**Project Title: Comprehensive Analysis and Classification of Fish Species Using Machine Learning and Feature Engineering**

**Comprehensive Project Summary**

**Overview**

This project focuses on classifying fish species from a dataset containing 1492 observations with six physical features—weight, three length-related measures (L1, L2, L3), height, and width—and a 7-class species label. The work involved training multiple machine learning classifiers, exploring feature importance and selection, simulating effects of noise addition, and analyzing classification confidence and possible label noise.

**1a) Classifier Training and Performance Evaluation**

**Data Characteristics and Preprocessing**

* The dataset displayed *significant class imbalance*, with species distribution ranging from ~29% (Bream) down to ~2% (Whitewhisk).
* Strong feature correlations (>90%) existed among Weight, L1, L2, and L3.
* Due to the small number of features, PCA was applied but explained variance suggested possible loss of separability in linear PCs.
* Non-linear dimension reduction methods (t-SNE and Kernel PCA) better captured underlying class structures, showing improved separability in some difficult cases (e.g., Smelt) but still overlapping classes (e.g., Perch, Roach, Whitewhisk).

**Classifier Training and Hyperparameter Optimization**

* Seven classifiers of varied nature were employed:
  + K-Nearest Neighbors (KNN) — instance-based, non-parametric.
  + Quadratic Discriminant Analysis (QDA) — generative parametric model.
  + Support Vector Machines (SVM) — linear and non-linear kernels.
  + Kernel Logistic Regression — linear and non-linear kernels.
  + Random Forest (RF) — ensemble of decision trees focusing on feature splits.
* Pipelines integrated scaling, PCA, and classification steps.
* Hyperparameters were optimized using GridSearchCV on stratified train-validation splits across 10 iterations.
* Predominantly, 5 principal components and non-linear kernels (RBF/poly) were chosen, with low variability across iterations.
* Random Forest hyperparameters showed more variability, consistent with its complexity.

**Performance Insights**

* Kappa scores, accounting for class imbalance, identified KNN (~0.93), RF (~0.88), and SVM with RBF kernel (~0.87) as top performers.
* Boxplots confirmed that KNN had marginally better stability and slightly higher median accuracy.
* Confusion matrices on unseen test data supported these findings.
* Common misclassifications:
  + Bream over-prediction due to being the dominant class.
  + Notable confusion between Silver Bream and Perch, reflecting their physical similarity.

**1b) Feature Importance and Feature Selection**

**Permutation Importance**

* Across models, **Height** was consistently the most important feature, followed by **Width** and the full body length measure **L3**.
* Weight and intermediate length measures (L1, L2) were generally less critical, except for QDA which showed an inconsistent priority likely linked to poorer performance.
* Low variability in importance scores over multiple iterations demonstrates stable feature relevance.

**Feature Selection Methods**

* Tested Variance Thresholding, F-score filtering, and exhaustive Best Subset selection on top models (KNN, RF, SVM).
* Variance filtering had little impact—best models kept all features despite differences in feature variance.
* F-score also did not improve models significantly; all features were retained.
* Best Subset method showed that:
  + RF preferred all features.
  + SVM selected slightly fewer features, excluding L2.
  + KNN selections were mostly comprehensive.
* Results imply all physical features contribute meaningfully.

**1c) Effects of Adding Simulated Noise Features**

**Uncorrelated Noise Features**

* Added between 0 to 50 uncorrelated Gaussian noise features with increasing mean and variance.
* Random Forest was notably *robust* to this noise, maintaining stable validation accuracy.
* SVM experienced moderate decreasing accuracy.
* KNN suffered dramatic degradation due to its reliance on distance metrics, which become less meaningful in noisy high-dimensional space.

**Correlated Noise Features**

* Added features correlated progressively stronger to randomly selected original features.
* Performance trends mirrored uncorrelated noise results but validation accuracy decreased slightly faster.
* Reinforces that KNN is the most sensitive to added noise.
* RF's ensemble feature splits help it resist noise.
* SVM’s RBF kernel preserves some robustness by capturing non-linear relations.

**1d) Classification Confidence and Label Noise Analysis**

* Utilized model confidence metrics to assess which observations could be classified with certainty.
* Ambiguous or set-valued predictions suggested possible mislabeled or borderline samples.
* Overlaps between classes such as Perch and Silver Bream increased potential mislabeling likelihood.
* Identifying these uncertain observations offers a pathway to dataset quality improvement.

**Final Remarks**

* Addressing class imbalance via stratified splits and applying Kappa as metric yielded reliable evaluation.
* Non-linear dimension reduction methods (t-SNE, Kernel PCA) reveal richer underlying data structure than PCA alone.
* Models KNN, RF, and SVM with RBF kernels emerged as best choices for this classification problem.
* Height, Width, and L3 are the most influential features for species distinction.
* Random Forest showed resilience to added noise, while KNN was highly sensitive to noise features.
* Strong evidence of mislabeled or ambiguous samples invites further dataset refinement.

**Key Interview Questions & Model Answers**

**Q1: How did you handle class imbalance in your classification pipeline?**  
*A1: We addressed class imbalance by applying stratified sampling during train-test splits to preserve class proportions. Additionally, Cohen’s Kappa score was used as the main evaluation metric to provide a balanced perspective beyond accuracy, reflecting performance across all classes fairly.*

**Q2: Why did you choose Kappa score over accuracy for performance evaluation?**  
*A2: Kappa accounts for chance-level agreement and balances the influence of dominant classes, unlike accuracy which can be misleading in imbalanced datasets. This provides a more reliable measure of true discriminatory performance across all classes.*

**Q3: Why did PCA fail to clearly separate the classes, and how did you address this?**  
*A3: PCA captures linear variance directions, but the underlying features showed non-linear class separability. Using non-linear methods such as t-SNE and Kernel PCA better revealed class clusters, supporting the preference for non-linear classifiers and kernels.*

**Q4: Which features were most important across models and why?**  
*A4: Height consistently emerged as the top feature, followed by Width and L3, likely because these physical sizes relate strongly to species differences. Weight and intermediate length measures were correlated with other features, thus contributing less additional unique information.*

**Q5: How did adding noise features affect classification performance?**  
*A5: Adding uncorrelated noise severely degraded KNN performance due to its reliance on Euclidean distances, while Random Forest maintained stable accuracy given its ensemble feature selection mechanism. SVM had moderate robustness aided by the RBF kernel's ability to capture complex patterns.*

**Q6: What evidence did you find of potential mislabeled data, and how might this impact your models?**  
*A6: The confusion matrices and low-confidence predictions concentrated misclassifications in closely related species like Silver Bream and Perch. These ambiguous or mislabeled samples reduce model certainty and accuracy, highlighting the value of incorporating confidence measures and possibly revisiting dataset labeling.*

**Q7: How can this analysis be extended to improve fish species classification?**  
*A7: Future work could integrate more sophisticated handling of ambiguous cases using set predictions or uncertainty modeling, address class imbalance with advanced techniques like SMOTE, and experiment with ensemble methods combining best-performing classifiers to further boost robustness.*

If you'd like, I can help prepare presentation slides or a detailed formatted report with this content. Just let me know!

Add to follow-up

Check sources