**Data 602: Midterm: Responses to Exercises**

1. **Exercise # 1: Naïve Bayes Model**
   1. Question 1-3: **Answered within the code**
   2. Question 4. What does the cross-validation outputs tell us?
      1. **Answer:** Cross validation is a procedure used to avoid overfitting and estimate the skill of the model on new data. The cross validation output tells us how well the model is performing with each new dataset extracted from within the test dataset when it is trained on the other folds extracted from the test dataset.
   3. Question 5. What can you say about the model?
      1. **Answer:** The mean accuracy of the cross validation scores is 95.3% which indicates that the model is performing well and is not overfitted.
2. **Exercise # 2: SVC Model**
   1. Question 6-11: **Answered within the code**
   2. **Answer:** Based on the accuracy (0.97), precision (0.98) and a recall (0.97), the model is performing very well on the train set**.**
3. **Exercise # 3: Random Forest Regression Model**
   1. Question 12,14, 16-18 and 20: **Answered within the code**
   2. Question 13. Is there any problem with the data?
      1. **Answer**: Based on the count function and the heatmap shown in the notebook, that there is no missing data or NaN in the dataset, so we can conclude that the dataset is clean.
   3. Question 15. Use a random forest regression model. Does it perform better than a Ridge regression Model? Why?
      1. **Answer**: The random forest regression model performed better than the ridge regression model because it had a lower mean absolute error, lower mean squared error and a lower root mean squared error compared to the ridge regression model (see table below). Additionally, random forest uses bootstrap aggregation and reduces variance resulting in a better model.

|  |  |  |
| --- | --- | --- |
|  | **Random Forest** | **Ridge** |
| mean absolute error | 51.9 | 56.1 |
| mean squared error | 3536.4 | 4350.9 |
| root mean squared error | 59.5 | 66.0 |

* 1. Question 19. Provide the feature importance values for the RF model. Which feature is the most important? Do you get the same importance with the Ridge Regression model? Explain why.
     1. **Answer**: The feature most important in the Random Forest model is Population Driver License (0.58). Although, the Ridge regression does not have "feature\_importances\_" attribute to identify the most important feature for the model, we can use the highest absolute value of coefficient in ridge regression to do the same. The feature identified in the Ridge regression model via coefficient calculation is also Population Driver License (coeff = 73.1). The same feature is identified by both models and being the most important.

1. **Exercise # 4: Ensemble Models**
   1. Question 21-24 and 27-29: **Answered within the code**
   2. Question 25. Determine the feature importance. Which one is the most important?
      1. **Answer:** Petal width (0.44) was identified as the feature importance using the random forest model.
   3. Question 26. Use scikit learn to determine the accuracy level. What is your assessment?
      1. **Answer:** In a Random Forest Model, each tree in the ensemble is built from a sample drawn with replacement from the training set. Also, a random subset of features is selected, further randomizing the tree. And due to the averaging of less correlated trees, the variance decreases, resulting in an overall better model. Given this understanding and the accuracy calculated (96%), we can determine that the model is performing well in predicting the target.
   4. Question 30. Did the GB or the XTrees model perform better than the Random Forest model? Do you have any reservations about the GB and/or the XTrees model? Why? What is the main difference between the RF and the GB models?
      1. **Answer:** The GB (98%) performed better than the Random Forest model (96%). Whereas XTrees (96%) performed just as well as the Random Forest (96%).
      2. Boosting algorithms like GB typically aid in reducing bias however they are more sensitive to overfitting if the data is noisy and also take longer to train the model because the trees are built sequentially whereas bagging algorithms like XTrees aid in reducing variance but don’t necessarily reduce bias.
      3. Gradient Boosting uses boosting and is aimed at reducing the bias whereas Random Forest is aimed at reducing the variance by bootstrap aggregation.