|  |  |
| --- | --- |
|  | **Introduction to**  **Business Analytics [MSBA]** |

|  |  |
| --- | --- |
| **Homework #2 Part 2** |  |

\_\_\_\_\_Qiyu Wang\_\_\_\_

(put your full name above (incl. any nicknames))

Note: This is an **individual** homework. Discussing this homework with your classmates is a **violation** of the Honor Code.

If you borrow code from somewhere else, please add a comment in your code to make it clear what the source of the code is (e.g., a URL would sufficient). If you borrow code and you don’t provide the source, it is a violation of the Honor Code.

Total grade: \_\_\_\_\_\_\_ out of \_\_\_70\_\_\_ points

***ATTENTION: HW2 has two parts. Please first complete the Quiz “HW2\_Part1” on Canvas. Then, proceed with Part 2 in the following page. You will need to submit (a) a PDF file with your answers and screenshots of Python code snippets as well as Rapidminer repositories and (b) the Python code and Rapidminer repositories.***

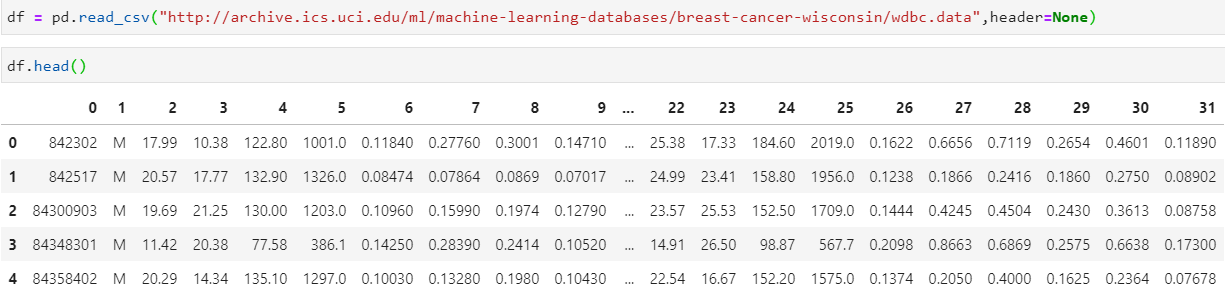
**(70 points) [Mining publicly available data. Please implement the following models with both Rapidminer and Python]**

**Please use the dataset on breast cancer research from this link:** [**http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data**](http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data) **[**Note: Rapidminer can import .data files in the same way it can import .csv files. For Python please read the data *directly from the URL* **without** downloading the file on your local disk.**] The description of the data and attributes can be found at this link:** [**http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names**](http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names)and is also provided as in the appendix of this homework assignment.

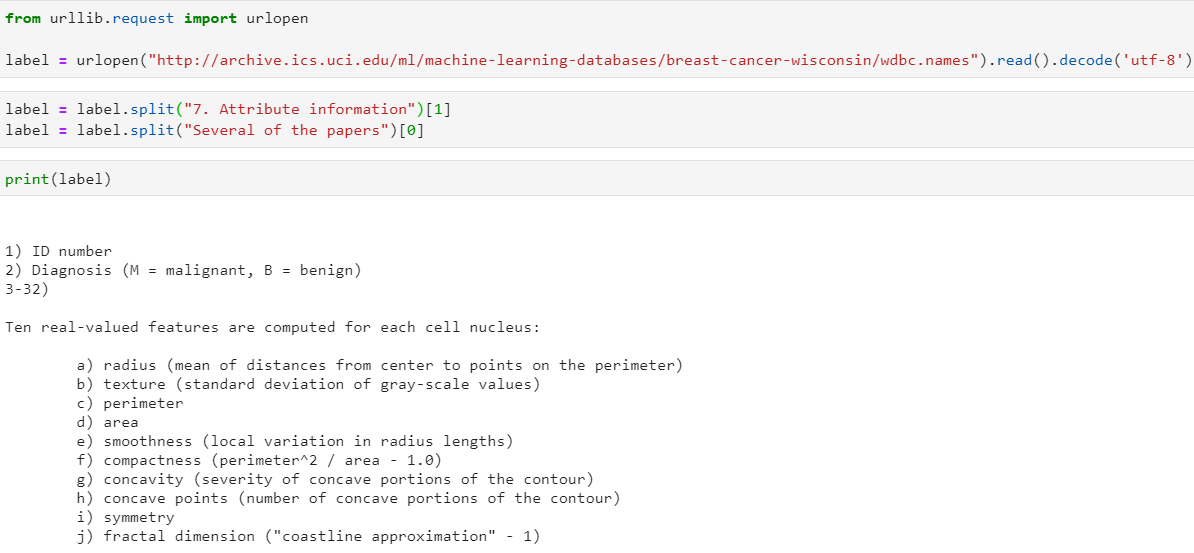
**Each record of the data set represents a different case of breast cancer. Each case is described with 30 real-valued attributes: attribute 1 represents case id, attributes 3-32 represent various physiological characteristics, and attribute 2 represents the type (benign B or malignant M) .**

**50 Points (Python):**

1. **(10 points) Load the data. Then, explore the data by reporting summary statistics and a correlation matrix. Show your code.**

I first read in the dataset.

Then I print out the top 5 tows to see how it looks like.



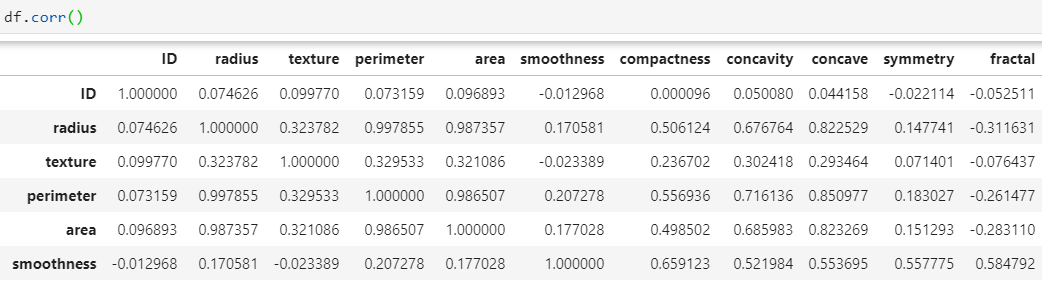
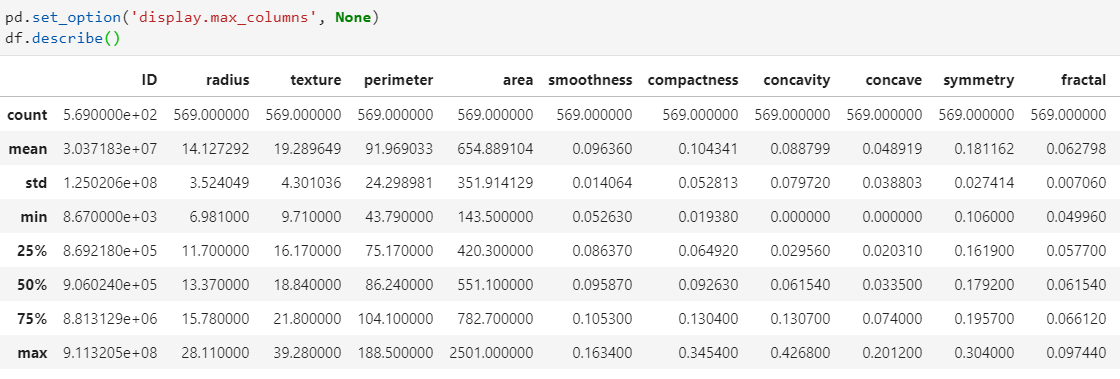
Then I read in the description of the dataset, and kept the lines containing labels and removed the rest



Then I used regex package to filter the extra letters and to keep only the labels, and create a list out of it.

Then I make the list of labels as the column names for the dataframe.

Finally, I created statistics summary table and correlation matrix.



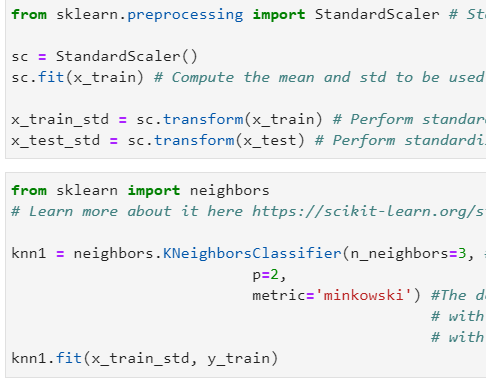
1. **(12 points) Perform a predictive modeling analysis on this dataset to predict the type (benign B or malignant M) using a k-NN technique (for k=3) and the Logistic Regression technique. Please be specific about what other parameters you specified for your models. Briefly discuss your modeling process (e.g., validation technique, any preprocessing steps, parameters used to build the models, etc.) and show your code. Report the estimated coefficients of the Logistic Regression technique.**



I first split the target variables and the rest.

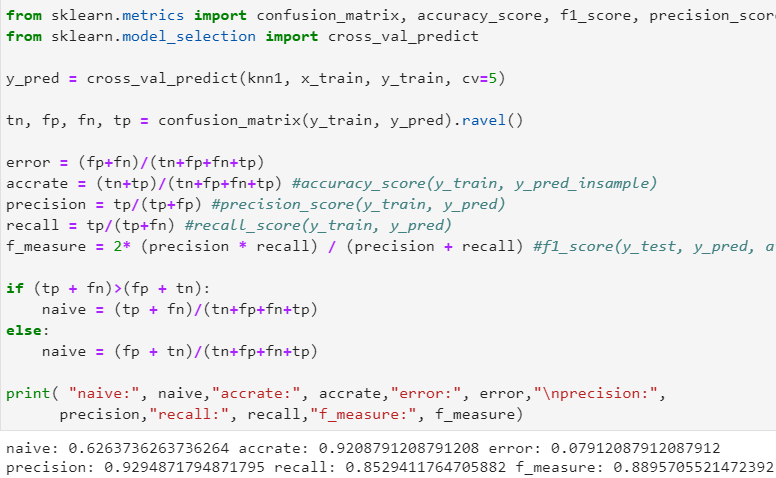
Then I split the test/train sets in a 80/20 manner, and I used stratify parameter to make sure the target: predictors ratios are same in both train/test sets.

And here, I checked the ratios in each set to make sure the splitter did its job.



Here, I standardized predictors from both train/test sets, because KNN and Regularizations for linear models are sensitive to scales.

Here I fitted a 3NN with Euclidean distance(p = 2, and metric = minkowski) with standardized train set.



Here, I applied 5-fold cross validation to test the performance on validation set, and got a accuracy of 92.08%.



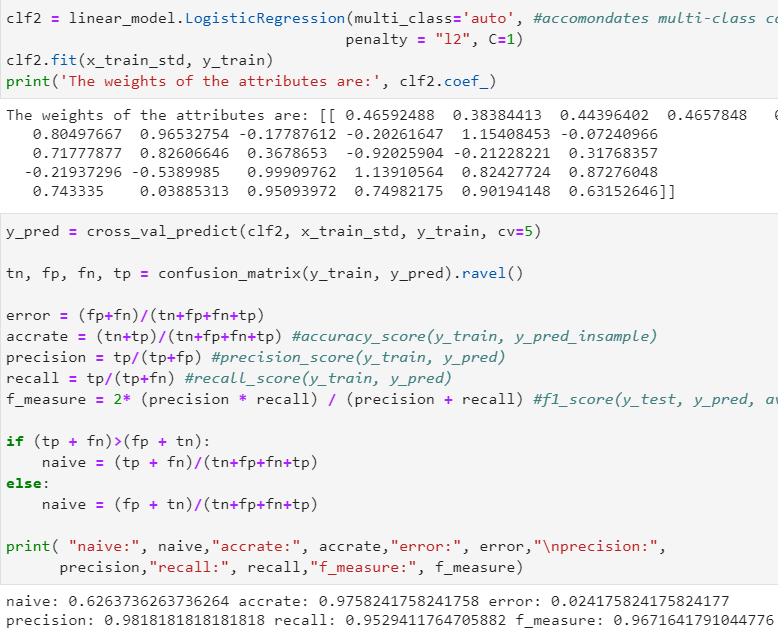
Then I train a second 3NN model using Manhattan distance (p = 1, metric = minkowski) with standardized training set.

Again, I applied 5-fold cross validation to test the performance on validation set, and got a accuracy of 92.47%, which is slightly better than former one.



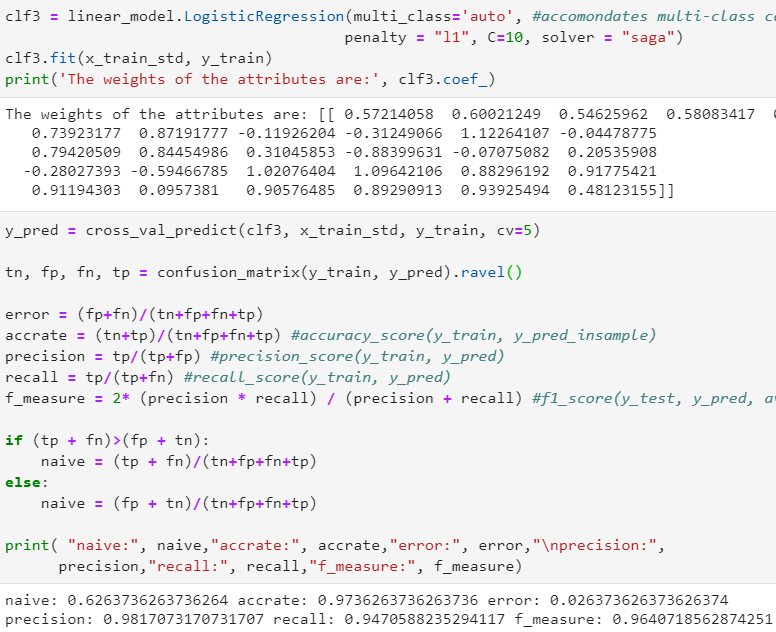
Here I trained a logistic basic regression with all variables available from the unstandardized training set. The c=1e5 meaning we are doing a plain logistic model without regularization. The coefficients are reported at this stage too.

I applied 5-fold cross validation to test the performance on validation set, and got a accuracy of 94.51%, which is better than 3NN’s.



Here I trained a logistic basic regression with all variables available from the standardized training set, using L2 ridge regularization. The coefficients are reported at this stage too.

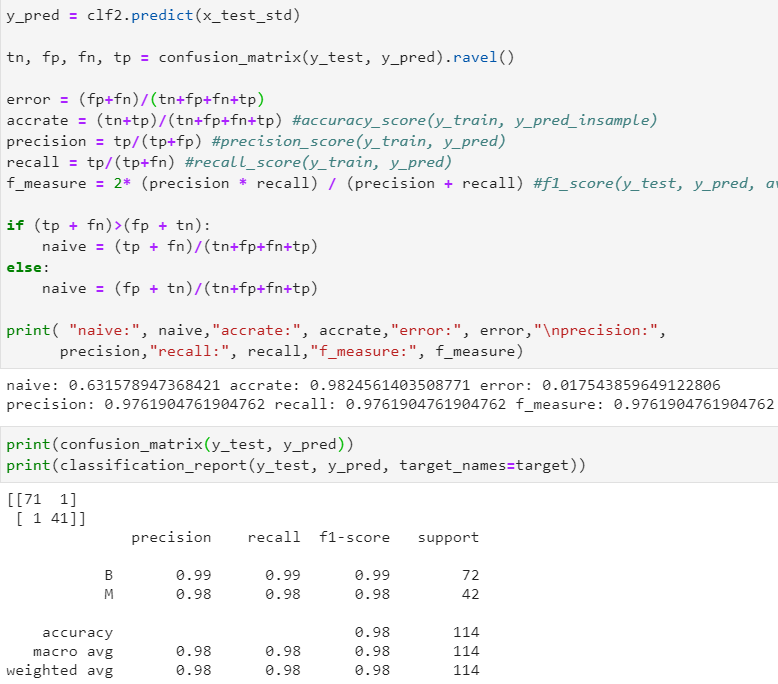
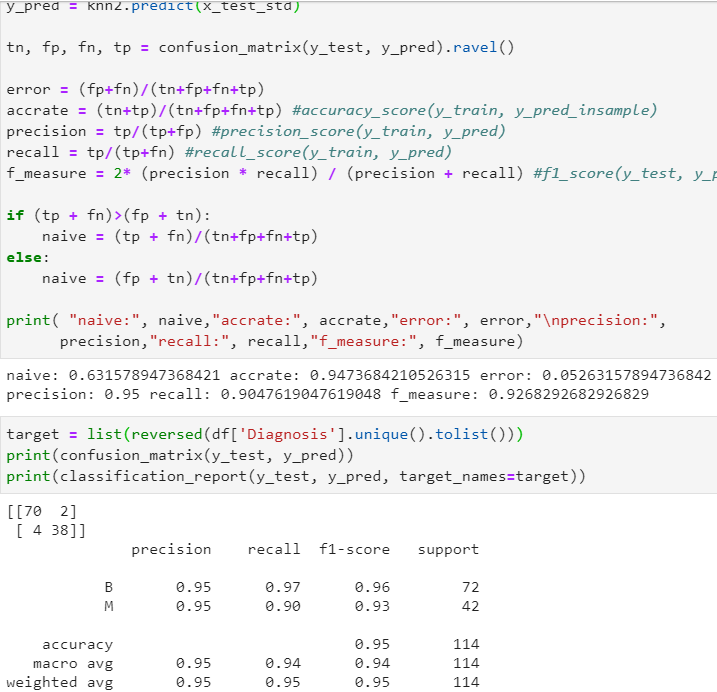
I tested multiple value on c\_parameter and when c = 1it gets best performance on 5-fold cross validation, and got a accuracy of 97.58%, which stands out as the best model so far.



Then, I trained a logistic basic regression with all variables available from the standardized training set, using L1 lasso regularization, which provides a automatic feature selection.

I tested multiple value on c\_parameter and when c = 10 it gets best performance on 5-fold cross validation, and got an accuracy of 97.36%, which is slightly weaker than the ridge regression.

1. **(13 points) Compare the generalization performance of the k-NN model with the Logistic Regression model. Make sure you report the confusion matrix, the predictive accuracy, precision, recall, and f-measure. Briefly discuss the results and show your code.**

Logit 3NN

Based on the performance metrics, 3NN model with Manhattan distance reported a accuracy of 94.7%, where has a precision of 95% for both M/B, and a recall of 97% for B and 90% for M. Logit model with Ridge regularization on the other hand, has a higher accuracy of 98.2%, where has a precision of 99% for B and 98% for M, and a recall of 99% for B and 98% for M. According to F-measure, we see that logit has a better capture of B/M than 3NN does. From those metrics we can see that logit model has an overall better performance than the 3NN model. However, both models performed pretty well, and way more accurate than the Native accuracy (benchmark).

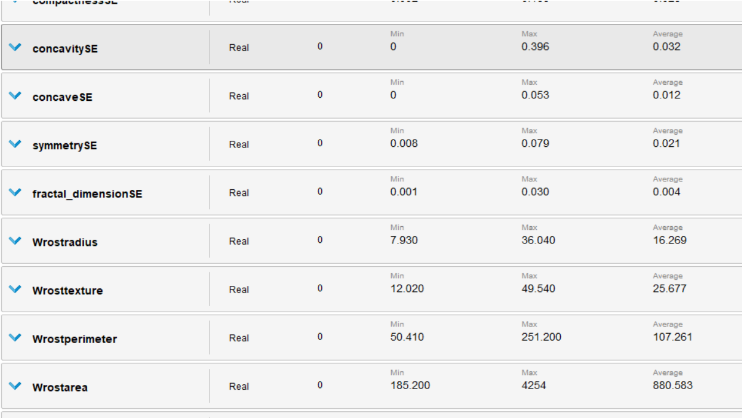
1. **(15 points) What generalization performance metric would you prefer to use in order to choose the best performing model in this context and why? Please be clear about any assumptions you might make when you choose the generalization performance metric you would prefer.**

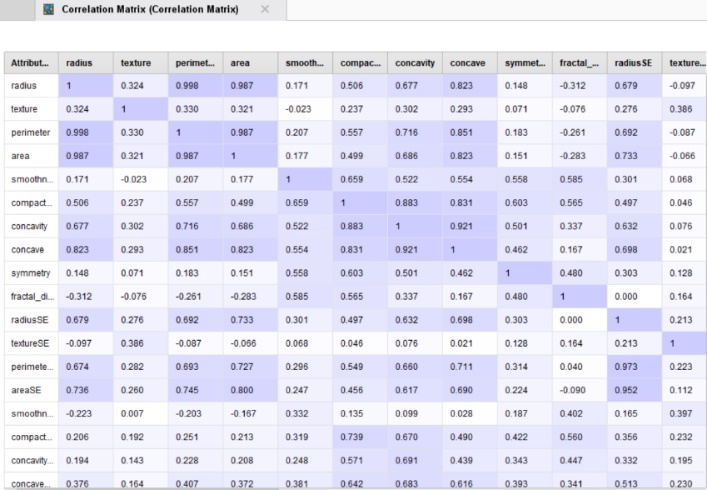
For this particular problem, which involve diagnose of cancer (M/B), I believe that recall(M) is the most important metrics here. Because it can get extremely risky and costly to miss a true malignant, which may risk a person’s life. By having a high recall score, we can make sure that we don’t make as less as possible mispredictions of true malignant, which can delay a patient’s treatment and reduce its chance of recovery. So for here, the logistic regression had a higher recall on malignant, thus, logistic model presents as the best model here.

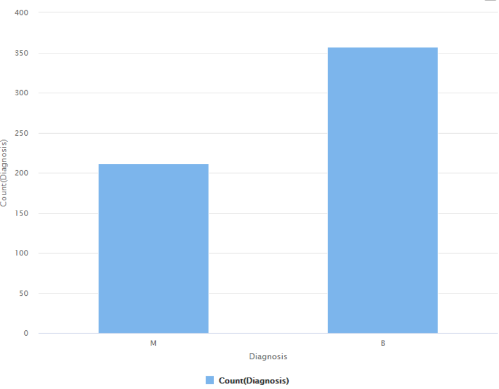
In addition to recall(M), I would say precision scores and accuracy are quite good metrics as well, where a high precision/accuracy scores might reduce unnecessary waste of medical resources and patients’ medical cost. But they are not as important as recall(M).

**20 Points (Rapidminer):**

**Perform a predictive modeling analysis on this dataset to predict the type (benign B or malignant M) using a k-NN technique (for k=3) and the Logistic Regression technique. Compare the generalization performance of the k-NN model with the Logistic Regression model. Make sure you report the confusion matrix, the predictive accuracy, precision, recall, and f-measure.**

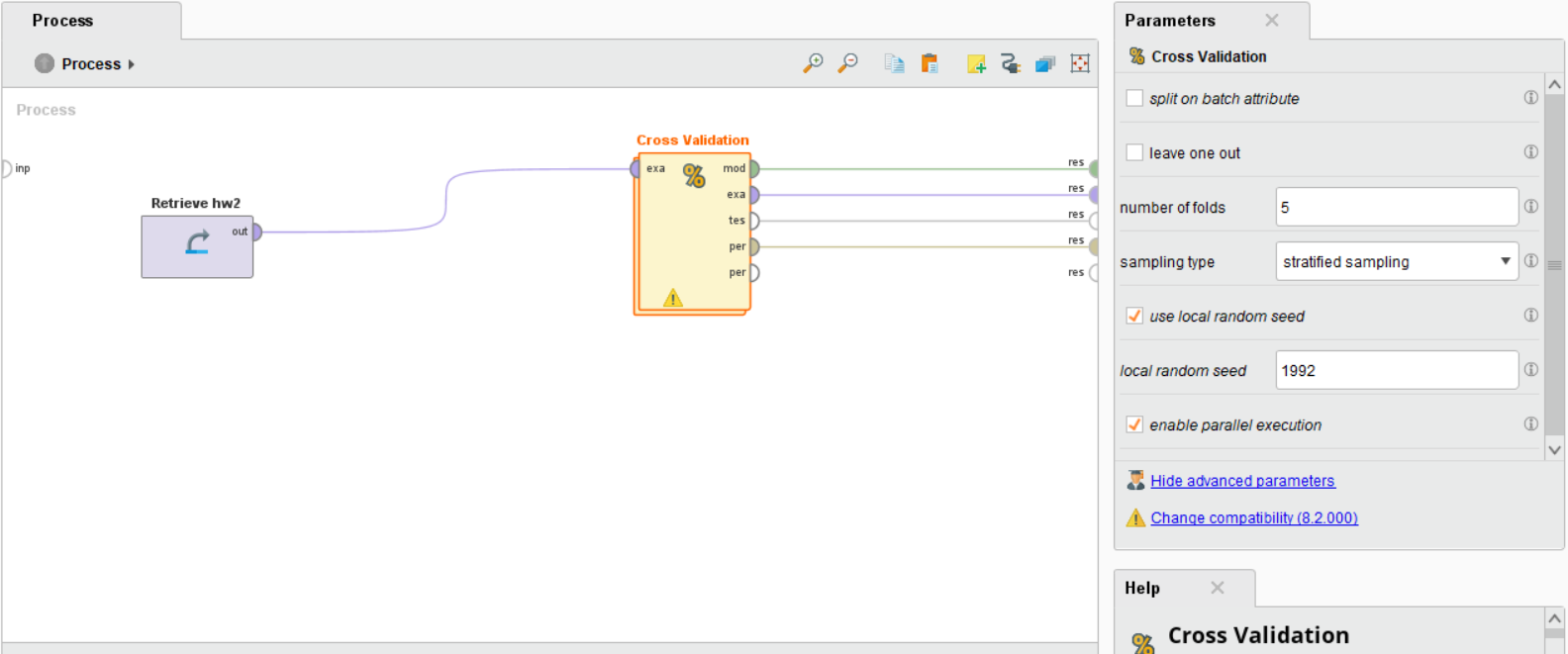
1. **[20 points] Please show below screenshots of the models you have built using Rapidminer, the results, and the parameters you have specified.**
   1. **[8 points] Data Preview**
      1. Screenshot of the summary statistics table in Rapidminer.

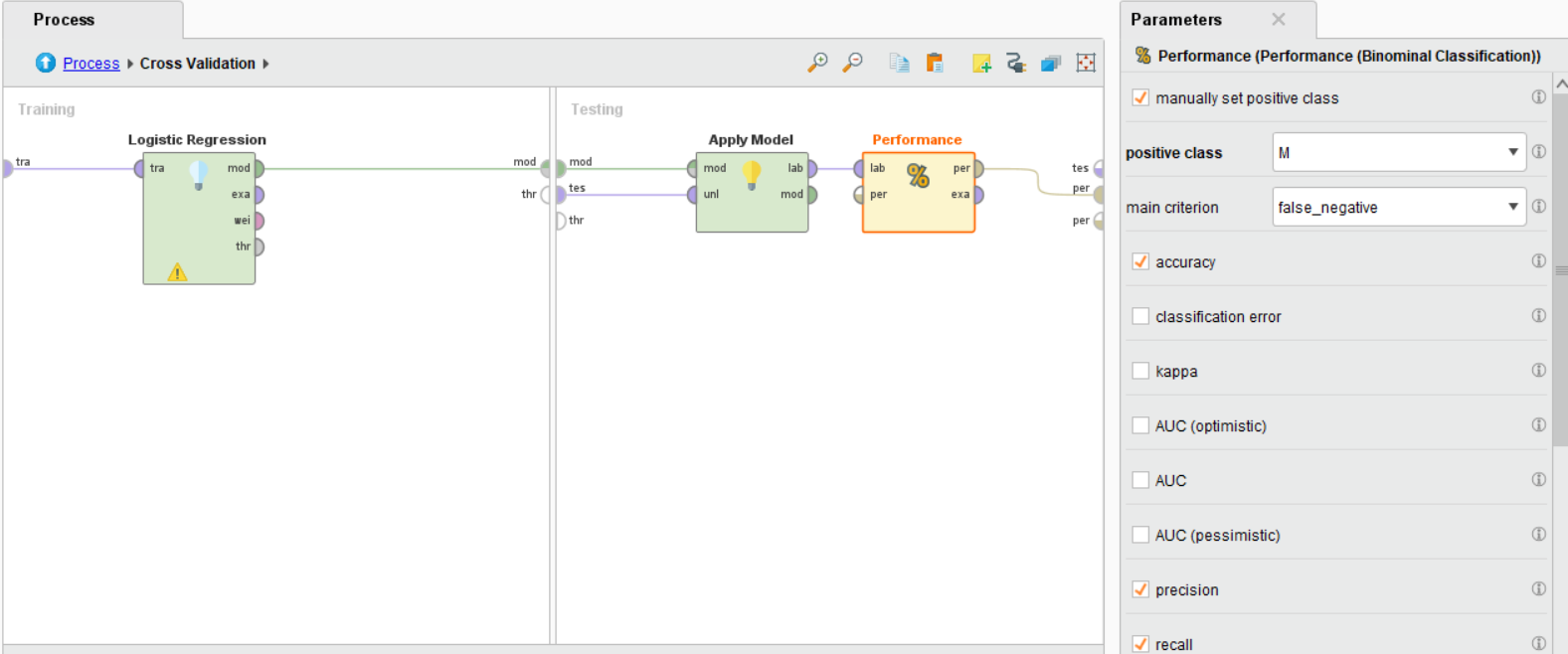




* 1. **[8 points] Logistic Regression**
     1. Screenshots for Logistic Regression Model Setup (Rapidminer Processes)

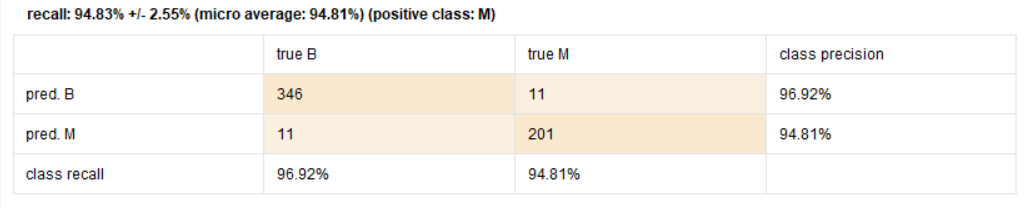
(Insert Shreenshots here – 2 screenshots are expected here; one for the upper layer and one inside the validation technique)





* + 1. Screenshot for Logistic Regression Performance

(Insert Shreenshot here)



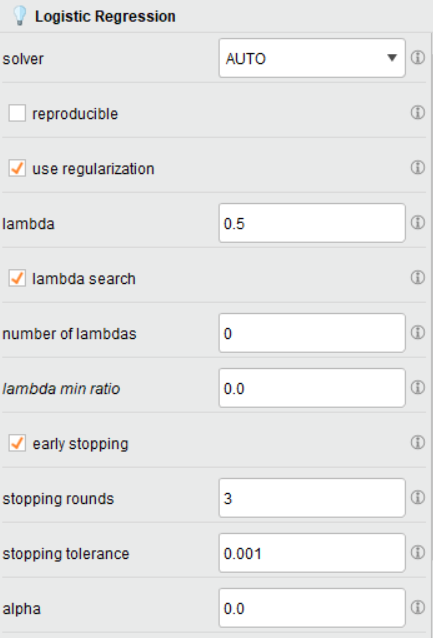
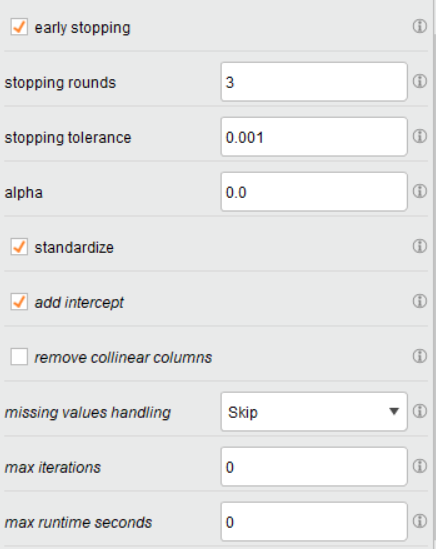
* + 1. Screenshot for Logistic Regression Results (Coefficients)

(Insert Shreenshot here)

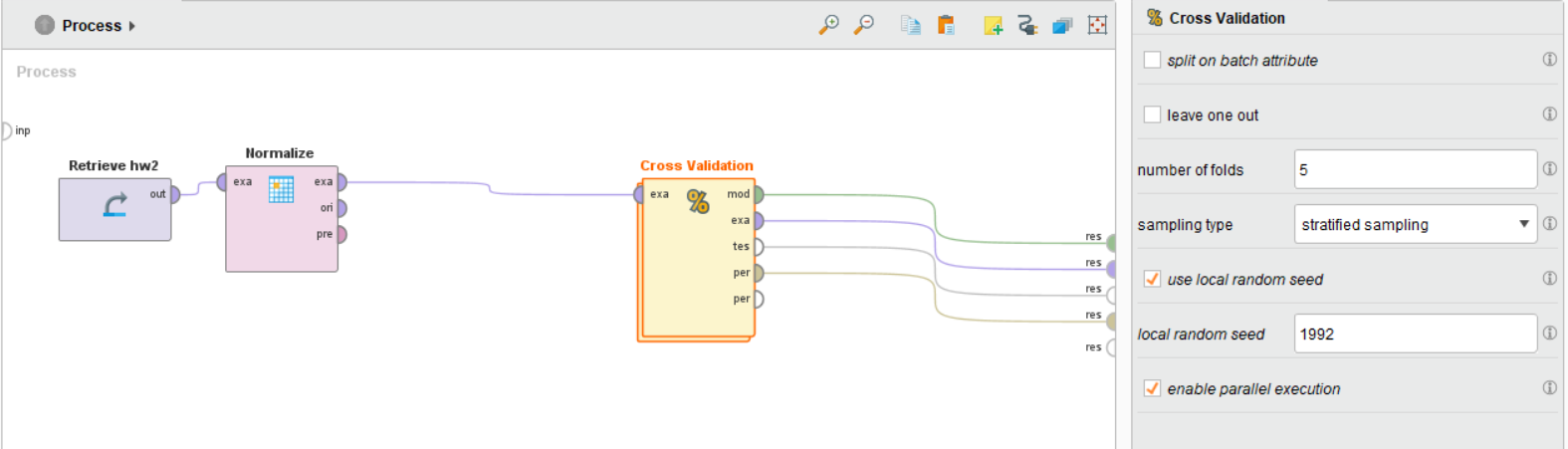
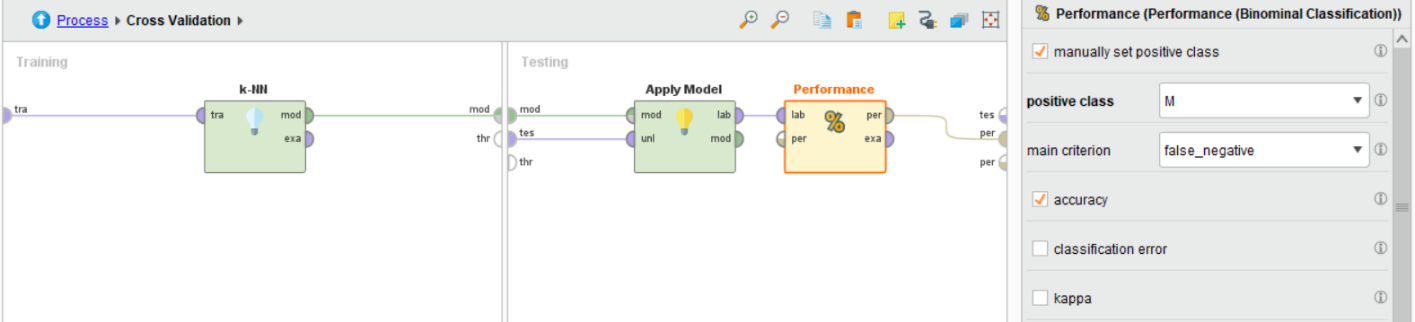
* + 1. Screenshot for Logistic Regression Rapidminer Operator Parameters (click on Logistic Regression operator and then take a screenshot of the Parameters window on the right)

(Insert Shreenshot here)

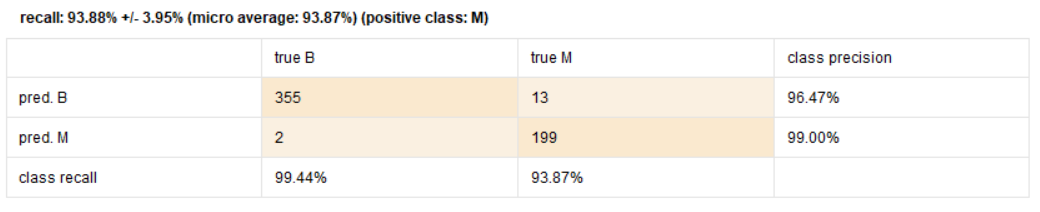
* 1. **[8 points] kNN**
     1. Screenshots for kNN Model Setup (Rapidminer Processes)

(Insert Shreenshots here– 2 screenshots are expected here; one for the upper layer and one inside the validation technique)



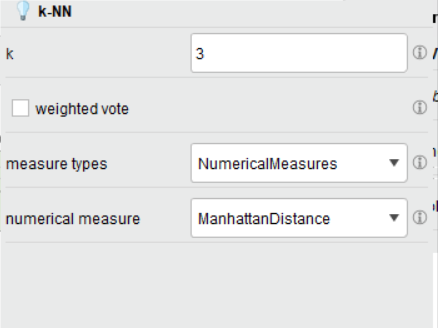
* + 1. Screenshot for kNN Performance

(Insert Shreenshot here)



* + 1. Screenshot for kNN Rapidminer Operator Parameters (click on kNN operator and then take a screenshot of the Parameters windows on the right)

(Insert Shreenshot here)



**Comparison between two models:**

Based on the performance metrics reported, we clearly see that 3NN model has a higher recall score on malignant. As we discussed, in the cancer diagnosis context, misprediction on true malignant is very risky for patients. Therefore, 3NN model is doing better for this particular problem.

Moreover, the 3NN model has a overall better capture of M/B, and overall better accuracy, which makes 3NN model a overall better model than the logistic model based on our RapidMiner reports.

**Appendix (Data Description)**

1.Title: Wisconsin Diagnostic Breast Cancer (WDBC)

Results:

- predicting field 2, diagnosis: B = benign, M = malignant

2. Number of instances: 569

3. Number of attributes: 32 (ID, diagnosis, 30 real-valued input features)

4. Attribute information

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three

largest values) of these features were computed for each image,

resulting in 30 features. For instance, field 3 is Mean Radius, field

13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

5. Missing attribute values: none

6. Class distribution: 357 benign, 212 malignant