CAIRO UNIVERSITY

FACULTY OF ENGINEERING

HEMN454 – Data Mining and Machine Learning in Healthcare

Task 3 – Support Vector Machine

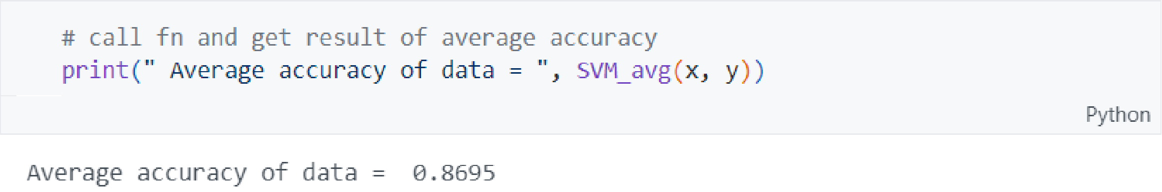
Amira Mahmoud Farid 1170498

Support Vector Machine

Code Description

Non – Scaled Data

1. Read Data
2. Assign "Foreign Worker" column to y "y_data"
3. Drop "Foreign Worker" from data and assign data to x "x_data"
4. Create SVM function
   Input: x_data, y_data
   Return: test accuracy
   4.1. Split the dataset into 0.6 training and 0.4 testing
   4.2. Train the SVM classifier with linear kernel
   4.3. Return the testing accuracy
5. Create average SVM function
   Input: x_data, y_data
   Return: average accuracy over 10 accuracies
   5.1. Repeat training process 10 times, save results in array, and take average accuracy of the array of 10 accuracies
6. Call function and get result of average accuracy

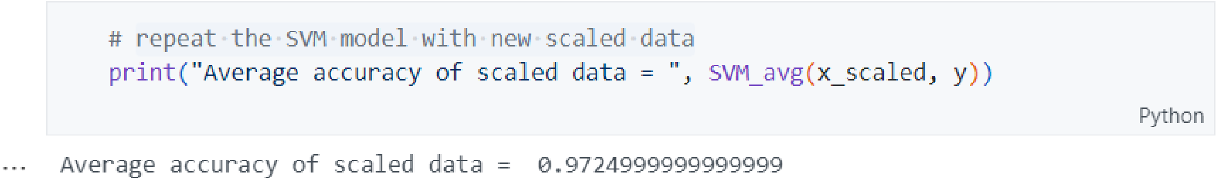


```
# call fn and get result of average accuracy
print(" Average accuracy of data = ", SVM_avg(x, y))
```

```
Average accuracy of data =  0.8695
```

Scaled Data

1. Pre-processing steps to standardize and normalize the dataset
2. Standardize features by subtracting the mean and scaling to unit variance
3. Normalization all dataset to range -1, 1
4. Repeat the SVM model with new scaled data and get average accuracy result



```
# repeat the SVM model with new scaled data
print("Average accuracy of scaled data = ", SVM_avg(x_scaled, y))
```

```
...  Average accuracy of scaled data =  0.9724999999999999
```

Conclusion

The scaled data is better than the non-scaled data. The accuracy of the scaled data is higher. Since the SVM considers the changes of 1 to be constant with respect to multiple features. the change of one is significant. therefore, the scaled data gives better accuracy.