NAME: JENNIFER HARRISON

ID: 223CS01000958

QUESTION 1

1. The methodology
2. The Random Forest model is a reasonable choice for binary classification problems, with n\_estimators determining complexity. The SVM model, with a linear kernel, is suitable for linearly separable data. A balanced C value of 1 balances performance and speed, ensuring a balanced approach to training data and avoiding overfitting. Higher C values can improve performance.
3. Analysis of the results

a)

|  |  |  |
| --- | --- | --- |
|  | Chess dataset | Letter Recognition dataset |
|  | SVM | Random Forest Classifier |
| Parameter | Kernel -linear | n\_estimatores = 100 |
| Training set size | 0.5 | 1.0 |
| Test set size | 0.5 | 1.0 |
| Accuracy test | 0.2824351 | 1.0 |
|  |  |  |
|  | SVM |  |
| Parameter | C = 1 | C=1 |
| Training set size | 0.7 | 1.0 |
| Test set size | 0.3 | 1.0 |
| Accuracy test | 0.5288 | 0.89875 |
|  |  |  |
|  | Random Forest Classifier |  |
| Parameter | n\_estimatores = 100 |  |
| Training set size | 0.5 |  |
| Test set size | 0.5 |  |
| Accuracy test | 0.824206 |  |

1. **Accuracy Test**

b)

Even when both methods are trained on the same size training set, the graph demonstrates that the random forest classifier outperforms the SVM classifier on the chess game dataset. This indicates that the random forest classifier is more adept at identifying patterns in the data and producing precise predictions when dealing with fresh data. There are several potential causes for this:

* Compared to SVMs, random forests are more resistant to overfitting.
* More intricate correlations between the features in the data can be discovered via random forests.
* For larger datasets, random forests are computationally more efficient than SVMs.
* All things considered, the random forest classifier is a solid option for machine learning applications requiring high accuracy and efficiency.

c) Increasing the size of the training set and reducing the number of estimators generally leads to better performance, but it also takes longer to train the model. I observed this when training a random forest model with n\_estimators on the chess dataset. It took longer to train than a linear SVM model, but it achieved higher accuracy on the held-out test set.

**Question 3**

Conclusion and future works

1. Yes, I think it's odd that the SVM's low accuracy on the chess dataset. It's likely that the intricate correlations between the features in the data are beyond the SVM's capacity to learn. Furthermore, the SVM's poor performance on the test data could be the result of overfitting to the training set.
2. Why, regardless of the number of estimators I adjust, the random forest classifier on the chess dataset attains 100% accuracy baffles me.

**REFERENCE**

*Show all columns of Pandas DataFrame in Jupyter Notebook*. (2023, March 20). GeeksforGeeks. https://www.geeksforgeeks.org/show-all-columns-of-pandas-dataframe-in-jupyter-notebook/

*‎Bard - Chat Based AI Tool from Google, Powered by PaLM 2*. (n.d.). Bard. https://bard.google.com/chat/303b4d3baf5b9ad0

Dreżewski, R., & Wątor, G. (2021). Chess as Sequential Data in a Chess Match Outcome Prediction Using Deep Learning with Various Chessboard Representations. *Procedia Computer Science*, *192*, 1760–1769. https://doi.org/10.1016/j.procs.2021.08.180