Project Title	**“Data Analysis Techniques Visualization for Breast Cancer Prediction and Classification with High Accuracy Rates”**		
Course Code	**CSCI-322**	**Course Name**	**Data Analysis**
Professor	**Dr. Mustafa Elattar**		
TA	**Eng. Ali Abdelmageed**	**Mentor Name**	**Eng. Ali Abdelmageed**
Team Name	**SHOA**		
Team Members	**Shrouk Hesham**	**Toqa Hamdy**	**Asmaa Muhammed**
	Omnia Salah	**Abdelrahman Mahmoud**	**Hussin fekry**
Problem Summary	Breast cancer is one of the leading causes of death in women worldwide. Around-one in 30 women is affected by breast cancer. It’s better to have knowledge of the type of cancer in the preliminary stages because early diagnosis leads to better outcomes. Mammography has helped in detecting breast cancer in the early stages which have reduced mortality. The diagnosis of breast cancer is dependent on a variety of parameters. We aim to create the best model for predicting breast cancer through preprocessing, feature extraction, data visualization and prediction using breast cancer data. Various visualization techniques like violin plot, grid plot, swarm plot and heat plot were utilized for proper feature extraction which has improved the accuracy of our results. For the purpose of prediction, we have used algorithms like the random forest, decision tree with single and multiple predictors, along with the commonly used statistical model, logistic regression model.		
Methodology	The overall methodology process is done as follows. The dataset that we are going to use is taken from UCI machine repository. The next step is called the selection of the data. The data selected possess various attributes (around 32). Some of these may be redundant, while others may be lacking. There is also the possibility of noise in the data. As a result, the data must be cleaned before it can be processed further (data wrangling). They can be cleaned only after we explore or analyze the data. The attributes are viewed, and data is checked for what must be taken care-off. The data is then pre-processed to reduce noise and outliers. Because the data has 32 attributes, there's a good chance that just a small portion of them will be needed for data mining. These steps are taken in the Data Visualization process where data is visualized, and useful information can be taken out of it. The Violin plot, Swarm plot, join plot, and Heat map are used to visualize the data. This is the only way to determine which attributes are truly important for further processing. Then, we move to the feature extraction. Then, the data classification.		
Achievements and Skills Gained	• We managed to predict the future behavior of breast cancer disease (Malignant). • various models are applied to the training and testing data to check which on has the better accuracy. • We learned how to filter redundant, missing, and noisy data. • How to visualize the data using several plots.		
Project Title	**Data Analysis Techniques Visualization for Breast Cancer Prediction and Classification with High Accuracy Rates**		

(Cont.)

Main Results	classification_and_fit_model(model, data, predictor1, outcome) Accuracy : 95.782% Cross-Validation Score : 75.439% Cross-Validation Score : 76.316% Cross-Validation Score : 81.287% Cross-Validation Score : 81.579% Cross-Validation Score : 81.723% classification_and_fit_model(model, data, predictor1, outcome) Accuracy : 95.782% Cross-Validation Score : 75.439% Cross-Validation Score : 76.316% Cross-Validation Score : 80.994% Cross-Validation Score : 81.368% Cross-Validation Score : 81.371% classification_and_fit_model(model, data, features_mean, outcome) Accuracy : 96.485% Cross-Validation Score : 85.965% Cross-Validation Score : 89.035% Cross-Validation Score : 91.813% Cross-Validation Score : 92.544% Cross-Validation Score : 92.973% classification_and_fit_model(model, data, predictor1, outcome) Accuracy : 95.782% Cross-Validation Score : 75.439% Cross-Validation Score : 76.316% Cross-Validation Score : 81.287% Cross-Validation Score : 81.579% Cross-Validation Score : 81.723% classification_and_fit_model(model, data, predictor, outcome) Accuracy : 93.849% Cross-Validation Score : 82.456% Cross-Validation Score : 86.842% Cross-Validation Score : 90.643% Cross-Validation Score : 90.789% Cross-Validation Score : 91.216%
Discussion and Conclusion	In conclusion, using data visualization to diagnose breast cancer can be used to exclude some features and determine which ones are significant. All of the models were evaluated in terms of accuracy and cross-validation. According to our findings, the Random-Forest Model with top 5 predictors 'concave-points_mean','area_mean','radius_mean','perimeter_mean','concavity_mean' is the best model to apply for diagnosing breast cancer. As indicated in the results, it has a prediction accuracy of ~95% and a cross-validation score of ~93% for the test data. As a conclusion, it is easy to conclude that the Random-Forest classifier is the most accurate of all these models in terms of accurately diagnosing breast cancer with careful feature selection using data visualization for this data set.
References	❖ Williams, K., P.A. Idowu, J.A. Balogun and A.I. Oluwaranti, 2015. Breast cancer risk prediction using data mining classification techniques. Trans. Netw. Commun. DOI: 10.14738/tnc.32.662. ❖ Wolberg, W.H., W.N. Street and O.L. Mangasarian, 1994. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer Lett., 77: 163-71. DOI: 10.1016/0304-3835(94)90099-x.
Future Work and Suggestions	In our project future work, we can develop a robust data analytical model which can assist in better understanding of breast cancer survivability, providing better insights into factors associated with patient survival, and establishing cohorts of patients that share similar properties. In addition, the results of the analysis can be used to segment the historical patient data into clusters or subsets, which share common values. Furthermore, providing enhanced computational analysis and interpretation and thus saving patients' lives.
Group Photo