**Predicting the risk of obesity based on daily and dietary behaviours**

Trina Chatterjee, Jessica Greenwell, Ruthvik Jujjavarapu, Yamini Vanama, Upamanyu Mondal, Lijitha Nannapaneni, Preethi Reddy Nomula

Indiana University-Purdue University, Indianapolis, IN 46202, USA,

[upmondal@iu.edu](mailto:upmondal@iu.edu), [rujujj@iu.edu](mailto:rujujj@iu.edu), [jemigree@iu.edu](mailto:jemigree@iu.edu), [trichatt@iu.edu](mailto:trichatt@iu.edu), [lijnann@iu.edu](mailto:lijnann@iu.edu), [yvanama@iu.edu](mailto:yvanama@iu.edu), [pnomula@iu.edu](mailto:pnomula@iu.edu)

**Abstract:** The major goal is to identify reliable risk factors for obesity, equip doctors with the knowledge to stop a chain reaction of chronic illnesses, and improve patients' quality of life. This analysis intends to evaluate the effectiveness of five distinct machine learning (ML) approaches, including multinomial logistic regression, decision trees, random forests, support vector machines, and K nearest neighbors. It also focused on identifying the incidence of obesity utilizing open source healthcare data, by using convenient method with advanced ML techniques to predict obesity in such an effort to transcend conventional prediction models.

**Keywords:** Obesity, risk factors of Obesity, food habits, physical conditions, public health, data analysis, SQL, python, data visualization, ML prediction.

**1. Project Scope:**

**1.1. Introduction:**

Obesity is commonly thought to simply be caused by overeating. While this is a significant contributing factor, it is not the only explanation. Comparable to other chronic health issues, obesity has numerous risk factors, such as, genetics, education, medications, physical activity, comorbidities, and socioeconomic status. Unfortunately, another external source that is arguably out of most individuals control is food ingredients, especially highly processed foods. “Worldwide obesity has nearly tripled since 1975…In 2016, more than 1.9 billion adults aged 18 years and older were overweight. Of these over 650 million adults were obese” [1]. Obesity is the primary cause of numerous major health problems, in addition to its cosmetic adverse effects. resulting in illnesses like cancer, diabetes, arthritis, heart disease, elevated blood pressure, sleep apnoea, and heart problems. These illnesses are therefore significant causes of death. “Most of the world's population live in countries where overweight and obesity kills more people than underweight” [1]. The healthcare sector will be able to preventively limit the rate of these resulting disorders from happening if researchers are successful in properly predicting the root causes of obesity.

**1.2. Aim:** The main aim is to find credible risk factors for obesity and give physicians the tools to prevent a cascade of chronic health issues and increase the quality of life for their patients.

Research question:

•To study the cause of obesity based on the risk factors related with food habits and physical conditions.

**Prediction:**

•Possibly predict the association of obesity with dependent and independent attributes such as Frequency of high-calorie food consumption, frequency of vegetable consumption, and frequency of main meals, alcohol consumption, transportation etc.

* The association will then be utilized to create a prediction model that will equip doctors with the knowledge and resources they need to counsel patients, stop a chain reaction of chronic health problems, and ultimately raise the standard of living for the patients.

**Null Hypothesis:** The independent attributes such as a person’s eating habits, and physical conditions do not have any association with obesity and the attributes cannot help predicting the chances of a person being obese.

**Alternate Hypothesis:** The independent attributes such as a person’s eating habits, and physical conditions have an association with obesity and the attributes can help predict the chances of a person being obese.

**1.3. Purpose:**

The purpose of this study is to find if the consumption of more calorie-dense, high-fat meals as well as a decline in physical activity as a result of sedentary jobs, new transportation technologies, and growing urbanization are some of the factors that contribute to obesity. According to CDC data “from 1999 –2000 through 2017 –March 2020, US obesity prevalence increased from 30.5% to 41.9%.”[2] Obesity affects both individuals and children and is a frequent, dangerous, and expensive chronic condition. Every day, new research addressing childhood obesity develops, particularly those that seek to identify influence variables and determine how to foresee the emergence of the disease under these factors. This is because obesity is a problem that has been continuously becoming worse. The data for this study was collected via a web platform using an anonymous survey. Seventeen attributes were developed based on answers to questions about subjects' gender, daily nutrition, time spent on technological devices, smoking, and drinking habits, choice of transportation, etc. [3] We would like to analyse the given variables in our dataset containing information on a person’s personal and dietary habits and predict a person’s chance of being obese. We are also interested in categorizing the risk factors based on their effect on obesity. With this, we would like to create a prediction model that will equip doctors with the knowledge and resources they need to educate the patient, prevent the onset of chronic health disease, and ultimately improve patients' quality of life.

**2. Methodology:**

**2.1. Steps of the study**: The study will involve quantitative research in which we will develop a model to forecast the likelihood of being obese based on habits and use descriptive statistics. We would like to detect the levels of obesity utilizing Structured query language(SQL) and Python based on the selected factors leading to obesity. For this, we have used technology tools like phpMyAdmin and Python Jupyter notebook.

The steps of our project include,

* Data collection
* Data Extraction and description
* Data Visualization
* Data analysis
* ML Model Building

**2.2. Team member’s Proposed Responsibilities**:

Following a thorough discussion based on the abilities of the team members in various skill areas, these are the duties we shared as a team.

|  |  |
| --- | --- |
| Name | Responsibility |
| Trina Chatterjee | Data collection |
| Jessica Greenwell | Data extraction and cleaning |
| Ruthvik Jujjavarapu | Exploratory Data Analysis |
| Upamanyu Mondal | Binary classification model development |
| Lijitha Nannapaneni | Data visualization |
| Preethi Reddy Nomula | Data visualization |
| Yamini Vanama | Data evaluation |

**2.3. Team Member’s Actual contributions:**

We first determined each member's areas of strength and distributed the work effectively. However, as the project progressed, a few adjustments to the team members' tasks were needed as mentioned below.

|  |  |
| --- | --- |
| Name | Responsibility |
| Trina Chatterjee | Data collection and Proofreading |
| Jessica Greenwell | Project presentation and Data extraction |
| Ruthvik Jujjavarapu | Evaluation of data and presentation |
| Upamanyu Mondal | Classification model building |
| Lijitha Nannapaneni | Data cleaning and visualization |
| Preethi Reddy Nomula | Data visualization and report |
| Yamini Vanama | Statistical data analysis and report |

**2.4. Project Challenges:**

Despite having excellent technical skills and excellent project group coordination, there were a few difficulties with the project. Only 500 rows are imported at first when the dataset is imported into SQL as a csv file, indicating that the data is too large. Later, we discovered that the data contains decimal values for a few columns after 500 rows, which are too huge to be imported. Therefore, we successfully imported the data into SQL after rounding it to two decimal places. Due to the increasing amount of data we are working with, we have run into a few problems with Jupyter Hub. We renamed the variables because it was challenging for us to grasp the actual variable names given that the data contained.

abbreviations for the column names. Another challenge is that the target column is the stages of obesity, which often have three stages, such as obesity stages 1, 2, and 3. In order to indicate a person's normal weight, excess weight, or underweight, we included a new column called BMI. Last but not least, as a team, we overcame all the difficulties by physically gathering or communicating via Zoom, which might be more practical for both online and offline group members.

**3. Data Collection:**

The source of the data is UCI machine learning repository. This dataset contains seventeen variables contributing to Obesity which are developed based on answers to questions about subjects' gender, height, weight, daily nutrition, time spent on technological devices, smoking, and drinking habits, choice of transportation, etc.

Website Link to the dataset: <https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+>

**4. Data Extraction and description:**

The first part of data extraction and storage involved determining attributes from our dataset, we had around 17 such attributes in total.

The attributes include:

* Frequent consumption of high caloric food (FAVC)
* Frequency of consumption of vegetables (FCVC)
* Number of main meals (NCP)
* Consumption of food between meals (CAEC)
* Consumption of water daily (CH20)
* Consumption of alcohol (CALC)
* Calories consumption monitoring (SCC)
* Physical activity frequency (FAF)
* Time using technology devices (TUE)
* Transportation used (MTRANS)
* Gender
* Age
* Height
* Weight
* Obesity stages

**4.1. Data Import:**

We have imported the data set in the form of CSV file into SQL using phpMyAdmin to look for the variables provided. Later on we connected the SQL to python using MySQLdb. We continued our next steps using pandas python library for reading the file in the python.

**Graphical user interface, application, table, Excel

Description automatically generated**

**Fig. 1.** Data imported into SQL

**4.2. Data cleaning:**

Data cleaning entails selecting the variables that are useful for predicting outcomes and answering research questions. We started out by looking at the dataset's null values. Finding the columns containing null values was part of this process. The data collection is free of null values, as we discovered. In the data, we also searched for duplicate values. A few duplicate values were discovered, so for accurate prediction, we removed them from the dataset.

After data cleaning, we have created a new column for BMI. And pre-processed the data as we have categorical variables in our data. We did label encoding to convert our categorical data into numerical data which aids us in easy data analysis.

Table

Description automatically generated

**Fig. 2.** Data Cleaning process

Table

Description automatically generated

**Fig. 3.** Dropping the duplicate values.

**5. Data Visualization:**

Exploratory data analysis is an approach to analysing and understanding data sets by using visualizations and summary statistics. It is a way to explore the data and discover patterns and relationships, and to identify potential outliers or anomalies.

In order to use categorical variables in statistical analysis or machine learning algorithms, they must be encoded as numerical values. This process is called label coding, and there are several ways to do it. One common method is to assign a unique number to each possible value of the categorical variable. For example, for Obesity\_stages, “Insufficient \_weight” was encoded as 0, “Normal\_weight" as 1, “Obesity\_Type\_I” as 2 and so on. Similarly, all other categorical variables were label encoded. This allows the categorical data to be used in numerical calculations, but it does not convey any information about the relative magnitude of the different categories.

One type of visualization used in exploratory data analysis is a count plot, which shows the count of observations in each category of a categorical variable. Count plots are useful for understanding the distribution of the data and identifying any imbalances or patterns in the data. For example, we have data on the gender of a group of people, a count plot can shows us how many people are male and how many are female. This can help us understand the makeup of the group and identify any potential biases or issues with the data. Some of the inferences that were made were that individuals with a family history of obesity were more likely to be obese, there were a greater number of obese people who did not monitor their daily calories and that very few participants were smokers.

Chart, bar chart

Description automatically generated

**Fig. 4.** Count plot of Gender and Obesity

Chart, bar chart

Description automatically generated

**Fig. 5.** Count plot of Obesity Stages.

Chart, bar chart

Description automatically generated

**Fig.6.** Count plot of Family History with overweight.

Chart

Description automatically generated

**Fig. 7.** Count plot of Consumption of food between meals Vs Obesity.

Chart, bar chart

Description automatically generated

**Fig. 8.** Count plot of Smoker Vs Obesity.

Chart, bar chart

Description automatically generated

**Fig. 9.** Count plot of Daily calories monitor Vs Obesity.

Chart, bar chart

Description automatically generated

**Fig. 10.** Count Plot of Physical activity frequency.

Chart, bar chart

Description automatically generated

**Fig. 11.** Count plot of time using technology devices.

Chart, bar chart

Description automatically generated

**Fig. 12.** Count plot of Alcohol V/s Obesity.

Following this, we plotted boxplots, which are a type of visualization that is used to summarize the distribution of a numeric variable and for identifying potential outliers. It is particularly useful for comparing the distribution of the variable across different categories. In the case of categorical variables, a separate boxplot is created for each category. The end of the box represents the upper and the lower quartiles, and the line within the box indicates the median.

Chart, box and whisker chart

Description automatically generated

**Fig. 13.** Box plot of Gender Vs Weight, Height, Age, BMI.

Chart, box and whisker chart

Description automatically generated

**Fig. 14.** Box plot of Consumption of food between meals Vs Age, Weight, Height, BMI.

We plotted a scatterplot to determine the distribution of our dataset. A scatterplot is a type of visualization that Is used to display the relationship between two numeric variables. When one or both variables being plotted are normally distributed, the points on the scatterplot will tend to cluster around a central point and will fall off symmetrically in both directions. If the points form a roughly symmetrical pattern around a central point, it may indicate that the variables are correlated, and that one variable is a good predictor of the other. From our scatterplot, we could infer that the distribution of BMI among the population was slightly positively skewed, i.e, the variables moved in the same direction. Most individuals with a high BMI fell under the age group of 20-30.

Chart, scatter chart

Description automatically generated

**Fig. 15.** Scatterplot of BMI distribution with Age and Gender.

We plotted histogram to know the distribution of Age and BMI and we found that the data is not normally distributed and is right skewed. Finally, we used Seaborn to plot pair plots in order to determine the pairwise distribution of the data in order to discover a trend for further study.

Chart, histogram

Description automatically generated

**Fig. 16.** Histogram showing Age Distribution.

Chart, histogram

Description automatically generated

**Fig. 17.** Histogram showing BMI distribution.

Scatter chart

Description automatically generated

**Fig. 18.** Pair plot showing pairwise distribution of data.

**6. Data Analysis:**

**6.1. Correlation Testing:**

We used the Spearman test to find the correlation among the variables. This method was used since BMI is not normally distributed. It was seen that the weight of a person, family history of obesity and age of a person is having the highest correlation with BMI. Whereas on the other hand physical activity, monitoring of daily calories, and consumption of food in between meals are negatively correlated.

**Chart

Description automatically generated with medium confidence**

**Fig. 19.** Heatmap for correlation test.

**6.2. Hypothesis Testing:**

As our data is not normally distributed and consists of various categorical variables we choose to go for non-parametric hypothesis testing.

1. **Kruskal Wallis Test:**

We performed a Kruskal Wallis Test between the variables: Alcohol, BMI, Frequency\_of\_consumption\_of\_vegetables, Number\_of\_main\_meals, Consumption\_of\_food\_between\_meals, Daily\_Water\_Intake, Physical\_activity\_frequency, Time\_using\_technology\_devices, Transportation\_used.

**The test results are as follows:**

Kruskal Result (Statistic = 3259.075, p-value = 0.0) for Alcohol and BMI.

Kruskal Result (Statistic = 3247.616, p-value = 0.0) for Frequency\_of\_consumption\_of\_vegetables and BMI.

Kruskal Result (Statistic = 3280.073, p-value = 0.0) for the Number of main meals and BMI.

Kruskal Result (Statistic = 3384.538, p-value = 0.0) for Consumption\_of\_ food\_between\_meals and BMI.

Kruskal Result (Statistic = 3204.620, p-value = 0.0) for Daily\_Water\_Intake and BMI.

Kruskal Result (Statistic = 3186.673, p-value = 0.0) for Physical\_activity\_frequency and BMI.

Kruskal Result (Statistic = 3205.989, p-value = 0.0) for Time\_using\_technology\_devices and BMI.

Kruskal Result (Statistic = 3305.996, p-value = 0.0) for Transportation\_used and BMI.

Text

Description automatically generated

**Fig. 20.** Kruskal Wallis test

We performed the Kruskal Wallis Test for all independent variables that are having more than three samples with the dependent variable i.e. BMI. And we found that the p-value is significantly lower than the statistic value for all the variables. Since our p-value is significantly lower than our test statistic, we have strong evidence against our null hypothesis.

**2**. **Mann-Whitney U Test:**

We also performed the Mann-Whitney U Test between the categorical variables: Gender (Male and Female), Smoker, Family\_history\_with\_overweight, No\_Family\_history\_with\_overweight, High\_caloric\_food\_intake, No\_High\_caloric\_food\_intake, Monitor\_Daily\_Calories, No\_Monitor\_Daily\_Calories.

**The test results are as follows:**

Mann-Whitney U Result (Statistic = 531382.5, p-value = 0.3439) for Male and Female BMI.

Mann-Whitney U Result (Statistic = 45100.5, p-value = 0.9689) for Smoker and Non-Smoker BMI.

Mann-Whitney U Result (Statistic = 551974.5, p-value = 2.2401) for Family history with overweight and No Family history with overweight BMI.

Mann-Whitney U Result (Statistic = 325378.5, p-value = 1.7426) for High Caloric Food Intake and No High Caloric Food Intake BMI.

Mann-Whitney U Result (Statistic = 325378.5, p-value = 1.7426) for Monitor Daily Calories and No Monitor Daily Calories BMI.

Text

Description automatically generated

**Fig. 21.** Mann Whitney U test

We performed the Mann-Whitney U Test for all independent categorical variables that are two sampled with BMI column. Apart from the variables relating to gender and smoking, we discovered that the p-value is less than 0.05. We can reject the null hypothesis except that the gender and smoking are not related in causing obesity.

**7. Machine learning - Model building:**

Firstly, for machine learning we used Scikit learn python library for building machine learning models.

Data can now be utilized to train a machine learning model after being collected and cleaned. With the help of this cleaned data, we trained the K nearest neighbors, random forest, decision tree, logistic regression, and support vector machine. we performed splitting the data into a 70:30 ratio, meaning that 70% of the data would be used to train the model and 30% would be used to evaluate how well the model performed. The performance of the algorithms was assessed using a confusion matrix, accuracy score, classification metrics and ROC curve.

Graphical user interface, text, application, email

Description automatically generated

**Fig. 22.** Code for Train and Test Data Split.

**7.1. Logistic Regression:**

By using the feature attributes(independent factors) as X and the obesity stages (the target/dependent variable) as Y, we created a Logistic Regression model. The risk of obesity can be predicted using the feature attributes (independent variables). We dropped the smoke column in the feature attributes as there is no relation between smoke and obesity which we found during our statistical tests. We have used multinomial class of logistic regression as our data contains more than 2 categorical samples and balanced class to balance class weight.

Graphical user interface, text, application, email

Description automatically generated

**Fig. 23.** Code for Logistic Regression.

**7.2. Random Forest:**

An approach for classifying data called Random Forest combines a variety of decision trees within to perform collective learning-based supervised machine learning. Each internal decision tree algorithm in a random forest model are considered weak classifiers, and their outputs are merged, or the average of all forecasts, to get the final prediction.

Graphical user interface, application

Description automatically generated with medium confidence

**Fig. 24.** Code for Random Forest model

**7.3. Decision tree:**

For this model we imported Decision tree classifier from sklearn.tree in python. We fit the data in to the X\_train and y\_train. We used random state of 42 which controls randomness and entropy criterion.

Graphical user interface, text, application, email

Description automatically generated

Fig. 25. Code for Decision tree model.

**7.4. Support Vector Machine:**

We utilized an exclusionary classifier called the Support Vector Classifier (SVC), which searches for the best hyperplane to effectively divide samples into distinct categories in hyperspace when it is provided trained data with labels.

Graphical user interface, text, application, email

Description automatically generated

**Fig. 26.** Code for Support Vector Machine.

**7.5. K nearest neighbors:**

We used the k-nearest neighbors (KNN) model, which calculated the probability that a data point would belong to one category or another depending on the group to which the data points closest to it belong.

Graphical user interface, text, application, email, Teams

Description automatically generated

**Fig. 27.** Code for K nearest neighbour

**7.6. Confusion Matrix:**

The effectiveness of the models was evaluated using the confusion matrix. In order to compare actual and projected values, it became essential to interpret the confusion matrix. This allowed us to analyse false positives and false negatives and gain a detailed picture of metrics such as recall, specificity, sensitivity and F1 score. Below are the confusion matrices which shows performance of each model.

Background pattern

Description automatically generated with medium confidenceA picture containing background pattern

Description automatically generated

A picture containing background pattern

Description automatically generatedA picture containing background pattern

Description automatically generated

A picture containing text, crossword puzzle

Description automatically generated

**Fig. 28.** Confusion matrix showing performance of various models.

**7.7. Receiver Operating Characteristic Curve (ROC):**

ROC curves are used to visualize the effectiveness of algorithms and classification settings. The ROC-AUC (Area under the curve) statistic is used to assess the classification model with the best performance. Using the area under the curve measurement and the micro average roc curve, we discovered that the model "Random Forest" is the best performing model. Given below are the ROC curves showing the model performances.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, line chart

Description automatically generatedChart

Description automatically generated

Chart

Description automatically generated

Fig. 29. ROC curves plotting model performances.

**8. Summary:**

We learned a lot about our study question after going through our methodology of data cleaning, data analysis, and data visualization. After reviewing the data and completing proper analysis and modeling, it is reasonable to conclude that the stated attributes contribute to obesity. Some of our core findings are shown below.

Research question:

To study the cause of obesity based on the risk factors related with food habits and physical conditions. These are the findings we obtained to support our research question.

1) Out of all the remaining models, Random Forest is the most accurate at predicting with a prediction rate of 94%.

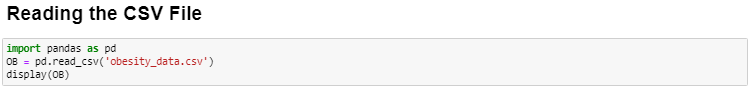
2) It can be observed that using the prediction models, we can predict the stages of obesity using the provided variables.

3) We conclude from our analyses that there is an association between obesity and dependent and independent variables such as frequency of high-calorie food intake, frequency of vegetable consumption, frequency of main meals, alcohol consumption, transportation.

**9. Limitations:**

The main limitation of this project is that it predicts the obesity of a person depending on their lifestyle and BMI. Though BMI is an important parameter to predict obesity it is not always correct to group person with higher BMI as obese. This is because the BMI of a person can be high even if they have higher muscle mass. So to ideally predict if a person is obese or not body fat percentage should be considered which is missing in the data that we used.

**Appendix:**



Text

Description automatically generated

Table

Description automatically generated



Graphical user interface, table

Description automatically generated



Text

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated



A picture containing text

Description automatically generated

Table

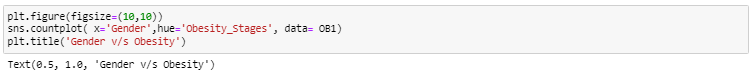
Description automatically generated with low confidence

Graphical user interface, text, application, email

Description automatically generated

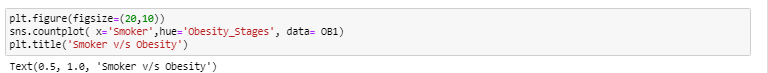
Table

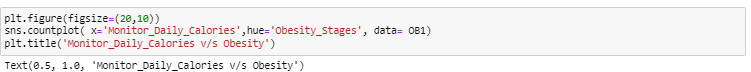
Description automatically generated







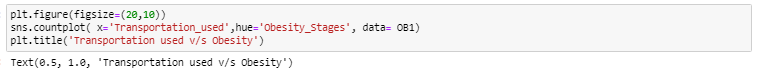






Text

Description automatically generated with low confidence



A picture containing shape

Description automatically generated

A picture containing shape

Description automatically generatedA picture containing text

Description automatically generated

A picture containing chart

Description automatically generatedA picture containing graphical user interface

Description automatically generated

Shape

Description automatically generated with medium confidenceShape

Description automatically generated with low confidence



Text

Description automatically generated

Text

Description automatically generated with medium confidence



A picture containing text

Description automatically generated

A picture containing chart

Description automatically generated



Rectangle

Description automatically generated with low confidence

Text

Description automatically generated with low confidence

Table

Description automatically generated with low confidence

Text

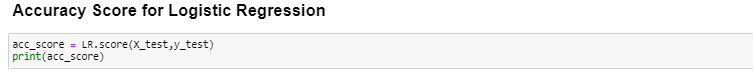
Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated



Graphical user interface, text, application

Description automatically generated

Rectangle

Description automatically generated with low confidence

Graphical user interface, text, application

Description automatically generated

Text, table

Description automatically generated with medium confidence

A picture containing text

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

Graphical user interface, table

Description automatically generated with medium confidence

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Table

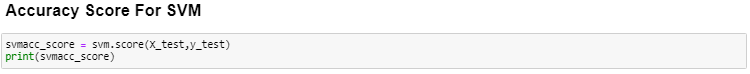
Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Shape, rectangle

Description automatically generated



Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidence

Graphical user interface

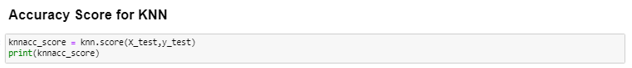
Description automatically generated with low confidence

Graphical user interface, application

Description automatically generated

Table

Description automatically generated with low confidence



Graphical user interface, text, application

Description automatically generated



Graphical user interface, text, application

Description automatically generated

Application

Description automatically generated with low confidence

Table

Description automatically generated

Text

Description automatically generated