Student number: 561228

Machine learning courswork

1) Problem Statement

In this study, the objective is to create a system that utilizes a high-resolution camera to automatically identify seven different types of dry bean seeds. For this purpose, the Super Vector Machine algorithm has been chosen for implementation and will be compared with another algorithm called Decision Tree Classification.

The database contains seven types of dry beans named: BARBUNYA, BOMBAY, CALI, DERMASON, HOROZ, SEKER and SIRA. Additionally, the database is classified based on the following features: Area, Perimeter, MajorAxisLength, MinorAxisLength, AspectRatio, Eccentricity, ConvexArea, EquivDiameter, Extent, Solidity, Roundness, Compactness, ShapeFactor1, ShapeFactor2, ShapeFactor3, ShapeFactor4, and Class (Anon., 2020). Overall, there are 13,611 grain pictures classified into 7 categories. These pictures have 16 features, 12 dimensions, and 4 shapes.

I chose The Super Vector Machine (SVM) algorithm and also it can be used for this study because of its capability to perform accurate classification on complex datasets. SVM is effective in high-dimensional spaces and can efficiently **handle nonlinear features** (Wahyu , et al., 2018). Additionally, SVM is known for its ability to generalize well on test datasets, making it a reliable choice for **automatic detection of different types of dry bean seeds based on high-resolution camera-captured images** (Wahyu , et al., 2018).

For the extra task, I chose for Implementing **the Decision Tree algorithm** because its offers several positive aspects as:
Interpretability: Decision Trees generate easily interpretable models in the form of tree-like structures, providing insights into the decision-making process and identifying relevant features for classifying dry bean seeds (Friedman, et al., 2009).
Computational Efficiency: Decision Trees are **relatively fast** and **efficient algorithms** in terms of **training and classification time** (Baykara, 2015). This is particularly advantageous when **dealing with large volumes of data captured by the high-resolution camera** (Baykara, 2015).
Handling Nonlinear Features: While Decision Trees are based on linear splits, they can capture **nonlinear relationships** through combinations of features (Baykara, 2015). This allows the algorithm to adapt to more complex patterns present in the dry bean seed data.

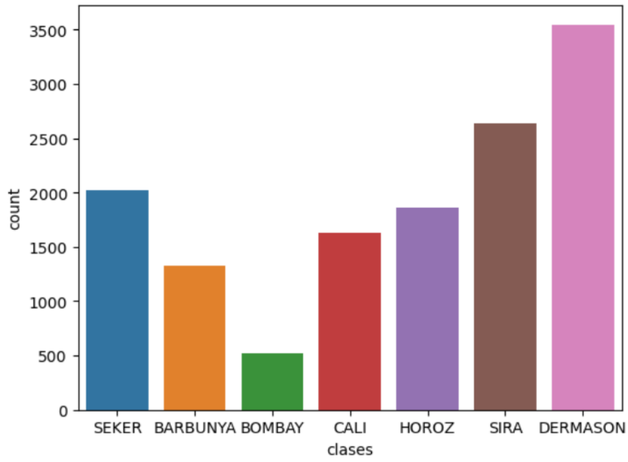
Comparison between SVM and Decision Tree: Implementing the Decision Tree algorithm enables a direct comparison with the Super Vector Machine (SVM) algorithm. This comparison helps evaluate and compare the accuracy and performance of both algorithms in the automatic detection of different types of dry bean seeds.

Firstly, Decision Tree Classification is a widely used and interpretable algorithm that can provide faster insights into the decision-making process (Mayasari, 2016), whereas SVM can have more computational cost. It creates a tree-like model of decisions and their possible consequences, making it easier to understand and visualize (Mayasari, 2016).

Secondly, comparing SVM with Decision Tree Classification allows for a comprehensive evaluation of the performance of different algorithms. It helps in determining which algorithm is better suited for the specific task of automatic detection of dry bean seeds based on high-resolution camera data.

Additionally, comparing the results obtained from both algorithms can provide valuable insights into their strengths and weaknesses. It helps in understanding the impact of different algorithmic approaches on the accuracy, efficiency, and interpretability of the seed detection system.

2) Data visualization:



According to the graph, there is an imbalance in the quantity of samples assigned to each type of dry beans, with the majority being classified as Dermason, while Bombay has the fewest number of samples.

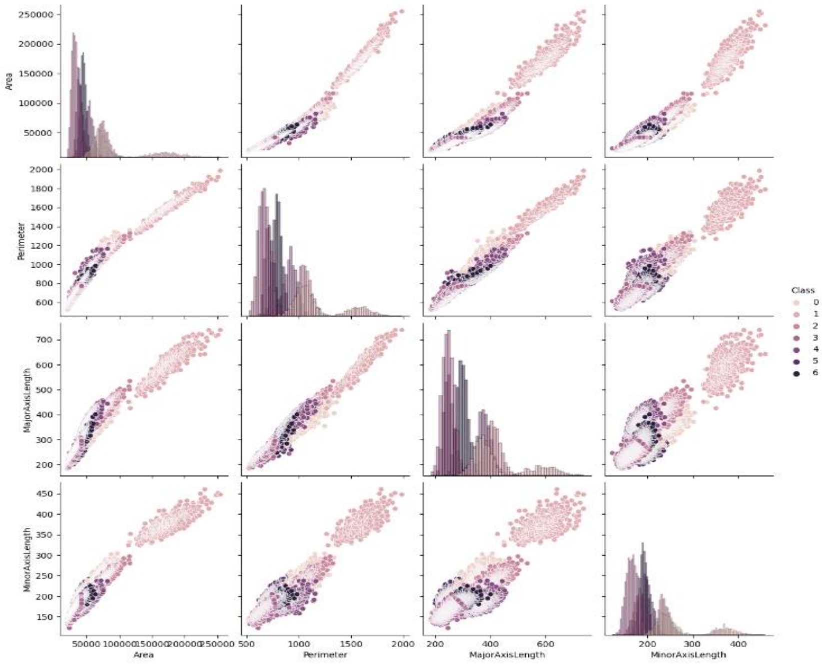
The dataset belongs to the multivariate category and is predominantly utilized for classification purposes. The dataset comprises categorical, integer, and real-valued attributes. There are **no missing values** present in the data eliminating the need for imputation or managing missing data.

To address these issues, the data has been **normalized** during pre-processing before applying the algorithms. This normalization step helps to balance the scales of each class and mitigate the dominance of certain features. By normalizing the data, we ensure that all features are on a similar scale, enabling the algorithms to converge more effectively and allowing all features to contribute meaningfully to the model's performance (Huang, et al., 2020)

For this analysis, **PCA is not required** due to the low dimensionality, and there is no evident need for it. Based on my observation, the understanding of the data features is at an appropriate level, and if PCA is applied, the original features can become more complex and less interpretable, leading to a loss of information, variance, or details from the original data (Jolliffe & Cadima, 2016).

Furthermore, decision trees are among the most robust algorithms and are inherently capable of handling high dimensionality (Baykara, 2015). Therefore, there is no need for a prior dimensionality reduction step. It's important to assess the specific requirements of your analysis, as well as the characteristics and limitations of the algorithms you plan to use, to determine whether PCA is necessary.

Upon further analysis of the summary statistics of the dataset, it became evident that the classes displayed diverse patterns and characteristics. The features demonstrated distinct means, standard deviations, and ranges, indicating the necessity for methodologies capable of capturing intricate relationships and managing varying data scales. Considering these findings, the decision tree model was initially selected for classification due to its proficiency in handling different feature types, conducting feature importance analysis, and exhibiting strong overall performance.



In the above diagrams, the density can be observed depending on the category, while for the other bar chart, this is because each seed is measured in different ways, hence **tuning** has been applied. This process was used to find the **optimal combination of hyperparameters** in the chosen machine learning models, as the seeds were parameterized using different units, such as area, length, diameter, solidity, among others (Bartz , et al., 2022). The “class” parameter was excluded and the seeds were appropriately parameterized for the tuning process.

Tree:

	precision	recall	f1-score	support
Accuracy	-	-	0.91	2709
Macro avg	0.92	0.92	0.92	2709
Weighted avg	0.91	0.91	0.91	2709

SVM

	Precision	Recall	F1-score	support
Accuracy	-	-	0.92	2709
Macro Avg	0.94	0.94	0.94	2709
Weighted avg	0.93	0.92	0.92	2709

In comparison with the different results from the confusion matrix, we can observe that SVM (Support Vector Machine) has higher scores in precision, accuracy, and F1-score (harmonic mean of precision and recall). This indicates that applying SVM to the trained data yields better results and a lower probability of having true data classified as false positives. SVM is more effective in identifying instances accurately and achieving a balance between precision and recall.

In comparison with the Decision Tree algorithm, SVM also demonstrates good scores. However, when selecting the model with the best evaluation, SVM would be the preferred choice.

3) Conclusion

As an Overview, SVM tackles the classification and regression challenges in machine learning by optimizing the decision boundary or fitting the data through techniques like maximizing the margin or minimizing the error (Wahyu , et al., 2018). It is a robust algorithm renowned for its capability to handle high-dimensional data effectively and deliver strong performance in both linear and non-linear situations (Wahyu , et al., 2018). Whilst decision trees handle the challenges of classification, regression, and feature selection in machine learning by creating a hierarchical structure of decision rules using the given features (Baykara, 2015). Thus, they provide interpretability, scalability, and the capability to handle categorical and numerical data.

There are many algorithms as an alternative, hence the selection of an algorithm is influenced by the particular needs of the problem, the properties of the data, and the balance between factors such as accuracy, interpretability, training time, and resource demands.

It is important to acknowledge **certain limitations**; The dataset lacks comprehensive contextual information about the domain or market circumstances in which the beans were collected. This absence of contextual information could restrict the suitability of the classification model to particular regions or time periods.

There is also important as a limitation that not all machine learning algorithms equally benefit from normalization. Certain algorithms, such as decision tree-based methods or rule-based methods, are less sensitive to differences in feature scales and may not require normalization as we said before (Huang, et al., 2020).

A drawback of Support Vector Machine is their computational complexity, particularly when working with large datasets (Mayasari, 2016). SVMs require solving a quadratic optimization problem, the computational cost of which increases with the number of training examples. As a result, SVMs may be less efficient when dealing with extensive datasets. Moreover, SVMs can be sensitive to the selection of hyperparameters, including the kernel function and regularization parameter, making it challenging to find the optimal values for these hyperparameters (Bartz , et al., 2022).

As a final conclusion, we can say that both models are good for classifying and predicting the type of seed. However, if we have to choose one based on its processing capacity, accuracy, and precision, SVM is the preferred choice. Despite having higher computational costs, it performs better in this case. For future studies, it would be beneficial to delve deeper into the problem by considering its context and being more specific about the requirements.

References

Anon., 2020. Dry Bean Dataset. [Online]
[Accessed May 2023].

Bartz , E., Bartz-Beielstein, T., Zaefferer, M. & Mersmann, O., 2022. Hyperparameter Tuning for machine and deep learning with R. Springer.

Baykara, B., 2015. Impact of Evaluation Methods on Decision Tree Accuracy. s.l.:s.n.

Friedman, J., Hastie, T. & Tibshi, R., 2009. The elements of statistical learning. Stanford.

Huang, L. et al., 2020. Normalization Techniques in Training DNNs: Methodology, Analysis and Application. Institute of Artificial Intelligence, Abu Dhabi, UAE, pp. 1-20.

Jolliffe, I. & Cadima, J., 2016. Principal component analysis: a review and recent developments. The royal society.

Mayasari, N., 2016. Comparison of Support Vector Machine and Decision Tree in Predicting On-Time Graduation. International Journal of Recent Trends in Engineering and Reaserch, 2(12), pp. 140-151.

Wahyu , R., Purnamia, S. W. & Rahayu, S. P., 2018. Boosting Support Vector Machines for Imbalanced Microarray Data. ELSEVIER, Volume 144, pp. 174-183.