In this assignment a decision tree classifier using C4.5 algorithm was used to classify the Breast Cancer Wisconsin Data Set (named "breast-cancerwisconsin.data"). Decision tree is a supervised learning algorithm that can be used to classify either discrete or continuous data. Although Python was chosen as the language of choice, it was difficult to find a proper standard library having an implementation of C4.5 algorithm. The version of C4.5 implementation that was used in this assignment is a work found on github at https://github.com/michaeldorner/DecisionTrees by Michael Dorner.

A decision tree is a rooted directed tree consisting of nodes representing an internal decision for testing of an attribute and leaves representing the resultant class. Splitting a node is decided by a set of algorithms like information gain, gini index, chi square etc. In this work C4.5 algorithm is chosen which a successor of ID3 algorithm. ID3 uses information gain as the tree splitting criteria and it does not apply any pruning procedure and do not handle numeric attributes or missing values. It builds a decision tree in a top-down way. At each node of the tree, only one property is evaluated based on maximizing information gain and minimizing entropy and the process is recursively followed until a sub-tree contains objects belonging to the same category. In contrast C4.5 is an improved version of ID3 extending the idea of information gain onto gain ratio while handling missing value and numerical continuous values as well. After the tree growing phase, an error based pruning is also performed on the tree. While ID3 is susceptible to over fitting problem, C4.5 overcomes this by implementing the bottom-up technique that is known as tree pruning.

In this C4.5 implementation, first the original data set (breast-cancer-wisconsin.data.txt) was loaded into pandas data frame including missing values, and the data set (698 total records) was split into training and test in the same ratio used on the research paper. However the used C4.5 algorithm did not support missing values on training data, though it supported missing values on testing data. Therefore the missing values on training data set were dropped to build the decision tree. Then the training data set was saved as a csv so that the library can load it and grow the decision tree by using entropy as the evaluation function. The build tree was then plotted using the in-built method to visualize the node distribution. Then the test data set was fed into the classify method to get the predicted values of the algorithm. Python's classification_report method was used to calculate precision, recall, f1-score, and support of the results. Overall accuracy score of the model is found to be around 90%. Above results are calculated without doing the tree pruning. On next stage, pruning is applied on the tree to represent a minimum entropy of 0.2 and the resultant decision tree was plot and observed to have minimized in size. The same procedure is followed on the pruned tree and the accuracy score of the model is observed to have reduced with pruning while increasing false positives and false negatives of the predictions.

When compared the results of the C4.5 with the logistic regression model, the latter was having a higher accuracy score and precision over the breast cancer data set when missing values are replaced with 1 in the data set. But the C4.5 algorithm was tested using a data set with missing values and it have given better results that logistic regression model. Below are the best values observed in each model for each parameters.

	Logistic Regression	C4.5
Precision, recall, f1 benign class	0.98, 0.92, 0.95	0.95, 0.90, 0.92
Precision, recall, f1 malignant class	0.86, 0.96, 0.91	0.82, 0.90, 0.86
Accuracy score	0.934	0.90