Logo, company name

Description automatically generatedSEP 786 (ML & AI)

(Course Project)

**(Maged Eid - ID:400488004)**

**Draft for (Nicky) BZhang Lab**

**Notes:**

**Obesity Dataset from UCI Website:**

* This data comes from the **UCI Machine Learning Repository.** This dataset includes data for the estimation of obesity levels in individuals from different countries, based on their eating habits and physical condition.
* The dataset has 2111 records.
* I used Python programming language along with many different ML libraries. I took screenshots of my codes from Jupyter Notebook with the output of it.
* I applied Logistic regression (**Classification**), Random Forest (**Classification**), Linear regression (**Regression**), and Decision tree (**Regression**) on the ObesityDataSet\_raw\_and\_data\_sinthetic.csv dataset. My goal was to predict the risk of obesity based on various factors. The random forest model was the most accurate, followed by logistic regression, linear regression, and decision tree regression. Overall, the study demonstrated the effectiveness of machine learning algorithms in predicting obesity risk.
* (MSE) tool for evaluating the performance of a regression model. (Confusion Matrix) tool for evaluating the performance of a classification model. I applied both methods to all the models whether it’s Classification or Regression out of curiosity and to see the output of it. I know that if you attempt to apply a confusion matrix to a regression problem or vice versa, the results would be meaningless, misleading, and could result in poor decision-making.
* I applied mean absolute error (MAE) tool to compare it with confusion (MSE) for regression models.
* I used (“test\_size=0.3”) which means that 30% of the data is used for testing, and the remaining 70% is used for training.
* I used the (“random\_state”) parameter. This parameter ensures that the data is split in the same way each time the code is run, which can help with the reproducibility of the results. I assigned it to number 42, because I saw it in a lot of examples online, after a search I did, I found that the choice of the number 42 as the seed value for the random\_state parameter in machine learning code is often attributed to its use in the book "The Hitchhiker's Guide to the Galaxy" by Douglas Adams. In the book, 42 is famously described as the "Answer to the Ultimate Question of Life, the Universe, and Everything". So, it’s just a random number I found in many examples, so I decided to use it too.
* The number of components I used for LDA is set to 2 (Two dimensions) using the parameter n\_components = 2 when initializing the LDA object to find the two most informative linear discriminants to project the data onto.
* I used different scikit-learn (sklearn) algorithms for tasks such as classification, regression, dimensionality reduction same as I did in Assignment 2, as well as tools for data preprocessing.
* I used LabelEncoder preprocessing function to convert categorical data, represented as text labels, into numeric values that machine learning algorithms can work with, and used StandardScaler preprocessing function to scale the data so that it has zero mean and unit variance. I used OneHotEncoder to encode categorical data as one-hot vectors. One-hot encoding is a way to represent categorical data as binary vectors, where each vector has a 1 in the position corresponding to the category and 0s in all other positions. This allows the categorical data to be used as input to machine learning algorithms that require numerical data. (I took all of that information from different sources especially google to help me to run my code smoothly)

In this study, I applied several machine learning algorithms, including logistic regression, random forest, linear regression, and decision tree regression, on the ObesityDataSet\_raw\_and\_data\_sinthetic.csv dataset. This dataset contains information about various factors that can contribute to obesity, such as age, gender, height, weight, and lifestyle habits. The goal was to build models that could predict the risk of obesity based on these factors.

I started by pre-processing the dataset, which involved cleaning the data, handling missing values, and converting categorical variables to numerical values. Then, I split the dataset into training and testing sets to evaluate the performance of the models.