BDA500 A3 Report

I declare this is my own work in accordance with Seneca College’s Academic Policies.  
No part has been copied manually or electronically from any other source (including  
web sites) or distributed to other students.  
Name: Jeff Sit Student Number: (152674198)

The stage before any of the steps starts with looking at the breast cancer dataset. The dataset contains over 500 rows and 30 feature columns such that the objective is to train a machine learning model that can accurately predict the target variable based on the features provided. The performance of this initial model will be evaluated, such that after, principal component analysis will be applied. The same steps will then be applied again to compare the differences in performance and metrics between multiple variants of the same data.

**Step 1** is the data preparation phase. Python libraries are imported and loaded while the data is loaded into a DataFrame using pandas and its built-in function “read.csv”. Additionally, there is no missing data and df.info() is used to confirm this.

Text

Description automatically generated

Next, the features and target are split for later use. The features specifically are scaled to ensure the numerous feature columns are within a comparable magnitude. With a separate dataframe containing the scaled data, the data preparation stage is now complete.  Text

Description automatically generated

**Step 2 and 3** is where the various classification models are first built. However, data first needs to be split into training and test data.

Text

Description automatically generated

While building and training each model, a timer starts and stops before and after each process to gather the training time required for each model type. This is done once per model. Text

Description automatically generated

**Step 4** simply evaluates the models created in the previous step. As a result, accuracy, precision, and recall are all assessed with the help of classification report.

A screenshot of a computer

Description automatically generated with low confidence

**Steps 5 to 8** repeat the steps performed from steps 2 to 4, but the data is modified in the way that PCA is applied to have the data retain a 95% variance rate.

**Steps 9 to 12** alsorepeat the steps performed from steps 2 to 4, but the data is now reduced to have 2 principal components.

**Step 13** visualizes the data from step 9, which plots 1 principal component against the other. Chart, scatter chart

Description automatically generated

**Steps 14 to 17** start with choosing the best performing model based on the original dataset. In this case, since logistic regression performed the best, a prediction is made using this model and its performance metrics are evaluated. Once this is complete, the same procedure is repeated on all logistic regression models per PCA affected datasets.

**Step 18** revolves around the writing of this report.

**Step 19)**

Looking at the logistic regression, k nearest neigbours, and support vector classification methods, on the surface it does not appear that there is a very significant difference. This is especially apparent when looking at the times to train, where the times are already so fast, they only differ in the 3rd or later decimal places. This is also applicable to the classification report metrics too, where all of the accuracy, precision, and recall values regardless of model were all extremely high across the board.

However, based on time, the difference became quite apparent between the SVC model and the others as the SVC model took over a second to train. While 1 second does not sound like a lot, relative to the speeds of the other 2 models, this is taking roughly a hundred to a thousand times as long as the others are to train. Thus, if the size of the dataset were to increase into the hundreds of thousands of rows, or possibly even millions, the decision of what type of model to train would likely be a much larger point of contention, as it should be a major priority to ensure the machine learning model is trained as efficiently as possible.

**Step 20)**

Based on the experiment performed, PCA will not always speed up training procedures. The extent of change in training time appears to depend on the machine learning algorithm. As an example, for the SVC model, times always decreased, regardless of how the dataset was being reduced. However, time was decreased more when reducing the dataset to 2 principal components, than when using the 95% variance dataset. Compare this to the logistic regression dataset, and times actually only increased as PCA was applied, although not substantially. The increases in time for the logistic regression model occurred in the 4th decimal place and beyond.

**Conclusion)**

Overall, PCA is an excellent tool to help in reducing the training times required by classification models. However, it should not be applied carelessly however as in some cases, PCA can produce the opposite effect, hence increasing training times. Although the times represented in this experiment differ insignificantly, the findings remain the same, and should therefore be applied responsibly. This is especially true during cases where the number of instances in the dataset begin to number far greater than just the hundreds.