**TANIYA PAL (s373094)**

**MODULE 5 ASSIGNMENT**

**Abstract**

Machine learning has developed over the years and has a wide range of applications, mainly in the area of analysis of big data. With machine learning, it is possible to accurately predict models of big data. We had been given four datasets consisting of data from E-nose, HPLC instruments as well as sensory scores and bacterial counts. An analysis was performed over the datasets and the algorithm which gives the best prediction was determined for each dataset.

**Introduction**

Several algorithms were used for the analysis. They are namely, K nearest neighbors, Support Vector machine, Random Forest and Partial Least Squares.

**K nearest neighbors**

The algorithm works by computing the distance between all the samples in the training and test dataset, giving in return k nearest neighbors or closest samples. Scaling is important for the algorithm to work accurately. It works well for both classification and numerical data giving back predictions with high accuracy, and it is also fast to train.

Though, knn is highly accurate it has some disadvantages too. It is computationally expensive, taking a lot of time. Also, depends on the suitable selection of k for it to predict accurately. Skewed datasets does not return accurate predictions.

K values ranges from 2 to 20. First, it calculates the similarity between the samples using the Euclidean or Manhattan distance function, then uses this information for predicting the similarity between new datasets. Takes the k closest samples and gives back the predictions of classes after considering the majority vote from the training data.

It is called lazy learning algorithm as it does not create any model, only measures similarities and distances between the samples. At the same time, it is powerful as it tries to find natural patterns in our model, instead of squeezing ad fitting the model.

Some of the applications of knn are pattern identification for genetic data, facial recognition, medical imaging, recognizing handwriting and many more.

**Support Vector Machines (SVM)**

Creates a liner decision surface also termed as hyperplane, between the classes, and separates the one which has largest distances. The classes that falls between the borders are called support vectors. The algorithm works equally well for binary and multiclass problems.

When the dataset is linearly separable, nit finds a hyperplane that separates the data linearly. It finds two more hyperplanes that consists of the support vectors. No samples should fall in the space between the hyperplanes. It maximizes the distance between the hyperplanes and minimizes the Euclidean norm.

When the dataset is nonlinearly separable, the algorithm allows some sample to fall in the spaces between the hyperplanes, now called as soft margins. The dataset is converted into a 3D dataset from a 2D one, in order to make it linearly separable.

In both the above cases, a cost value is applied when the algorithm violates the constraints, and this cost value is minimized by the algorithm in order to get an accurate prediction.

**Random Forest**

Random Forest, a supervised learning algorithm, functions by building multiple decision trees, in order to get more accurate prediction by merging them together. Can be used for both classifier and regression problems. Since it includes multiple decision trees in the model, an additional randomness is added to the model as a result. It investigates a subset of features instead of searching for the most important feature, thus adding diversity which results in a better model. Also, we can also have a look up at the importance of the features and this helps to drop the features which don’t contribute enough to the model.

**Partial Least Square regression**

This algorithm builds a linear regression model by projecting the predicted variables and observable variables to a new space, it does not build a hyperplane between response and independent variables. It finds fundamental relationship between latent variables X and Y. This type of regression is suitable when observations exceeds matrix of predictors, a multicollinearity among X values is observed.

**Results:**

**Objective 1: Building classification model with Enose and sensory data**

Analysis with k nearest neighbors algorithm:

R packages used for this analysis are:

* “CLASS”
* “GMODELS”
* “CARET”
* “RPART”
* “MIXOMICS”
* “RPART.PLOT”

**Model tuning**

Accuracy of the knn model was tested over values of k ranging from 1 to 20. The accuracy values from the confusion matrices was plotted against each k value.

Chart, scatter chart

Description automatically generated

From the plot, the highest accuracy was observed at k with a value of 16. So, the knn model was generated with k =16.

**Generating the model**

Chart

Description automatically generated

On generating the model over 100 iterations, the accuary eas seen the highest at around 50th iteration. The highest accuracy was achieved around 0.8.

Chart

Description automatically generated

The plot shows the cumulative mean accuracy varies over the 100 iterations.

Chart, histogram

Description automatically generated

The number of misclassified samples were plotted for each iteration. The number of misclassified samples were beyond 20.

**Analysis wit Support Vector machine algorithm:**

**Model tuning**

The model tuning for SVM was not done and the analysis was done with a cost (C) value equal to one.

**Generating the model:**

Chart

Description automatically generated

After 100 iterations, the accuracy was seen the highest around 80th iteration. The maximum accuracy achieved is around 0.6.

Chart

Description automatically generated

The cululative mean accuracy was nearly the same for all 100 iterations.

Chart, histogram

Description automatically generatedThe highest number of misclassified sample was seen around 85th iteration. The maximum number of misclassified samples is around 20.

**Analysis with Random Forest algorithm:**

Chart, histogram, scatter chart

Description automatically generated

It can be seen from the plot that the error rate has not stabilized at the end of 500 trees. We can rectify this, by increasing the number of tress further.

Chart, histogram

Description automatically generated

The plot highlights the importance of the variables or features in prediction of the model.

**Conclusion:**

We can conclude at the end of the analysis, both SVM model and knn model performs the same in predicting the model as the accuracy of the knn model was higher than that of svm model over 100 iterations. At the same time, the number of misclassified sample was less for svm model than the knn model. The performance of the model also depends on the selection of the hyperparameters so a different algorithm might work the best for the dataset at different hyperparameters.

**Objective 2:Building classification model with HPLC and sensory data**

**Analysis with knn algorithm**

A plot was generated between the accuracy generated over values of k ranging from 1 to 20.Chart, line chart

Description automatically generated We see that the highest accuracy can be achieved at k=6.

A picture containing text, antenna

Description automatically generatedAfter 100 iterations, it was seen that the model gives the highest accuracy over 50th iteration. The highest accuracy achieved is around 0.8.

Chart, histogram

Description automatically generated

The plot shows that the cumulative mean accuracy varies over 100 iterations.

Chart, histogram

Description automatically generated

The maximum number of misclassified sample is beyond 20.

**Analysis with support vector machine algorithm:**

Chart

Description automatically generated

The highest accuracy was achieved around 80th iterations with a value of 0.6.

Chart

Description automatically generated

The cumulative mean accuracy was equal over 100 iterations.

Chart, histogram

Description automatically generated

The number of misclassified samples was beyond 20.

**Analysis with Random Forest algorithm:**

Chart, scatter chart

Description automatically generated

It can be seen from the plot that the error rate has not stabilized at the end of 300 trees. We can rectify this, by increasing the number of trees further.

**Chart, bar chart

Description automatically generated**

The plot highlights the importance of the variables or features in prediction of the model.

**Conclusion**

The knn model has the highest accuracy value so we can conclude it is the best suited model for this dataset.

**Objective 3: Building regression model with Enose and bacterial count data for total viable count and Pseudomonads data**

**For TVC Data:**

**Analysis with knn algorithm**

**Chart, scatter chart

Description automatically generated**

The RMSE was lowest at k=3. So, the analysis was done at k=3.

Chart, line chart

Description automatically generated

The actual values of the test dataset and the predicted values of the model for one iteration were plotted against each other and it gave a low RMSE value. So, it confirmed k with a value of 3 is a good choice.

Chart, line chart

Description automatically generated

After 100 iterations, the same plot and RMSE value was observed. The low RMSE value confirmed the accuracy of the model was high.

**Analysis with Random Forest algorithm**

Chart, histogram

Description automatically generated

The plot shows the error does not vary much over the number of trees.

Chart, line chart

Description automatically generated

The low RMSE value confirmed the high accuracy of the model when the actual values of the test dataset and the predicted values of the model were plotted together.

**Analysis with Partial least squares algorithm**

Chart, line chart

Description automatically generated

The error was seen lowest at number of components equal to one. So, the model was generated with one principal component.

Chart, line chart

Description automatically generated

The RMSE value was low but higher than the previous two models.

**Conclusion:**

Thus we can conclude that knn is the best prediction model for this dataset as it has the lowest RMSE value.

**For Psedomonad data**

Chart, scatter chart

Description automatically generated

We see that the accuracy is highest at k equal to 2. So the analysis was done with k equal to 2.

Chart, line chart

Description automatically generated

A RMSE value of 0.83 of the predicted values was observed after prediction of the model over 100 ietartions.

Random Forest

Chart, scatter chart

Description automatically generated

The RMSE value was observe lowest at a value of mtry equal to 11. So, the analysis was done with mtry equal to 11.

Chart, line chart

Description automatically generated

After 100 iterations, the RMSE of the predicted values was observed to be approximately 0.53.

**Analysis with Partial least squares algorithm:**

Chart, line chart

Description automatically generated

The error was observed lowest at number of principle components equal to one. So, the analysis was done with one principal component.

Chart, line chart, histogram

Description automatically generated

A high RMSE value was observed for the predicted values for this model.

**Conclusion**

Thus, we can conclude that random forest is the best suitable for this dataset as it has the kowest RMSE value.

**Objective 4:Building regression model with Enose and bacterial count data for total viable count and Pseudomonads data**

**For TVC Data:**

**Analysis with knn algorithm:**

Chart, scatter chart

Description automatically generated

The RMSE for the k values was lowest at k equal to 14, so the analysis was performed with k equal to 14.

Chart, line chart

Description automatically generated

A RMSE value of 1.3 was observed for this model.

**Analysis with Random Forest algorithm**

Chart, scatter chart

Description automatically generated

We see the RMSE value is lowest at mtry equal to 8. So the analysis was done with mtry equal to 8.

Chart, line chart

Description automatically generated

A RMSE value of 1.05 approximately was observed after prediction of the model.

**Analysis with Partial least squares algorithm**

Chart, line chart

Description automatically generated

The RMSE was lowest at number of principal component equal to one so the analysis was done with the same value.

Chart, line chart

Description automatically generated

A RMSE value of 1.57 was observed at the end of the prediction of the model.

**Conclusion:**

Thus, we can conclude knn model is the best suited for this dataset has the predicted model has the lowest RMSE value.

**For Pseudomonads data:**

**Analysis with knn algorithm**

Chart

Description automatically generated

The analysis was performed with k equal to 14 as it had the lowest RMSE value.

Chart, line chart

Description automatically generated

The RMSE value for the model was found to be approximately 1.21.

**Analysis with Random Forest algorithm**

Chart, scatter chart

Description automatically generated

The analysis was performed with mtry equal to 8 as it had the lowest RMSE value.

Chart, line chart

Description automatically generated

A RMSE value of approximately 1.15 was obtained at the end of the prediction.

**Analysis with Partial Least Square algorithm:**

Chart, line chart

Description automatically generated

The RMSE value was found to be lowest at number of components equal to 2. So, the analysis was performed with two principal components.

Chart, line chart

Description automatically generated

A RMSE value of approximately 3.11 was obtained at the end othe prediction.

**Conclusion:**

Thus, we can conclude knn model is the best suited for this dataset has the predicted model has the lowest RMSE value.