Detecting Breast Cancer Type

1. Explain the problem you are trying to solve and why it is important.

   Cancer is a disease that is fatal but if it is detected early has a higher chance of survival for patients. The problem at hand, we need to create an algorithm that help detect type of cancer according to the classes classified to help the administering the correct diagnosis.

2. Briefly summarize (for a non-data scientist) how each algorithm works (no more than one paragraph each).

   SVC: This is a class in SVM module. In our problem we used linear SVC and the objective of a Linear SVC (Support Vector Classifier) is to fit to the data provided and returns a "best fit" hyperplane that divides, or categorizes, provided data. After getting the hyperplane, feed some features to the classifier to see what the "predicted" class is this will make algorithm rather suitable in our case.

   Decision Tree Algorithm is belongs to the supervised machine learning models. That means data is provided and the models learns from it. Decision tree tries to classify an object by eliminating other possibilities that least describes the object and finally a decision is made by what is left e.g. if a toaster is the object one can eliminate possibilities of it being a sandwich, none-metallic, uses electricity etc.

   MultinomialLogisiticRegression can be said to be an extension of logistic regression which adds support for multi-class classification problems. Unlike other extensions, this model doesn't require the problem to be transformed into multiple binary classification problem but it changes the loss function to cross-entropy loss and predict probability distribution. Entropy is just the number of bits required to transmit a randomly selected event from a probability distribution.

3. Briefly describe what you did to tune to each model and how the hyper-parameters you tuned impact that model.

   In all the developed models, gridSearchCV was used. GridSearchCV is a very common hyper-parameter selection technique, and the benefit of this tuner is that it searches exhaustively and it create a certainty that each combination has been compared. It can further fine-tune a model when a small search space is defined.

4. Include the results in well-formatted tables – both the values and the standard deviations of those values

SVC

1-	0.20918367
2-	0.23469388
3-	0.35672515
4-	0.35672515
5-	0.59064327
6-	0.21052632
7-	0.61988304
8-	0.61988304
9-	0.41520468
10-	0.18970588
Mean	0.38031740766482025
Standard Deviation	= 0.1664347929133342

MLR:

1-	0.9744898
2-	1.0
3-	0.97076023
4-	0.97076023
5-	1.0
6-	0.70760234
7-	0.94152047
8-	0.9122807
9-	0.94152047

10-	0.97205882
Mean	0.9390993063892225
Standard Deviation	0.08129159955019517

Decision Tree

1-	0.41326531
2-	0.74489796
3-	0.73684211
4-	0.64912281
5-	0.73684211
6-	0.15204678
7-	0.21052632
8-	0.32748538
9-	0.9122807
10-	0.13382353
Mean	0.5017132993548297
Standard Deviation	= 0.272136523489935

5. For each metric, determine which model is the best and whether or not its performance is statistically significantly better than any of the other models ($p < 0.05$)

Assuming that we conducted this test with a significance level of $p < 0.05$, we can reject the null-hypothesis that two models perform equally well on this dataset, since the p-value ($p < 0.005$) is smaller than p.

The best model is MLR with an accuracy of 96%

And it is statistically significant than the other models
Decision Tree vs MultinomialLogisticRegression
F statistic: -0.446
p value: 0.657

MLR vs SVC
F statistic: 1.080
p value: 0.283

Decision Tree vs SVC
F statistic: 0.555
p value: 0.580