeholders to make informed decisions based on data-driven insights.

However, creating impactful and ethical data visualizations requires careful consideration of various theoretical and practical guidelines, as well as potential biases and ethical implications. Additionally, automating data mining using Python can be a valuable approach to enhance the efficiency and accuracy of data visualization and analysis, particularly for complex datasets.

In summary, data visualization is a vital tool for machine learning, and by following best practices and ethical principles, we can maximize its potential to derive insights and improve decision-making in various domains.

# 7. REFERENCES

[1] “Kaggle: Your Home for Data Science.” <https://www.kaggle.com/> (accessed Apr. 27, 2023).

[2] Song, C., Ristenpart, T. and Shmatikov, V. (2017) ‘Machine Learning Models that Remember Too Much’, in Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York, NY, USA: Association for Computing Machinery (CCS ’17), pp. 587–601. Available at: <https://doi.org/10.1145/3133956.3134077>.

[3] How to develop machine learning models for healthcare | Nature Materials. Available at: <https://www.nature.com/articles/s41563-019-0345-0> (Accessed: 27 April 2023).

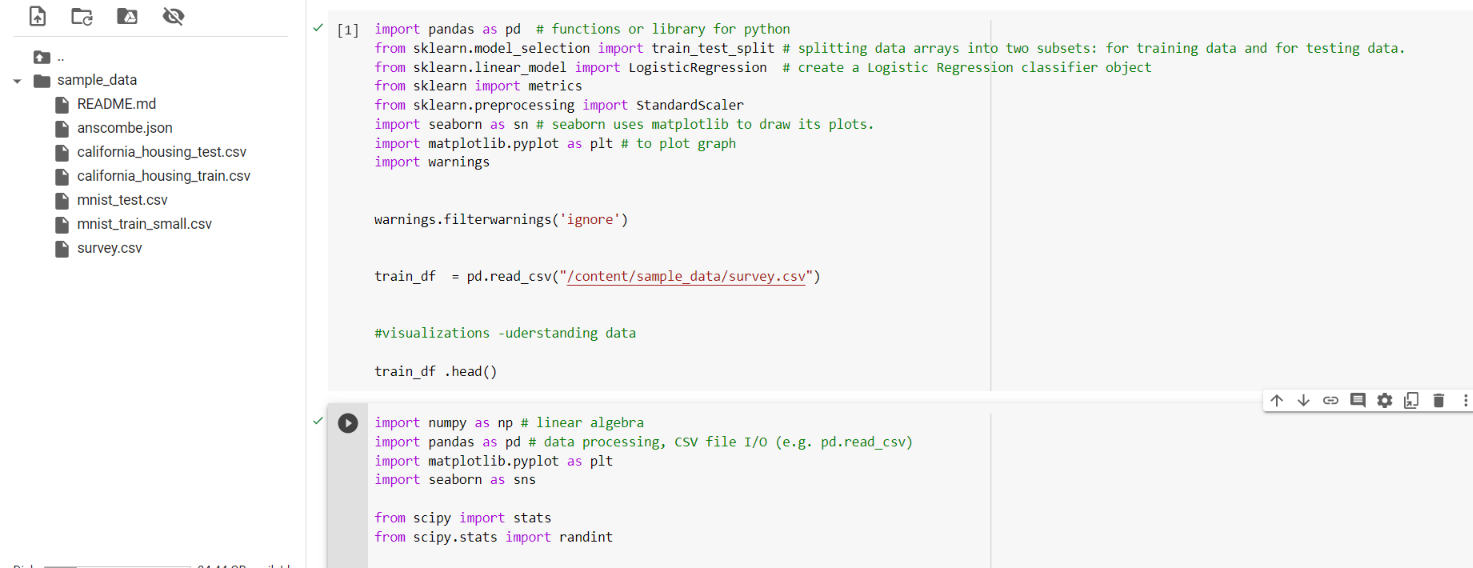
[4] Machine Learning Techniques for Prediction of Mental Health | IEEE Conference Publication | IEEE Xplore. Available at: <https://ieeexplore.ieee.org/abstract/document/9545061> (Accessed: 27 April 2023).

[5] Srividya, M., Mohanavalli, S. and Bhalaji, N. (2018) ‘Behavioral Modeling for Mental Health using Machine Learning Algorithms’, Journal of Medical Systems, 42(5), p. 88. Available at: <https://doi.org/10.1007/s10916-018-0934-5>.

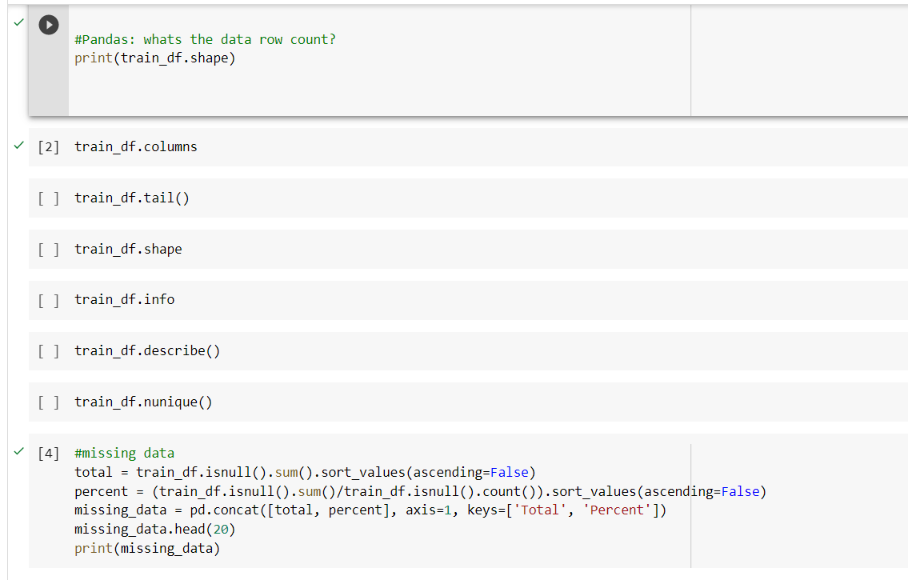
[6] Mohanapriya, M. and Lekha, J. (2018) ‘Comparative study between decision tree and knn of data mining classification technique’, Journal of Physics: Conference Series, 1142(1), p. 012011. Available at: <https://doi.org/10.1088/17426596/1142/1/012011>.

[7] A simple method of sample size calculation for linear and logistic regression - Hsieh - 1998 - Statistics in Medicine - Wiley Online Library (no date). Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)10970258(19980730)17:14%3C1623::AID-SIM871%3E3.0.CO;2-S> (Accessed: 27 April 2023).

# 8. APPENDIX

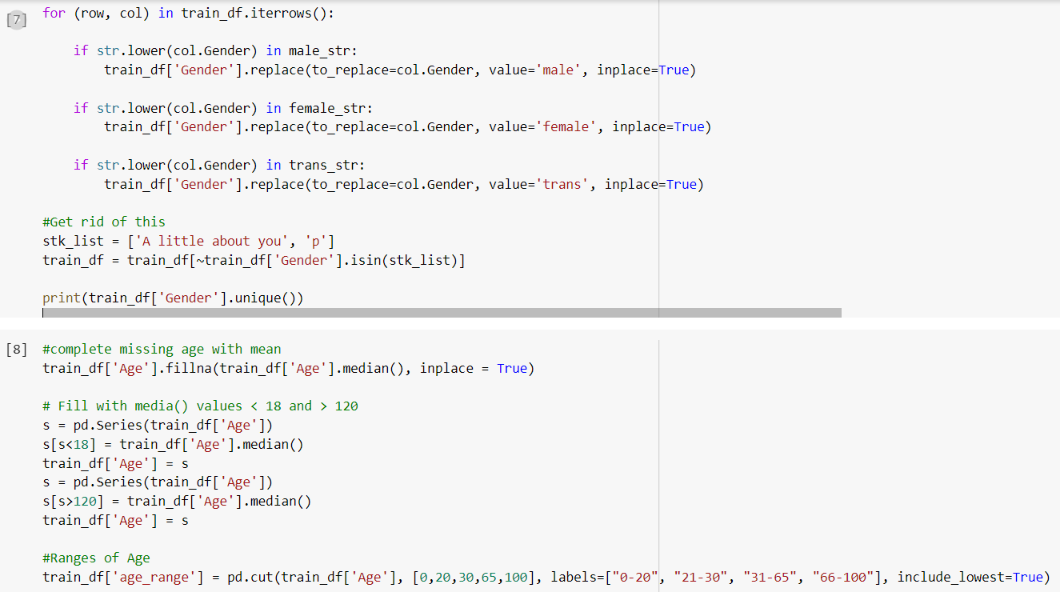


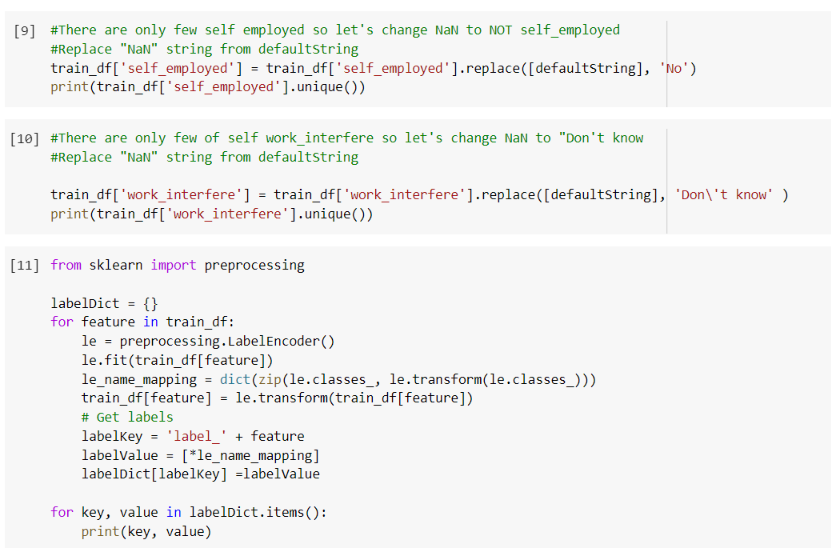








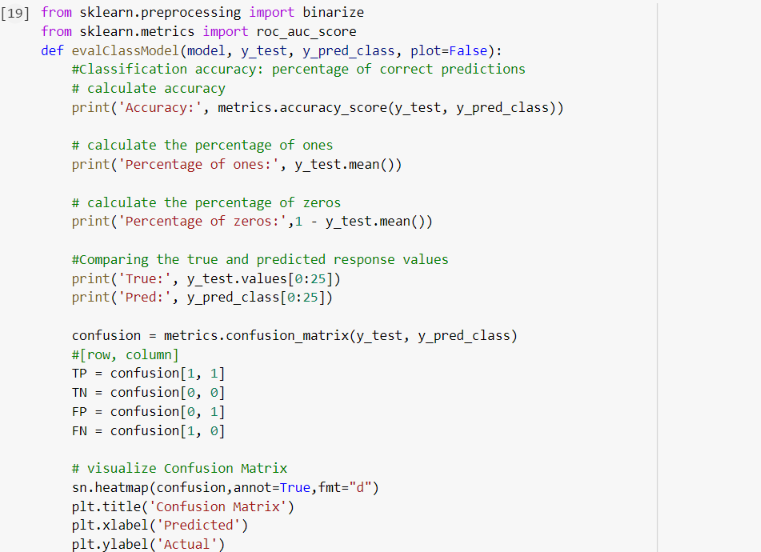


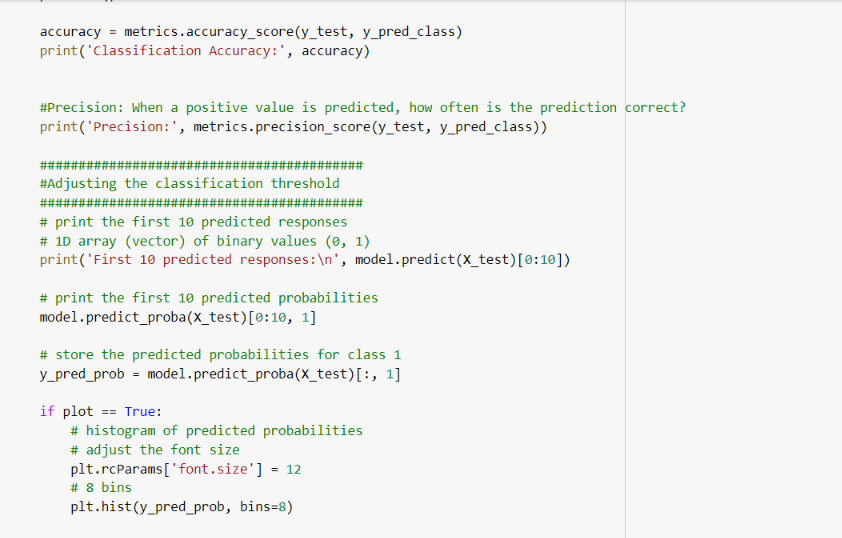


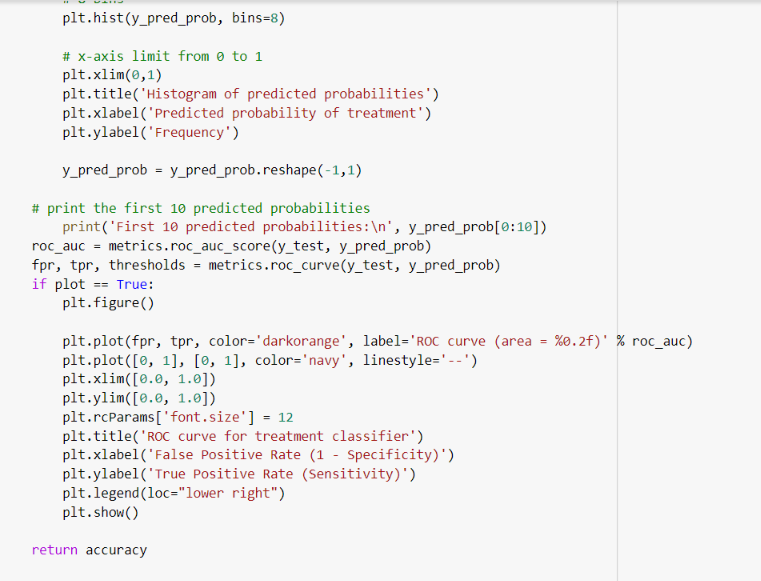




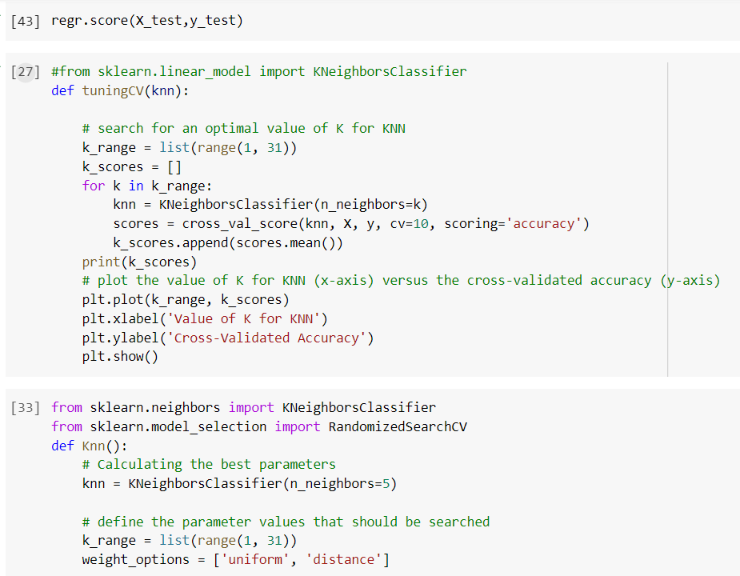


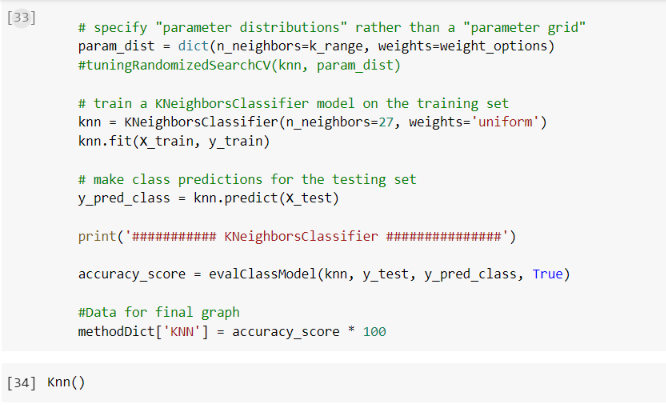
















# 9. COLAB LINK

<https://colab.research.google.com/drive/1Fb579KUbnQj8MQ8EBhcRZhReQEd_06_R?usp=sharing>