**Machine Learning Part:**

Wisconsin Breast Cancer data set consists of 30 features obtained from a digitized image of a breast mass that could act as predictors in our predictive model. The objective is to predict the stage of the breast cancer that could be either benign (B) or malignant (M). Therefore, we predict a binary outcome (B/M) and multiple machine learning models could be useful in our case. Below, I briefly overview the machine learning models that will be used in this capstone project and the challenges associated with them:

1. **Logistic Regression:** Logistic regression is the method of choice since the outcome of prediction is a binary parameter. This analysis is a supervised problem as the output dataset is provided and could be used to train the machine learner. The dataset is split into two subsets of training set and test set (the split ratio is defined based on the value of 1/sqrt(number of predictors) which would be 20% for test set in our case). The most important step in this part is to select the subset of predictors that lead to the best predictive model. It has to be noted that many of the variables are highly correlated to each other as they measure various morphology parameters that are related (e.g. area is correlated with perimeter). We would use different algorithms to choose the best subset of predictors including best subset regression algorithm and also removal of highly correlated variables with a specific threshold. The measures we use to define the best model include: accuracy, sensitivity, specificity, misclassification error, area under curve (AUC), BIC, AIC, and adjusted R squared.
2. **K-means clustering:** This method is used to classify our data points into two main clusters (B and M). K-means clustering is an unsupervised algorithm as it does not require the data to be labelled. However, we would still split the data to assess the model in predicting the data that it has never seen before. The outcome measures we use include: AUC, accuracy, sensitivity, specificity, and misclassification error. In this dataset, we already knew the number of clusters (B and M), otherwise we had to use algorithms to identify the optimal number of clusters.
3. **K-nearest neighbor algorithm:** This method is considered as supervised learning as a subset of data (training set) with their outcome variable is used to train the machine learner. The outcome measures we use include: AUC, accuracy, sensitivity, specificity, and misclassification error. We also use k-fold cross validation method to identify the optimal number of nearest neighbors. Additionally, the method will be used to define the number of folds that yield minimum misclassification error.
4. **Classification and Regression Tree and Random Forest:** This method is a more advanced algorithm for classification that could be considered as a supervised learning algorithm. We use random forest method to train a machine learner using the training set and assess the predictive power of the model in the test set based on these measures: AUC, accuracy, sensitivity, specificity, and misclassification error. Nodesize is a parameter needed to be determined for this algorithm and we identify its optimal value by relating the misclassification error to that parameter.