Age prediction of the Abalone dataset with Random Forrest Regression

Id: 39603547
Lancaster University
Leipzig, Germany

***Abstract*—The prediction of the age of abalones is an important challenge in the data science domain, which aims to find more straight-forward alternatives to the biological expensive and time-consuming microscopy techniques. The prediction is instead based on some simple physical measurements of abalones. In this study, multiple regression algorithms are exposed to this dataset and compared for their performance. A clear objective of this report is to describe the influence of the features data distribution (especially within outliers) in the model's performance, as well as try to discuss which machine learning algorithms are more convenient based on their current predictive power, error minimization and principal basis. The main objective of this paper is to provide and justify reliable predictions on the age of the abalones.**

I. Introduction

Abalones are marine species belonging to the invertebrates group of mollusks. These marine animals present a flattened calcified shell similar to the carapace of snails. They are normally found in reef habitats close to the coasts. Abalone's body shape confers on him consistency towards waves and sea tides [1].

Abalones are provided with numerous physical characteristics of interest which are mainly studied from a biological point of view. However, predicting some of its features using biological techniques represents a serious challenge. In this situation, relying on data science is usually an interesting and necessary approach. In this report, the analysis focuses on predicting the age of abalones through the use of some of its descriptive physiscal characteristics and comparing the results attained from different models application to the dataset. The interest in studying a population such as the abalones lies on providing useful information to improve their biodiversity and conservation [2] or understanding the environmental impact that represent fish farms [3].

II. The Dataset

The dataset provides observations of the physical measurements of 4000 abalones. The features recovered from the abalones refer to specific physical characteristics (TABLE I). The age of the abalones can be obtained by adding 1.5 to the number of rings present in their physical structure. It appears that predicting age using the number of rings does not represent any challenge, as the relationship between them is completely deterministic.

The dataset suggests that the analysis must focus in predicting the age through the rest of the recorded features of abalones. Furthermore, the number of rings of an abalone can only be obtained by counting them through microscopical techniques. For this purpose, the shell of abalones is cut and staining methods are performed to differentiate the rings to other internal structures. This represents a very time-consuming and boring process which can be relieved by developing an algorithmic model predicting age towards the physical measurements that are widely easy to obtain [4].

Feature	Type	Description	Units
Sex	Categorical	M, F, and I (Infant)	-
Length	Continuous	Longest shell measurement	mm
Diameter	Continuous	Perpendicular to length	mm
Height	Continuous	With meat in shell	mm
Whole weight	Continuous	Whole abalone weight	grams
Shucked weight	Continuous	Weight of meat	grams
Viscera weight	Continuous	Gut weight (after bleeding)	grams
Shell weight	Continuous	After being dried	grams
Number of rings	Integer	+1.5 gives the age in years	-

TABLE I
Abalones dataset features description.

III. Data Pre-processing

A. Data distribution inspection

The dataset provides 9 features, as depicted in TABLE 1, among which the Sex of the abalones is described as a nominal categorical variable [5] and the rest as numerical continuous features. The observations containing missing values from the original dataset are removed and the ranges of the continuous variables scaled. The target variable can be appended to the dataset by simply adding 1.5 to the number of rings of the abalones.

The data distribution of the continuous numerical features can be inspected through box-plots, which are the most prevalent visualization techniques employed in anomalies detection[6]. The total amount of outliers indicates the possibility of finding further problems in the model's performance as the dataset frequents a high variability in all the features (Fig.1).

B. Anomalies treatment

The outliers effect on the model's performance lies on the total proportion of outliers in the dataset and the robustness towards anomalies of the chosen model. The hypothesis formulated is that the Abalones dataset presents a considerably high number of outliers. The elimination of all these data points is suspicious of having negative consequences on the

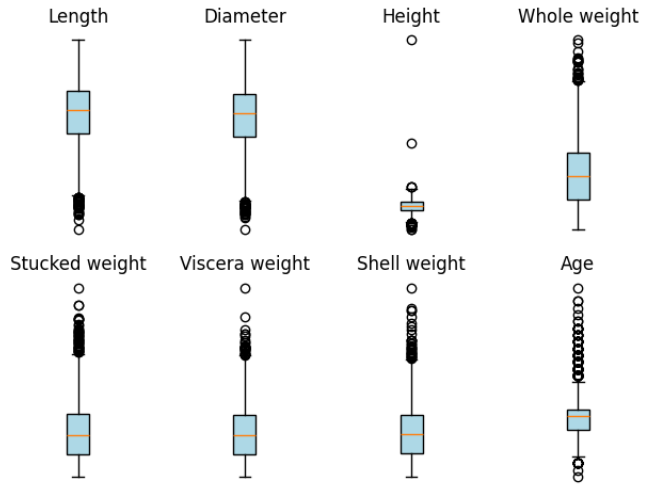

Fig. 1. Box-plots for the numerical features of the Abalones.

models fitted to data. This would explain why the possibility of studying the model's performance through the outliers removal is not discarded. For that, the feature's data distribution is then addressed by the histograms and Quantile-Quantile-plots (qq-plots) visualization techniques (Fig.2).

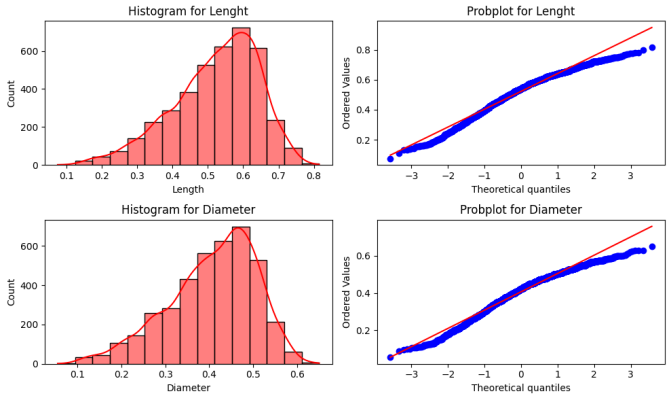

Fig. 2. Data distribution for some features of the Abalones

The bell-shape forms in the continuous variables shown by the histograms and the center fit of the qq-plots reinforce the speculations about high outliers prevalence in the dataset. Either the histograms and the probabilistic plots tails deviate from the depicted behavior from the other observations. It can be assumed that the features follow a Normal distribution far from the outliers. The detection and deletion of the outliers is carried out by two main statistical techniques, the z-score normalization (1) and the inter-quartile range (IQR) (2) [7].

$$z = \frac{x - \mu}{\sigma} \quad \text{(Z-score normalization)} \tag{1}$$

$$\text{IQR criteria:} \quad \begin{aligned} x &< Q_1 - 1.5 \cdot IQR \\ x &> Q_3 + 1.5 \cdot IQR \end{aligned} \tag{2}$$

The z-normalization (1) transforms the data to have a mean of 0 and a standard deviation of 1. In a Normal distribution, 99.7% of the data is found within 3 standard deviations from the mean [8]. All the observations out of this range are considered as anomalies. Nevertheless, the Abalones dataset features are not assumed to follow a completely straight Normal distribution and so the thresholds imposed to detect the outliers within a variable strongly depend on the skewness this variable presents. For feature values showing negative skewness (larger left tail) in their histogram, the lower limit adds either half or one standard deviation to the threshold. The same procedure is applied to variables with negative skewness (larger right tail), where the upper limit is increased by a half or one entire standard deviation.

As the criteria related to the Z-score normalization is highly subjective in thresholds estimation, IQR (2) seems a more reasonable method to confront the outliers detection challenge. This technique states that data points showing values inferior to the difference between the first quartile minus one and a half times the IQR or superior to the addition between the third quartile plus one and a half times the IQR are detected as outliers [9]. Once these anomalies are removed from the dataset, the missing values are replaced by mean imputation, which consists of filling up the missing values by the mean of the variables distribution without accounting for the outliers. In this case, the imputation technique is decided to be the mean because of the normality assumption on data [10].

C. Feature scalling

Two resulting datasets are obtained, one dataset conserving the original data distribution and another with the outliers replaced by the imputed mean. The following procedure applied to data is feature scalling. This procedure normalizes all the features of the dataset (without accounting for the age). The concrete approach selected for that task is the Min-Max Scalling (3) which transforms the data to a scale of 0,1. Applying this technique to the dataset ensures the equal contribution of features to the selected model learning process[11].

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3}$$

D. Feature engineering

The Sex of the abalones is the only categorical variable and it has to be treated before fitting the models to data. Most of Machine Learning models do not accept categorical variables as model inputs. A common alternative to this situation is conducting a one-hot encoding labeling on these categorical variables. This approach introduces new binary variables to the dataset, which takes the value of 1 if the category given appears or the value of 0 in the opposite case. Instead of that, the abalones Sex is replaced by an integer depending on the category it represents (Male:0, Female:1 and Infant:2). In this way, no additional variables are introduced in the dataset.

E. Feature correlation analysis

The features correlation understanding is key in the model selection process. The Pearson's correlation measures the

linear relationships between the features [12]. All these calculations can be represented in a heat-map (Fig.3). The r values (from the Pearson's correlation coefficient) can take values from a perfect negative correlation (r = -1) to a perfect positive correlation (r = 1), going through the possibility of not detecting linear relationship at all (r = 0).

No linear relationships are depicted between the age and the other physical measurements in the Abalones dataset (apart from the redundant relationship with the number of rings). However, some linear correlation can be observed among the predictors themselves. This leaves an open door towards the possibility of combining several variables for the age prediction purpose. From that statement, it can be drawn that linear models could be discarded from the possible available models. The non linear relationships present between the predictors and the target variable do not leave any chance about the use of simple statistical approaches such as linear regression models [13]. The analysis is redirected to Machine Learning related models for the abalones age prediction.

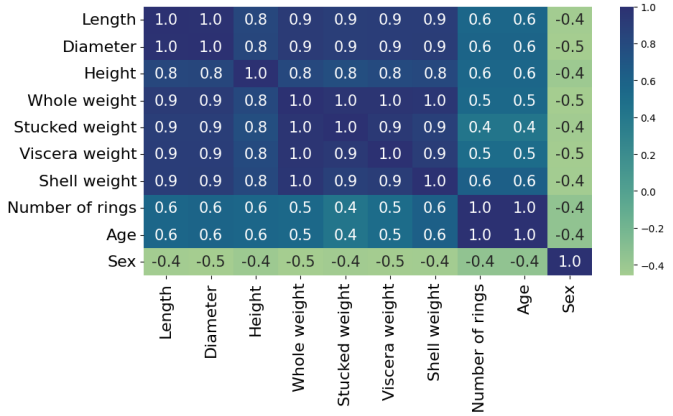

Fig. 3. Heatmap of the scaled dataset.

IV. Evaluation metrics

The predictive accuracy of machine learning models is commonly evaluated by performance metrics. These metrics rely on differences encountered while comparing the expected outcomes and the predicted results obtained from a model or assessing how good does a model fit the data (goodness of fit). The evaluation metrics used in this report are the Mean Squared Error (MSE), the Mean Absolute Error (MAE) and the R squared (R^2) [14].

The MSE computes the average of the square differences between the predictions and the real observations . It is usually applied in regression models and penalizes large prediction errors. The MAE metric is also commonly employed in regression models and calculates the average of the absolute value of the differences between the predicted results and the real ones. The R^2 metric contributes by explaining how good regression models explain the variability of the target variable with respect to the other features. In other words, what is the global amount of variance that can be explained by the predictors.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

V. Age prediction by the Random Forest Regression algorithm

A. Model justification

Ensemble machine learning techniques combine predictions of multiple individual models to elevate a machine learning algorithm overall performance [15]. Random Forest regression machine learning algorithm (RFR) is considered to be an ensemble model. It is based on the integration and application of multiple decision trees. Even though this algorithm is mostly used in classification tasks, it shows some advantages while being applied towards regression problems. The ensemble models rely in the idea of joining different regression models, from which one might present a weaker and performance, and computing the overall performance of the ensemble.

RFR has the ability to handle non linear patterns within the features. They demonstrate more robustness and resistance to overfitting than the normal multiple decision trees and are also less sensitive to outliers [16]. One of their main disadvantages is their lack of explainability compared to regular decision trees. Their predictive accuracy is higher and the feature importance [17] of the model can be displayed. This is a measure of the weight that has every independent variable (predictor) on the target variable prediction (TABLE II).

B. Model construction

1) Data Partition: A proportion of 80 % of the dataset is used for training and the remaining subset for testing the model. The number of rings of the abalones is removed from the predictors set. Before fitting the model to the data, the two resulting subsets for training and testing are scaled by a z-normalization [18].

2) Model 1: The first model adopts a RFR algorithm from the scikit-learn library. To try to enhance the model's performance, a Principal Component Analysis (PCA) analysis is performed on the dataset variables. PCA is a dimensionality reduction technique that aims to capture the maximum variance among data using the fewest possible number of components. The original features are then transformed to fewer variables which are uncorrelated [19]. Thus, there is no noise introduced by the redundancy of correlated features. In the Abalones dataset, the number of principal components to explain a 99% of the variance within the data is 5 (Fig. 4). In PCA, each principal component explains a certain proportion of the variability of the data. Even though the model's predictive power remains the same, PCA still improves the model

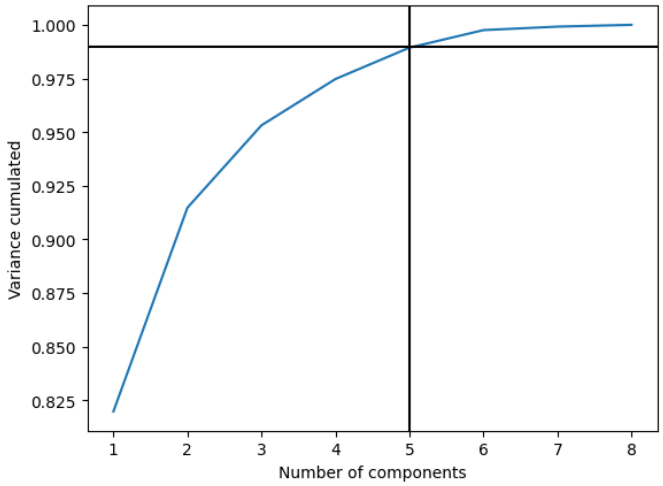

Fig. 4. PCA analysis on the scaled dataset

computational expense by reducing the number of variables for an equal outcome.

PCA does not achieve any improvement. This could be related to the fact that models such as RFRs already manage properly by themselves high-dimensional dataset without the needing of dimensionality reduction approaches. Furthermore, PCA does not guarantee that the dimensions that contain the highest variability within data are also the ones that generate the most impact in model's predictive accuracy.

Feature selection appears scene as a possible solution to boost the model predictive accuracy. The non linear correlations between several features might cause inaccurate predictions. RFR offers the possibility of visualizing importance of features in the model predictions. In other words, trying to create combinations of the variables having the highest importance or deleting the variables which are almost irrelevant may increase the predictive power of the model. Unfortunately, any combination of features involving the deletion of other variables in the model fit implies a decrease in the accuracy.

Feature	Importances
Length	0.028972
Diameter	0.034103
Height	0.032491
Whole weight	0.059907
Stucked weight	0.162721
Viscera weight	0.044522
Shell weight	0.612246
Sex	0.025039

TABLE II
FEATURE IMPORTANCES.

t-Distributed Stochastic Neighbor Embedding (tSNE) is a dimensionality reduction technique applied to the Abalones dataset. tSNE embeds high dimensional data to a lower and visual dimensional space, such as 2 or 3 dimensions. This procedure builds clusters within the data points and is normally used in datasets presenting non linear relationships among features [20]. Different combinations of tSNE most influencing hyperparameters are tested. The model's performance does not respond positively to any of them and the performance remains poor (TABLE III).

Hyperparameter	Description	Best
`n_components`	Dimension of the embedded space	2
`perplexity`	Number of nearest neighbors	5
`learning_rate`	Controls size of steps minimizing loss f	1000
`n_iter`	Number of iterations	500

TABLE III
OPTIMAL HYPERPARAMETERS COMBINATIONS OF T-SNE.

The last technique to improve the model is the cross validation for hyperparameter optimization. It consists of trying different hyperparameter available options (TABLE IV) and preserve the one that maximizes the performance and minimizes the error [21]. This approach barely improves the model and is highly time-consuming and computationally expensive. The hyperparameter tuning is one of the most common approaches applied to a model when it comes to search the best predictions.

Hyperparameter	Description	Best
`max_depth`	Maximum tree depth	10
`min_samples_leaf`	Min nº of samples for a leaf node	4
`min_samples_split`	Min nº of samples for internal node	10
`n_estimators`	Number of trees	300

TABLE IV
OPTIMAL HYPERPARAMETERS COMBINATIONS FOR RANDOM FORREST REGRESSOR.

3) Model 2: The second model is a RFR algorithm built through the tensor flow library based on Neural Networks. The first model evaluation is performed on the scaled dataset with the defaults parameters. The obtained results don 't differ from the ones in the sklearn RFR. The keras tuner is a meta-learning algorithm that finds the optimal hyperparameters (TABLE V) values of a base learner. This algorithm is applied to the model providing a wide range of possibilities for all the parameters of the keras RFR. The tuner looks for a given number of trials and keeps going through all the parameters until reaching the ones that maximize the accuracy of the fitted model. Configuring the tuner with all the different choices entails more possibilities to find the optimal solution. The influence of the tuner on the predictive accuracy is undisputed, as the model performance increases considerably even though it is still not reliable [22].

VI. SUPPORT VECTOR REGRESSION ALGORITHM

A. Model justification

Support Vector Machine (SVM) is a supervised algorithm normally involved in classification tasks, but also applied to regression problems (SVR). These models are effective at solving problems on datasets with small number of samples (such as the Abalones dataset), presenting non linearity among features or that prone to overfitting. The main assumption SVR holds is that all the samples come from the same distribution.

SVM manages to adapt to regression by implementing an epsilon-insensitive loss function. To achieve the prediction in non linearity conditions, SVR offers the possibility to use

Hyperparameter	Description
num_candidate_attribute_ratio	Proportion of attributes.
use_hessian_gain	Use of Hessian G.
split_axis	Axis splitting method.
apply_link_function	Apply link function.
growing_strategy	Tree-growing strategy.
max_depth	Depth of the tree.
max_num_nodes	Maximum number of nodes.
shrinkage	Learning rate.
subsample	Samples in bootstrap.
use_regularization	Regularization.
num_trees	Number of trees.
bootstrap_training	Bootstrap sampling.
l2_regularization	L2 regularization.

TABLE V
KERAS TUNER HYPERPARAMETERS DESCRIPTION TABLE.

diverse kernel functions that transform the sample variables to another feature space. The Abalones dataset contains some irregularities and heterogeneous information which conduce to a more sophisticated approach involving Multiple-Kernel Learning (MKL). MKL is more flexible than the traditional simple kernel processing method and enhances the generalization performance and interpretability of the model [23].

B. Model construction

L1 Multi-Kernel Learning Support Vector Regression Ensemble Algorithm (L1 MKL-SVR) treats the Abalones dataset age prediction problem by applying to the features space a transformation through a combination of more than one kernel functions, aiming to have a better effect on the model prediction performance. The data partition process and preprocessing steps are equally followed by the RFR and L1 MKL-SVR.

Three different kernels are applied simultaneously to the training data. A set of possible combinations for all the hyperparameters (TABLE VI) is created to cover every possible composition of the kernels. The model runs a high enough number of iterations where the combinations are randomly chosen. A SVR model is uploaded for each of the kernel-type functions (radial-basis, lineal and polynomial functions). These three models are introduced together with L1-Ridge regession model into an ensemble algorithm. The L1-Ridge regression model is implemented in the ensemble algorithm to ensure that the model does not fail capturing also the linear patterns. The ensemble model then fits the data and carries the predictions. A voting regression is disposed inside the ensemble model to boost the model performance. The best kernel parameter arrangements are conserved and the model with the better prediction accuracy is conserved [23] .

Parameter	Description
C	Regularization parameter.
epsilon	Margin of tolerance. (epsilon-insensitive loss function)
gamma	Kernel coefficient.
degree	Degree of the polynomial kernel function.

TABLE VI
KERNEL FUNCTION PARAMETERS WITH DESCRIPTIONS.

VII. MODELS COMPARISON

A. Outliers effect discussion

The removal of outliers impacts the models decreasing the evaluation metrics with respect to the original scaled dataset. In the RFR with outliers the R^2 is equal to 0.53 and the MAE 1.63 while in the SVR without outliers the R^2 is 0.5 and the MAE is 1.31 (TABLE VII). A general trend can be extracted from these observations. On one hand, the decrease of R^2 is associated to the contribution of the outliers to the dataset variance. By removing these extreme values, some variability is lost within the data and consequently, the goodness of fit decreases. In the other hand, the error distance metrics (MAE and MSE) decrease because the removal of outliers generates simultaneously a decrease in the error of the predictions.

The other main problem to be addressed is the influence of the number of outliers in the final choice. Is this number high enough to directly condition the decrease in the predictive accuracy. The obtained R^2 suggests that the number of outliers is high enough to consider that these anomalies are an important part of the data structure. When these values are eliminated, the model struggles to capture the global patterns in data. The Abalones dataset contains too many outliers to be deleted but, at the same time, too many to be ignored.

B. Models predictions

There are not significant variations in the predictive performances between the RFR models imported from the different libraries. The main differences arise from their basic principles and implementation processes. Scikit-learn RFR is based on traditional machine learning approach while the keras implementation entails the use of neural networks (NN) combined with other standard models. Keras RFR surpasses scikit-learn predictive power by integrating deep NN within its structure. NN can learn more complex data representations and capture more closely non-linear patterns than RFR do. Moreover, the keras tuner provides a more flexible hyperparameter optimization procedure than the classical approach in machine learning consisting of cross validation.

Model	MSE	R^2	MAE
SVR with outliers	5.19	0.57	1.55
SVR without outliers	3.11	0.48	1.33
sklearn-RFR with outliers	5.34	0.53	1.63
sklearn-RFR without outliers	2.77	0.50	1.31
keras-RFR with outliers	4.88	0.60	1.53
keras-RFR without outliers	2.81	0.53	1.31

TABLE VII
EVALUATION METRICS OF DIFFERENT MODELS.

RFR and SVR show also almost identical results. Both models fail to accurately predict the age of the abalones (RFR: 0.53 and SVR: 0.57) (TABLE VII). The analysis lies on depicting which could be the reasons that led these models to their poor performance. RFR and SVR can be designed to deal with non linear data, as RFR is an ensemble of multiple decision trees and SVM can transform the feature space using kernels. RFR and SVR completely differ towards

their response to noise effect. RFR shows robustness while SVR is sensible to noise and outliers.

Before the analysis, the characteristics from both algorithms were suiting the analysis and RFR and SVR seemed very promising regression algorithms that could achieve the correct prediction of the ages of the abalones. But as the evaluation metrics show in TABLE VII, the prevalence of non linear relationships between the features and the target variable is imposed against the adjustement of the hyperparameters or the robustness of the fitted models.

VIII. Conclusion

The age prediction in the Abalones dataset represents a serious challenge within the data science paradigm. Predicting age through physical measurements of the abalones other than the number of rings raises an important obstacle, dealing with complex and non linear relationships among features. In other regression problems, techniques that normally enhance algorithm's performance such as dimensionality reduction or hyperparameter optimization are capable of improving the models. In this report, both RFR and SVR have failed to predict accurately the age. The unavoidable presence of outliers in the data diminishes the model's ability to fit data. The handling of the anomalies opens up to an interesting dichotomy for any regression problem that comes in further studies. The decision of removing or not the outliers is mediated by a trade-off between decreasing the accuracy and reducing the prediction error or the inverse situation.

Deep Learning Neural Networks might be the key in future analysis on the age prediction of abalones, as they manage to capture closely non linear patterns among data. The only model in this analysis consisting of a Neural Network architecture (keras RFR) reached the best evaluation metrics. Trying to create efficient Neural Networks architectures to improve age prediction in the Abalones dataset is a future line that should not be ruled out to solve this kind of regression problems.

References

[1] P. A. Cook, *Introduction, taxonomy, and general biology of abalone*, in *Developments in Aquaculture and Fisheries Science*, vol. 42, Elsevier, 2023, pp. 1–8. DOI: https://doi.org/10.1016/B978-0-12-814938-6.00001-4.

[2] K. Kerlin, *Ocean Acidification Creates Legacy of Stress for Red Abalone: Reducing Exposure at Crucial Stages Can Help Save Red Abalone*, December 6, 2023. Available: https://www.ucdavis.edu/climate/news/ocean-acidification-creates-legacy-stress-red-abalone.

[3] B. D. Russell, S. D. Connell, C. Mellin, B. W. Brook, O. W. Burnell, and D. A. Fordham, *Predicting the Distribution of Commercially Important Invertebrate Stocks under Future Climate*, PLoS ONE, vol. 7, no. 12, pp. e46554, Dec. 2012. Available: https://doi.org/10.1371/journal.pone.0046554.

[4] UCI Machine Learning Repository, *Abalone Dataset*. Available: https://archive.ics.uci.edu/dataset/1/abalone. Accessed: Dec. 13, 2024.

[5] D. Chen and C. Anderson, *Categorical Data Analysis*, preprint of an Elsevier book chapter, 11 pages. Available: https://arxiv.org/abs/2409.02942. Accessed: Dec. 13, 2024.

[6] V. Chandola, A. Banerjee, and V. Kumar, *Anomaly Detection: A Survey*, ACM Computing Surveys, vol. 41, no. 3, Jul. 2009, doi: 10.1145/1541880.1541882.

[7] I. Hodge and M. Austin, *Outlier Analysis*, 1st ed., Springer, 2017.

[8] SciPy Documentation, "scipy.stats.zscore," *SciPy*, [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.zscore.html. [Accessed: Dec. 13, 2024].

[9] Datagy, "Pandas IQR: Calculate the Interquartile Range in Python," *Datagy*, May 10, 2023. [Online]. Available: https://datagy.io/pandas-iqr/. [Accessed: Dec. 13, 2024].

[10] A. Desiani, N. R. Dewi, A. N. Fauza, N. Rachmatullah, M. Arhami, and M. Nawawi, "Handling Missing Data Using Combination of Deletion Technique, Mean, Mode and Artificial Neural Network Imputation for Heart Disease Dataset," *Science and Technology Indonesia*, vol. 6, no. 4, pp. 303-312, Oct. 2021. https://doi.org/10.26554/sti.2021.6.4.303-312.

[11] Scikit-learn, "sklearn.preprocessing.MinMaxScaler," *Scikit-learn User Guide*, Accessed: Dec. 2023. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[12] L. Breiman, "Statistical Modeling: The Two Cultures," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001.

[13] M. Goyal, M. Pandey, and R. Thakur, "Exploratory Analysis of Machine Learning Techniques to predict Energy Efficiency in Buildings," *IEEE*, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9197976.

[14] S. Thomas, "Understanding the Pearson Correlation Coefficient," *Outlier Articles*, Nov. 4, 2023. Accessed: Dec. 2023. [Online]. Available: https://articles.outlier.org/pearson-correlation-coefficient#section-what-is-a-correlation.

[15] N. Van Otten, "Top 6 Most Powerful Ensemble Learning Techniques Explained & Algorithms That Implement Them," *Spot Intelligence*, Aug. 9, 2023. Accessed: Dec. 2023. [Online]. Available: https://spotintelligence.com/2023/08/09/ensemble-learning/.

[16] Y. Zhang, S. Wang, T. Fu, and J. Hu, "Application and Research Based on Random Forest Regression and CRITIC Methods," in *Proceedings of IEEE*, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10361767.

[17] Scikit-learn, "RandomForestRegressor," *Scikit-learn Documentation*, 2023. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#sklearn.ensemble.RandomForestRegressor.feature_importances_.

[18] Scikit-learn, *train_test_split*, Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html, Accessed: December 2023.

[19] A. Rehman, A. Khan, M. A. Ali, M. U. Khan, S. U. Khan, and L. Ali, "Performance Analysis of PCA, Sparse PCA, Kernel PCA and Incremental PCA Algorithms for Heart Failure Prediction," in *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, IEEE, 2020, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/9179199.

[20] M. Mittal, P. G. J., G. P. M. S., R. M. Devadas, L. Ambreen, and V. Kumar, "Dimensionality Reduction Using UMAP and TSNE Technique," in *2023 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, IEEE, 2023, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/10690797.

[21] A. R. Laeli, Z. Rustam, S. Hartini, F. Maulidina, and J. E. Aurelia, "Hyperparameter optimization on support vector machine using grid search for classifying thalassemia data," in *Proceedings of the 2020 International Conference on Data Science and Its Applications (ICoDSA)*, IEEE, 2020. Available: https://ieeexplore.ieee.org/document/9317227

[22] TensorFlow, "TensorFlow Decision Forests: Tuner," TensorFlow Documentation. [Online]. Available: https://www.tensorflow.org/decision_forests/api_docs/python/tfdf/tuner. [Accessed: Dec. 13, 2024].

[23] X. Xie, K. Luo, and G. Wang, "A New L1 Multi-Kernel Learning Support Vector Regression Ensemble Algorithm With AdaBoost," *IEEE Access*, vol. 10, pp. 15063-15072, Feb. 2022, doi: 10.1109/ACCESS.2022.3151672.