Open Access Library Journal
2022, Volume 9, e8414
ISSN Online: 2333-9721
ISSN Print: 2333-9705

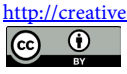
Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest

Azad Abdulhafedh

University of Missouri, State of Missouri, USA
Email: dr.azad.s.a@gmail.com

How to cite this paper: Abdulhafedh, A. (2022) Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest. *Open Access Library Journal*, **9**: e8414.
https://doi.org/10.4236/oalib.1108414

Received: February 1, 2022
Accepted: February 15, 2022
Published: February 18, 2022



Abstract

Statistical techniques are important tools in modeling research work. However, there could be misleading outcomes if sufficient care is undermined in choosing the right approach. Employing the correct analysis in any research work needs deep knowledge on the differences between these tools. Incorrect selection of the modeling technique would create serious problems during the interpretation of the findings and could affect the conclusion of the study. Each technique has its own assumptions and procedures about the data. This paper compares common statistical approaches, including regression vs classification, discriminant analysis vs logistic regression, ridge regression vs LASSO, and decision tree vs random forest. Results show that each approach has its unique statistical characteristics that should be well understood before deciding upon its utilization in the research.

Subject Areas

Applied Statistical Mathematics, Mathematical Analysis

Keywords

Supervised Learning, Logistic Regression, Discriminant Analysis, KNN, Ridge Regression, LASSO, Decision Tree, Random Forests, PCA, Clustering

1. Introduction

Selection of the correct statistical approach is vital in any research work. The

wrong selection would lead to incorrect interpretation of the results and inadequate findings. In some research situations, there could be some confusion on choosing the most appropriate technique for the analysis, because different techniques seem to be applicable. In order to overcome such problems, the researcher should be aware of the major differences between possible statistical modeling approaches that could be applied simultaneously [1]. In addition, the researcher should have clear idea of the variables that will be used in the research work, whether they are categorical or nominal, ordinal, or rank-ordered, interval, or ratio-level. Moreover, the type of data is also a fundamental concept in the analysis, for example the techniques appropriate to interval and ratio variables are not suitable for categorical or ordinal variables [2]. Besides, the researcher should have good knowledge of parametric methods and non-parametric methods. Non-parametric techniques must be used for categorical and ordinal data, but for interval & ratio data they are generally less powerful and less flexible and should only be used where the standard parametric test is not appropriate—e.g., when the sample size is small [2]. Sample size calculation or power analysis is directly related to the statistical technique that is chosen, because the sample size calculation is based on the power (typically 0.80 is desired), and the effect size (typically a medium or large effect are selected; the larger the effect, the smaller a sample is needed) [1] [2] [3].

2. Supervised Learning Methods vs Unsupervised Learning Methods

Machine learning uses two types of techniques: supervised learning, which construct a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns in input data.

The goal of supervised learning methods is to build a model that makes predictions based on evidence in the data. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response. Supervised learning uses classification and regression techniques to develop predictive models [1] [2] [3].

Examples of Supervised Learning:

1) House prices:

If there is data about the houses, such as the square footage, number of rooms, features, whether a house has a garage or not, and so on. We then need to know the prices of these houses by leveraging data coming from thousands of houses, with their features and prices. Now, we could train a supervised learning model to predict a new house's price based on the examples observed by the model.

2) Weather condition:

In order to make correct predictions for the weather, we need to consider various inputs. For instance, historical temperature data, amount of precipitation, wind, snow, and humidity. In this situation, we could predict tomorrow's

temperature by a regression model. We could also predict the weather condition whether it is going to snow or not tomorrow by a binary classification problem.

Unsupervised learning methods find hidden patterns or intrinsic structures in data. It is used to draw inferences from datasets consisting of input data without labeled responses. Clustering and association are the most common unsupervised learning technique. It is used for exploratory data analysis to find hidden patterns or groupings in data. Clustering will split the dataset into groups based on their similarities. Association will identify sets of items which often occur together in the dataset [1] [2] [3].

Unsupervised learning uses unlabeled dataset with unknown output values for all the input values. As there are no known output values that can be used to build a logical model between the input and output, some techniques are used to mine data rules, patterns, and groups of data with similar types. These groups help the end-users to understand the data better as well as find a meaningful output. Once a model learns to develop patterns, it can predict patterns for any new datasets. This process does not figure out the right output, but it explores the data and draw inferences to describe the hidden structures in the data. All features in the dataset are equally important [1] [2] [3] [4].

Clustering, and association are examples of unsupervised learning. Clustering will split the dataset into groups based on their similarities. Association will identify sets of items which often occur together in the dataset.

Examples of Unsupervised learning:

1) Market Segmentation Analysis:

Some companies use this process to segment its customers to better adjust products and offerings. In this process, we might recognize multiple characteristics for potential customers, such as zip code, family income, and shopping habits. We might assume that the customers fall into different groups, such as high spenders versus low spenders. If the information about each customer's spending patterns were available, then a supervised analysis would be possible. However, this information is not available, so we can try to cluster the customers on the basis of the variables measured, such as similar behavior or demographic information.

2) Bank's Loan Repaying:

If we want to predict how capable an applicant is of repaying a loan from a perspective bank, then we need to collect a lot of information about each application to make predictions, such as the applicant's average monthly income, debt, and credit history. However, not all of it is relevant for predicting an applicant's credit risk score. For instance, does an applicant's age make any difference while deciding whether the applicant can repay the loan? Is the applicant's gender important for determining the credit risk score? Probably not. Hence, it is important to understand that not every feature adds value to modeling the problem. Therefore, eliminating these features is an essential part of the unsupervised learning through the feature selection process.

3) People that buy a new house also tend to buy new furniture. This is an example of association in unsupervised learning.

The main differences between supervised and unsupervised learning methods are summarized in **Table 1** [1] [2] [3] [4].

3. Regression vs Classification

Both regression and classification problems are examples of supervised learning method, in which [1] [2] [3] [4] [5]:

- Regression problems are used to predict output values based on previous data observations. In regression problems, we often use quantitative variables that take on numerical values, such as a person's age, height, or income, the value of a house, and the price of a stock. In other words, Regression techniques predict continuous responses, for example, changes in temperature or fluctuations in power demand.
- Classification problems are used where the output variable can be categorized, such as yes or no, pass or fail. In classification problems, we often use qualitative variables that take on values in one of different classes, or categories, such as a person's gender (male or female), and the type of product purchased (type A, B, or C). The classification problems act similar to a classifier that can have two or more levels, and the levels may or may not be ordinal. In other words, Classification techniques predict categorical responses, for example, whether an email is genuine or spam [1] [2] [3] [4] [5].

Table 1. Comparison between supervised learning and unsupervised learning methods.

Supervised Learning	Unsupervised Learning
Deals with known and labelled data	Deals with unknown and unlabeled data
Input variables and output variables are specified	Only input data are specified
The ultimate goal is to determine the function so well that when new input dataset is given, then it can predict the output	The ultimate goal is to find the hidden patterns or underlying structure in the given input data in order to learn about the data
Uses training data to learn a relationship between the input and the outputs	Does not use output data
It is a Predictive Modeling technique which predicts the future outcomes	It is a Descriptive Modeling technique which explains the hidden relationship between the data elements
More accurate results are obtained as input data and corresponding output are well known, and the software only needs to give predictions	Less accurate results are obtained as the input data are unlabeled. Thus, the software has to first understand and label the data and then give predictions
Learning method takes place offline	Learning method takes place in real time
It includes classification and regression algorithms	It includes clustering and association algorithms
Complex in Computation	Less Computational Complexity

There are different types of Regression in machine leaning, including the following types [4] [5]:

1) Linear Regression
2) Multiple Regression
3) Polynomial Regression
4) Logistic Regression
5) Quantile Regression
6) Stepwise Regression
7) Ridge Regression
8) Lasso Regression
9) Elastic Net Regression
10) Principal Components Regression
11) Support Vector Regression
12) Ordinal Regression
13) Poisson Regression
14) Negative Binomial Regression
15) Quasi Poisson Regression
16) Cox Regression
17) Tobit Regression
18) Bayesian Regression
19) Least Absolute Deviation (LAD) Regression

There are different types of classification algorithms in machine learning, including [4] [5] [6] [7] [8]:

1) Naive Bayes Classifier
2) Logistic Regression
3) Decision Tree
4) Random Forests
5) Support Vector Machines
6) K-Nearest Neighbor
7) K-Means Clustering

Some methods can be used for both regression and classification, and they have same or close names as shown in **Table 2** [5] [6] [7] [9] [10].

Table 2. Common methods used in both regression and classification.

Regression	Classification
Simple Linear Regression	Binary Logistic Regression
Multiple Linear Regression	Multinomial Logistic Regression
Support Vector Regression	Support Vector Machine
Decision Tree Regression	Decision Tree Classification
Random Forest Regression	Random Forest Classification
Neural Network	Neural Network

4. Can Linear Regression Be Used for Classification? Answer: No

Linear regression model is not suitable for a classification problem for the following reasons [1]-[10]:

4.1. Model Output

Linear regression is suitable for predicting output that is continuous value, such as predicting the price of a property, the age or weight of a person. Its prediction output can be any real number, ranging from negative infinity to infinity. Classification is used when the dependent variable (target) is categorical/qualitative. For example, predict whether a customer will make a purchase or not, or if an email is spam (1) or not spam (0). So, classification predicts a probability range between 0 to 1.

4.2. Model Objective

The objective of a linear regression model is to find a relationship between the input variables and a target variable. It predicts the best fit line for the dataset, which aims to minimize the distance between the predicted value and actual value. The objective of a classification model is to classify or categorize the target variable into classes or labels based on input variables.

4.3. Range of Predicted Values

In classification problems, we are interested in the probability of an outcome occurring. Probability is ranged between 0 and 1, where the probability of something certain to happen is 1, and 0 is something unlikely to happen. But in linear regression, we are predicting an absolute continuous number, which can range outside 0 and 1. For example, if we use linear regression to model a binary outcome, it is entirely possible to have a fitted regression line which is outside the (0,1) range or probabilities.

4.4. Sensitivity to Unbalanced Observations

When observation in one class is higher than the observation in other classes then there exists a class imbalance. This is common in classification problems, such as fraud detection, spam filtering and disease screening. Linear regression is sensitive to unbalanced data, as it might predict the majority class with high accuracy but fail to capture the minority class.

4.5. Variance of Residual Errors

Linear regression assumes that the variance of random errors is constant (also called homoscedasticity). In classification models, both mean and variance depend on the underlying probability. Any factor that affects the probability will change not just the mean but also the variance of the observations, which means the variance is no longer constant. As a result, we cannot directly apply linear

regression because it will not be a good fit.

4.6. The Distance Ordering among Predictors

Linear regression falsely assumes an equal-distance order among the predictors. For instance, if we coded three medical symptoms as 1, 2, and 3, the order in which the symptoms are labeled may affect the outcome of the linear regression model, which is intuitively unreasonable. Even if we were to convert the three values into three binary predictors (*i.e.*, whether a patient has a symptom), the linear regression will still suffer from the interpretation problem, because when the fitted value comes beyond the [0,1] interval, it is difficult to interpret the result as a probability (e.g., how likely is a patient diabetic).

For Example:

Suppose that we are trying to predict the severity level of vehicle accidents at a road. In this example, there are three possible severity outcomes of the accidents: fatality, injury, and vehicle damage without any fatality or injury (also called property damage only, PDO). We could encode these outcomes as 1 if fatal, 2 if injury, 3 if PDO. Using this coding, least squares could be used to fit a linear regression model to predict the severity level on the basis of a set of predictors, such as vehicle speed, vehicle type, driver's age, driver's gender, width of road lanes, etc. Unfortunately, this coding implies an ordering on the outcomes. However, in practice this is not always the case. For instance, if we choose different order coding, which would imply a different relationship among the three conditions. Each of these coding would produce different linear models that would lead to different sets of predictions on test observations. For this reason, the linear regression model is not suitable, and proper classification models are necessary to categorize the output feature.

5. Linear Regression vs Logistic Regression

Linear Regression and Logistic Regression are the two Machine Learning Algorithms which used in supervised learning. Since both the algorithms are supervised in nature hence these algorithms use labeled dataset to make the predictions. But the main difference between them is how they are being used. The Linear Regression is used for solving Regression problems whereas Logistic Regression is used for solving the Classification problems [1] [2] [3].

Linear regression is used for predicting the continuous dependent variable from the independent variables. The goal of the Linear regression is to find the best fit line that can predict the output for the continuous dependent variable. If single independent variable is used for prediction, then it is called Simple Linear Regression and if there are more than two independent variables then it is called Multiple Linear Regression. By finding the best fit line, algorithm establishes the relationship between dependent variable and independent variable. And the relationship should be of linear nature. The output for Linear regression should only be the continuous values such as price, age, salary, etc. [1]-[14].

Logistic regression is used for Classification problems. The output of Logistic Regression problem can be only between the 0 and 1. Logistic regression can be used where the probabilities between two classes is required, either 0 or 1, true or false, etc. Logistic regression is based on the concept of Maximum Likelihood estimation, which uses the sigmoid function and the curve obtained is called as sigmoid curve or S-curve. **Table 3** shows the main differences between Linear regression and Logistic regression [1]-[14].

6. Linear Discriminant Analysis (LDA) vs Logistic Regression (LR)

Linear discriminant analysis (LDA) and logistic regression (LR) are both used for classification problems in machine learning. Assumptions of multivariate normality and equal variance-covariance matrices across groups are required before proceeding with LDA, but such assumptions are not required for LR and hence LR is considered to be much more robust than LDA. While both methods are applicable in many instances, it is important to understand the key differences between them [9]-[14]:

- While both techniques require a categorical dependent variable, LR is preferred when the dependent variable is dichotomous (Dichotomous variables are categorical variables with two categories or levels), while LDA is preferred when it is nominal (more than two groups).
- LR accepts continuous as well as categorical predictor variables while DFA accepts only continuous (or dummy) and no categorical predictors.
- LR is more appropriate when the researcher is interested in the underlying structure of the prediction ("what are the most important predictors?" or "what is the role that different variables play in the prediction), rather than in the specific prediction of which group people belong to which is the emphasis of LDA.

Table 3. Linear regression vs Logistic regression.

Linear Regression	Logistic Regression
Used for solving Regression problem	Used for solving Classification problems
Predict the value of continuous variables	Predict the values of categorical variables
Find the best fit line, which can easily predict the output	Find the S-curve, which can easily classify the samples
Least square estimation method is used for estimation of accuracy	Maximum likelihood estimation method is used for estimation of accuracy
The output must be a continuous value, such as price, age, etc.	The output must be a categorical value such as 0 or 1, Yes or No, etc.
It is required that relationship between dependent variable and independent variable must be linear	It is not required to have the linear relationship between the dependent and independent variable
There may be some collinearity between the independent variables	There should not be any collinearity between the independent variable

- LDA requires multivariate normality while LR is robust against deviations from normality.
- LR generally requires a larger sample size than LDA.
- When classes are well-separated, parameter estimates for logistic regression LR will be unstable. LDA does not suffer from this problem.
- If sample size n is small and distribution of X is approximately normal in each of classes, LDA is more stable than logistic regression.

7. Linear Discriminant Analysis (LDA) vs Quadratic Discriminant Analysis (QDA) vs K-Nearest Neighbor (KNN)

Linear Discriminant analysis (LDA) is a method that can be used for both classification and dimensionality reduction (*i.e.*, reduce the number of features to a more manageable number before classification) in machine learning. Quadratic discriminant analysis (QDA) is a variant of LDA that allows for non-linear data [1] [2] [3] [4] [5].

Both LDA and QDA approaches assume that the observations from each class are drawn from a normal distribution. But, in LDA, we assume the normal distributions for different classes have a common variance-covariance matrix, whereas in QDA, we assume the normal distributions for different classes have different variance-covariance matrices. QDA will work best when the variances are very different between classes and we have enough/very-large observations to accurately estimate the variances. LDA will work best when the variances are similar among classes or we don't have enough data to accurately estimate the variances. When the number of predictors is large the number of parameters, we have to estimate with QDA becomes very large because we have to estimate a separate covariance matrix for each class. This can lead to high variance and so we have to be careful when using QDA. Therefore, the LDA is less flexible than QDA because we have to estimate fewer parameters. This can be good when we have only a few observations in our training data set so we lower the variance. However, when the assumption of a common covariance matrix is violated, or if the training set is large, then LDA will suffer from high bias and QDA might be a better choice [15]-[21].

The abbreviation KNN stands for "K-Nearest Neighbor". It is a supervised machine learning algorithm. The algorithm can be used to solve the classification problem statements. The number of nearest neighbors to a new unknown variable that has to be classified is denoted by the symbol "K". The KNN algorithm employs the same principle. Its aim is to locate all of the closest neighbors around a new unknown data point in order to figure out what class it belongs to. It's a distance-based approach.

KNN is a non-parametric approach; no assumptions are made about the shape of the decision boundary. Therefore, we can expect KNN to perform better than LDA and logistic regression when the decision boundary is non-linear.

However, KNN does not show which predictors are important; we do not get a table of coefficients [17]-[22].

QDA could be a compromised method between the non-parametric KNN method and the linear LDA and logistic regression approaches, because QDA assumes a quadratic decision boundary, so it can accurately model a wider range of problems than can the linear methods [1] [2] [3] [13] [19] [20] [21].

Despite the fact that KNN is more flexible than QDA, however, QDA can still perform better than KNN if there are a limited number of training observations because it does make some assumptions about the form of the decision boundary [1] [2] [3] [4] [20] [21].

8. Ridge Regression vs LASSO

Ridge regression and lasso are both regularization (shrinkage) methods. LASSO regression stands for "Least Absolute Shrinkage and Selection Operator". One problem that often occurs in practice with multiple linear regression is multicollinearity when two or more predictor variables are highly correlated to each other, such that they do not provide unique or independent information in the regression model. This can cause the coefficient estimates of the model to be unreliable and have high variance. Two methods we can use to get rid of this issue of multicollinearity are ridge regression and lasso regression. The benefit of ridge and lasso regression compared to least squares regression lies in the bias-variance tradeoff. We know that the mean squared error (MSE) is a metric we can use to measure the accuracy of a given regression model. The basic idea of both ridge and lasso regression is to introduce a little bias so that the variance can be substantially reduced, which leads to a lower overall MSE. Both lasso regression and ridge regression are known as regularization methods because they both attempt to minimize the sum of squared residuals (RSS) along with some penalty term. In other words, they constrain or regularize the coefficient estimates of the model [1] [2] [3].

Ridge regression is an extension for linear regression. It's basically a regularized linear regression model. An important fact about ridge regression is that it enforces the β coefficients to be lower to avoid over-fitting, but it does not enforce them to be zero. That is, it will not get rid of irrelevant features but rather minimize their impact on the model [1] [2] [3] [15]-[20].

Lasso is another extension built on regularized linear regression, but with a small twist. Lasso works by not only punishing high values of the coefficients β but actually setting them to zero if they are not relevant. Therefore, we might end up with fewer features included in the model than we started with.

We expect that lasso would perform better if there were a relatively small number of predictors. Ridge regression would perform better when there were many predictors. But since the number of predictors is usually unknown in real data sets, therefore cross-validation can be used in order to determine which approach is better on a particular data set [1] [2] [3] [16] [17] [18] [19].

The ridge regression cannot zero out coefficients; thus, it either end up including all the coefficients in the model, or none of them. In contrast, the LASSO does both parameter shrinkage and variable selection as it can zero out some coefficients in the model [1] [2].

The lasso produces simpler and more interpretable models because it involves only a subset of the predictors. In contrast, the ridge regression includes all the coefficients in the model, and hence produces a complex model that is less interpretable [1] [2].

Since the lasso assumes that a number of the coefficients equal zero. Hence, we expect that ridge regression outperforms the lasso in terms of prediction error, because ridge regression will not exclude any predictors from the model [1] [2].

When there is excessively high variance, the lasso can produce a reduction in variance at the expense of a small increase in bias, and hence lasso can generate more accurate predictions than the ridge regression [1] [2] [3] [4].

In both the ridge and lasso, we use a grid of λ values, and compute the cross-validation error rate for each value of λ. We then select the tuning parameter value for which the cross-validation error is smallest [1] [2].

Ridge regression uses L2 penalty, which adds "squared magnitude" of coefficients as penalty term to the loss function. Lasso uses L1 penalty, which adds "absolute value of magnitude" of coefficients as penalty term to the loss function. L1 can force some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large [1] [2] [3] [4] [5].

9. Bagging vs Boosting vs Random Forest

Bagging, random forest, and boosting are all ensemble techniques where a set of weak learners are combined to create a strong learner that obtains better performance than a single one. Bagging, random forests, and boosting can use trees as building blocks to construct better prediction models. The main causes of error in learning are due to noise, bias, and variance. Ensemble methods help in minimizing these factors. These methods are designed also to improve the stability and the accuracy of Machine Learning algorithms [22]-[32].

Bootstrap aggregation, or bagging, is a suitable procedure for reducing the variance of a statistical learning method, especially within the context of decision trees. Bagging involves creating multiple copies of the original training data set using the bootstrap, fitting a separate decision tree to each copy, and then combining all of the trees in order to create a single predictive model. Bagging can improve the accuracy by combining together hundreds or even thousands of trees into a single procedure. However, it can be difficult to the resulting model. Thus, bagging improves prediction accuracy at the expense of interpretability. But we can obtain an overall summary of the importance of each predictor using the RSS (for bagging regression trees) or the Gini index (for bagging classification trees) [22]-[32].

Random Forest is an extension over bagging, where in addition to taking the random subset of data, it also takes the random selection of features rather than using all features to grow trees. Random forests can create an improvement over bagging by using a random small tweak that decorrelates the trees. The predictions from the bagged trees will usually be highly correlated, random forests overcome this problem by forcing each split to consider only a subset of the predictors. In addition, random forests use a small subset size compared to bagging, which could lead to a reduction in both test error and OOB error over bagging. Random forests can handle higher dimensionality data very well, and also maintain accuracy for missing data [22]-[32].

Boosting is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model. Thus, each model boosts the performance of the ensemble. In other words, while the training stage is parallel for Bagging (*i.e.*, each model is built independently), Boosting builds the new learner in a sequential way. Bagging and Boosting get N new learners by generating additional data in the training stage. N new training data sets are produced by random sampling with replacement from the original set. By sampling with replacement some observations may be repeated in each new training data set. In the case of Bagging, any element has the same probability to appear in a new data set. However, for Boosting the observations are weighted and therefore some of them will take part in the new sets more often. In Bagging the result is obtained by averaging the responses of the N learners (or majority vote). However, Boosting assigns a second set of weights, this time for the N classifiers, in order to take a weighted average of their estimates. Both Bagging and Boosting decrease the variance of the single tree estimate as they combine several estimates from different models. So, the result may be a model with higher stability. However, Boosting could generate a combined model with lower errors as it optimizes the advantages and reduces errors of the single model. By contrast, if the difficulty of the single model is over-fitting, then Bagging is the best option. Boosting will not help to avoid over-fitting. So, Bagging tries to recover over-fitting problem while Boosting tries to reduce bias [22]-[32].

10. Decision Tree vs Random Forest

Decision Trees are a non-parametric supervised learning method used for both classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. It is a flowchart-like diagram that shows the various outcomes from a series of decisions. It can be used as a decision-making tool, for research analysis, or for planning strategy. The main advantage of a decision tree is that it can be fit to a dataset quickly and the final model can be neatly visualized and interpreted using a "tree" diagram. The main disadvantage is that a decision tree tends to overfit a

training dataset, which means it's likely to perform poorly on unseen data. It can also be heavily influenced by outliers in the dataset [33]-[38].

Random Forest is a powerful supervised machine learning algorithm that combines multiple decision trees to create a "forest." It can be used for both classification and regression problems. The logic of the Random Forest model is that multiple uncorrelated models (the individual decision trees) perform much better as a group than they do alone. When using Random Forest for classification, each tree gives a classification. The forest chooses the classification with the majority of the outputs. When using Random Forest for regression, the forest picks the average of the outputs of all trees. That is there is low (or no) correlation between the individual models that make up the larger Random Forest model. While individual decision trees may produce errors, the majority of the group will be correct, thus moving the overall outcome in the right direction. The benefit of random forests is that they tend to perform much better than decision trees on unseen data and they're less affected by outliers. The disadvantage of random forests is that there's no way to visualize the final model and they can take a long time to build if the dataset is extremely large. **Table 4** compares decision trees and random forests [33]-[38].

11. Principal Component Analysis vs Clustering

Principal Component Analysis (PCA) is an unsupervised statistical technique that can be used for dimension reduction, feature extraction, and data visualization. PCA can analyze the data to identify patterns in order to reduce the dimensions of the dataset with minimal loss of information. High dimensionality means that the dataset has a large number of features, which could produce overfitting. PCA can also obtain important variables (in form of components) from a large set of variables available in a data set. It extracts low dimensional set of features by taking a projection of irrelevant dimensions from a high dimensional data set to capture as much information as possible. With fewer variables obtained, visualization also becomes much more meaningful. The new projected variables (principal components) are uncorrelated with each other and are ordered so that the first few components retain most of the variation present in the original variables. By reducing the dimensions of learning data sets, PCA provides an effective and efficient method for data description and classification. In

Table 4. Decision tree vs. random forest.

terms	Decision Tree	Random Forest
Interpretability	Easy	Harder
Overfitting	Likely	Unlikely
Outliers	Highly affected by outliers	Robust against outliers
Accuracy	Can vary	Higher accuracy
Computation	Quick to build	Slow to build

addition, PCA can be employed in exploratory data analysis to reveal outliers and departures from a normal distribution. Moreover, PCA is also useful for constructing predictive models, as in principal components analysis regression (also known as PCA regression) [39]-[46].

Example: considering a dataset composed by a set of properties from vehicle features. These properties describe each vehicle by its size, color, circularity, compactness, number of seats, number of doors, size of trunk and so on. However, many of these features will measure related properties and so will be redundant. Therefore, we should remove these redundancies and describe each vehicle with less properties. This is what PCA aims to do. For instance, considering the number of wheel as a feature of cars and buses, almost every example from both classes has four wheels, hence we can tell that this feature has a low variance, so this feature will make bus and cars look the same, but they are actually different from each other. If we consider the height as a feature, cars and buses have different values for it, the variance has a great range from the lowest car up to the highest bus. Clearly, the height of vehicle is a good property to separate them. Therefore, PCA will look at the variance of this feature to split between these two classes [39]-[46].

Clustering is an unsupervised learning method, in which we draw references from datasets consisting of input data without labeled responses. Clustering is used to find meaningful structure, explanatory processes, generative features, and groupings inherent in a set of data. This means that clustering would divide the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them [39]-[46].

Through the use of clustering, attributes of unique entities can be profiled easier. It can also help in dimensionality reduction if the dataset is comprised of too many variables. Irrelevant clusters can be identified easier and removed from the dataset. The main types of clustering in unsupervised machine learning include K-means, hierarchical clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Gaussian Mixtures Model (GMM).

In K-means clustering, data is grouped in terms of characteristics and similarities. K is a letter that represents the number of clusters. For example, if K = 15, then the number of desired clusters is 15. If K = 20, then the number of desired clusters is 20 [39]-[46].

The Hierarchical Clustering is used when constructing a hierarchy (of clusters). This algorithm will only end if there is only one cluster left. Unlike K-means clustering, hierarchical clustering doesn't start by identifying the number of clusters. Instead, it starts by allocating each point of data to its cluster. The representations in the hierarchy provide meaningful information. It doesn't require the number of clusters to be specified. However, Hierarchical models have an acute sensitivity to outliers. In the presence of outliers, the models don't perform well [39]-[46].

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering that involves the grouping of data points close to each other. We mark data points far from each other as outliers. It then sorts data based on commonalities. It doesn't require a specified number of clusters and offers flexibility in terms of the size and shape of clusters. However, it's not effective in clustering datasets that comprise varying densities [42]-[48].

The Gaussian Mixture Models (GMM) is an advanced clustering technique in which a mixture of Gaussian distributions is used to model a dataset. These mixture models are probabilistic. GMM clustering models are used to generate data samples, each data point is a member of all clusters in the dataset, but with varying degrees of membership. The probability of being a member of a specific cluster is between 0 and 1. In Gaussian mixture models, the information includes the latent Gaussian centers and the covariance of data. This makes it similar to K-means clustering. It offers flexibility in terms of size and shape of clusters. Also, membership can be assigned to multiple clusters, which makes it a fast algorithm for mixture models [42]-[48].

Comparing the PCA with Clustering, we can realize that the goal of the clustering algorithm is to partition the objects into homogeneous groups, such that the within-group similarities are large compared to the between-group similarities. The principal components, on the other hand, are extracted to represent the patterns encoding the highest variance in the data set and not to maximize the separation between groups of samples directly. The results from PCA and hierarchical clustering support similar interpretations. However, PCA represents the data set in only a few dimensions, some of the information in the data is filtered out in the process. The discarded information is associated with the weakest signals and the least correlated variables in the data set, and it can often be safely assumed that much of it corresponds to measurement errors and noise. This makes the patterns revealed using PCA cleaner and easier to interpret than those seen in the clustering techniques [42]-[48].

12. Conclusion

Selection of the correct statistical method is important in any research work. The wrong selection would cause inadequate findings. The researcher should be knowledgeable about the major differences between possible statistical methods that could be applied in research. This paper presented the main differences between supervised learning methods and unsupervised learning methods. Supervised learning methods build a model that makes predictions based on evidence in the data, while unsupervised learning methods draw inferences from datasets without labeled responses. The paper overviewed the differences between Linear Regression and Logistic Regression and showed that Linear Regression is used for solving Regression problems whereas Logistic Regression is used for solving the Classification problems. Also, the paper examined both Linear discriminant analysis (LDA) and logistic regression (LR), which are used for classification

problems in machine learning. It was clear that the assumptions of multivariate normality and equal variance-covariance matrices across groups are required before proceeding with LDA, but such assumptions are not required for LR and hence LR was considered to be more robust than LDA. The paper indicated that multicollinearity, when two or more predictor variables are highly correlated to each other, can cause the coefficient estimates of the model to be unreliable and have high variance. Two methods were suggested to get rid of the issue of multicollinearity, namely ridge regression and lasso regression. The paper also showed that bagging, random forest, and boosting are all ensemble techniques where a set of weak learners are combined to create a strong learner that obtains better performance than a single one. These ensemble methods were designed to improve the stability and the accuracy of Machine Learning algorithms. In addition, the paper explained both decision trees and random forest algorithms. It showed that decision trees are non-parametric supervised learning methods used for both classification and regression. The tree is a flowchart-like diagram that shows the various outcomes from a series of decisions. It can be used as a decision-making tool, for research analysis, or for planning strategy. Random Forest is a powerful supervised machine learning method that combines multiple decision trees to create a "forest." It can be used for both classification and regression problems. Lastly, the paper shadowed on both the Principal Component Analysis (PCA) and Clustering as unsupervised statistical techniques. It was explained that PCA can be used for dimension reduction, feature extraction, and data visualization. PCA can analyze the data to identify patterns in order to reduce the dimensions of the dataset with minimal loss of information. Clustering on the other hand could be used to find meaningful structure, explanatory processes, generative features, and groupings inherent in a set of data, which means that clustering would divide the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Gareth, J., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning: With Applications in R. Springer, Berlin, Heidelberg.

[2] Hastie, T., Tibshirani, R. and Friedman, J. (2008) The Elements of Statistical Learning. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7

[3] Brett, L. (2015) Machine Learning with R. Packt Publishing Ltd., Birmingham.

[4] Freedman, D.A. (2009) Statistical Models: Theory and Practice. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511815867

[5] Venables, W.N. and Ripley, B.D. (2002) Modern Applied Statistics in S. 4th Edition, Springer, New York. https://doi.org/10.1007/978-0-387-21706-2

[6] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[7] Washington, S.P., Karlaftis, M.G. and Mannering, F. (2010) Statistical and Econometric Methods for Transportation Data Analysis. 2nd Edition, Chapman Hall/CRC, Boca Raton.

[8] Abdulhafedh, A. (2017) Road Crash Prediction Models: Different Statistical Modeling Approaches. *Journal of Transportation Technologies*, **7**, 190-205. https://doi.org/10.4236/jtts.2017.72014

[9] Passos, I.C., Mwangi, B. and Kapczinski, F. (2016) Big Data Analytics and Machine Learning and beyond. *Lancet Psychiatry*, **3**, 13-15. https://doi.org/10.1016/S2215-0366(15)00549-0

[10] Abdulhafedh, A. (2017) Incorporating the Multinomial Logistic Regression in Vehicle Crash Severity Modeling: A Detailed Overview. *Journal of Transportation Technologies*, **7**, 279-303. https://doi.org/10.4236/jtts.2017.73019

[11] Heinze, G. and Schemper, M. (2002) A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine*, **21**, 2409-2419. https://doi.org/10.1002/sim.1047

[12] Abdulhafedh, A. (2022) Incorporating Multiple Linear Regression in Predicting the House Prices Using a Big Real Estate Dataset with 80 Independent Variables. *Open Access Library Journal*, **9**, Article No. e8346. https://doi.org/10.4236/oalib.1108346

[13] Abdulhafedh, A. (2016) Crash Severity Modeling in Transportation Systems. PhD Dissertation, University of Missouri, Columbia, MO, USA. https://mospace.umsystem.edu/xmlui/browse?authority=b5818edd-97e5-439f-a994-206bab12f712&type=author

[14] Yeh, I.-C. and Lien, C.-H. (2009) The Comparisons of Data Mining Techniques for the Predictive Accuracy of Probability of Default of Credit Card Clients. *Expert Systems with Applications*, **36**, 2473-2480. https://doi.org/10.1016/j.eswa.2007.12.020

[15] Abdulhafedh, A. (2017) Road Traffic Crash Data: An Over-view on Sources, Problems, and Collection Methods. *Journal of Transportation Technologies*, **7**, 206-219. https://doi.org/10.4236/jtts.2017.72015

[16] Abdulhafedh, A. (2021) Vehicle Crash Frequency Analysis Using Ridge Regression. *International Journal for Science and Advance Research in Technology*, **7**, 254-261.

[17] Pasha, G.R. and Shah, M.A. (2004) Application of Ridge Regression to Multicollinear Data. *Journal of Research* (*Science*), **15**, 97-106.

[18] Lin, T.-H., Li, H.-T. and Tsai, K.-C. (2004) Implementing the Fisher's Discriminant ratio in a k-Means Clustering Algorithm for Feature Selection and Data Set Trimming. *Journal of Chemical Information and Modeling*, **44**, 76-87. https://doi.org/10.1021/ci030295a

[19] Wiener, A.L.A.M. (2002) Classification and Regression by Random Forest. *R News*, **2**, 18-22.

[20] Agresti, A. (2013) Categorical Data Analysis. John Wiley & Sons, New Jersey.

[21] Dobson, A.J. and Barnett, A. (2008) An Introduction to Generalized Linear Models. CRC Press.

[22] Abdulhafedh, A. (2016) Crash Frequency Analysis. *Journal of Transportation Technologies*, **6**, 169-180. https://doi.org/10.4236/jtts.2016.64017

[23] Lorenzen, T.J. and Anderson, V.L. (1993) Design of Experiments: A No-Name Approach. CRC Press. https://doi.org/10.1201/9781482277524

[24] Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978) Statistics for Experimenters. John Wiley & Sons, New York.

[25] Cochran, W.G. and Cox, G.M. (1992) Experimental Designs. John Wiley & Sons, New York.

[26] Lawless, J.F. (2002) Statistical Models and Methods for Lifetime Data. John Wiley & Sons, New York. https://doi.org/10.1002/9781118033005

[27] Miller, R. (1998) Survival Analysis. John Wiley & Sons, New York.

[28] Bajwa, S.J. (2015) Basics, Common Errors and Essentials of Statistical Tools and Techniques in Anesthesiology Research. *Journal of Anaesthesiology Clinical Pharmacology*, **31**, 547-553. https://doi.org/10.4103/0970-9185.169087

[29] Kim, K.S., Choi, H.H., Moon, C.S. and Mun, C.W. (2011) Comparison of k-Nearest Neighbor, Quadratic Discriminant, and Linear Discriminant Analysis in Classification of Electromyogram Signals Based on the Wrist-Motion Directions. *Current Applied Physics*, **11**, 740-745. https://doi.org/10.1016/j.cap.2010.11.051

[30] Abdulhafedh, A. (2021) Incorporating K-Means, Hierarchical Clustering and PCA in Customer Segmentation. *Journal of City and Development*, **3**, 12-30.

[31] Wang, S., Li, D., Song, X., Wei, Y. and Li, H. (2011) A Feature Selection Method Based on Improved Fisher's Discriminant Ratio for Text Sentiment Classification. *Expert Systems with Applications*, **38**, 8696-8702. https://doi.org/10.1016/j.eswa.2011.01.077

[32] Abdulhafedh, A. (2017) How to Detect and Remove Temporal Autocorrelation in Vehicular Crash Data. *Journal of Transportation Technologies*, **7**, 133-147. https://doi.org/10.4236/jtts.2017.72010

[33] Sexton, J. and Laake, P. (2009) Standard Errors for Bagged and Random Forest Estimators. *Computational Statistics & Data Analysis*, **53**, 801-811. https://doi.org/10.1016/j.csda.2008.08.007

[34] Tanha, J., van Someren, M. and Afsarmanesh, H. (2015) Semi-Supervised Self-Training for Decision Tree Classifiers. *International Journal of Machine Learning and Cybernetics*, **8**, 355-370. https://doi.org/10.1007/s13042-015-0328-7

[35] Chapelle, O., Sindhwani, V. and Keerthi, S. (2008) Optimization Techniques for Semi-Supervised Support Vector Machines. *Journal of Machine Learning Research*, **9**, 203-233.

[36] Joachims, T. (1999) Making Large Scale SVM Learning Practical. In: S*upport Vector Learning*, MIT Press. Cambridge, 169-184.

[37] Maaten, L.V.D. (2014) Accelerating t-SNE Using Tree-Based Algorithms. *Journal of Machine Learning Research*, **15**, 3221-3245.

[38] Athey, S., Tibshirani, J. and Wager, S. (2019) Generalized Random Forests. *The Annals of Statistics*, **47**, 1148-1178. https://doi.org/10.1214/18-AOS1709

[39] Scornet, E., Biau, G. and Vert, J.-P. (2015) Consistency of Random Forests. *The Annals of Statistics*, **43**, 1716-1741. https://doi.org/10.1214/15-AOS1321

[40] Gan, H., Sang, N., Huang, R., Tong, X. and Dan, Z. (2013) Using Clustering Analysis to Improve Semi-Supervised Classification. *Neurocomputing*, **101**, 290-298. https://doi.org/10.1016/j.neucom.2012.08.020

[41] Altman, D.G. and Bland, J.M. (2009) Parametric vs. Non-Parametric Methods for Data Analysis. *BMJ*, **338**, Article No. a3167. https://doi.org/10.1136/bmj.a3167

[42] Afifi, A., Clark, V.A. and May, S. (2004) Computer-Aided Multivariate Analysis. 4th Edition, Chapman & Hall/CRC, Boca Raton.

[43] Abdulhafedh, A. (2017) A Novel Hybrid Method for Measuring the Spatial Autocorrelation of Vehicular Crashes: Combining Moran's Index and Getis-Ord *Gi Statistic. *Open Journal of Civil Engineering*, **7**, 208-221. https://doi.org/10.4236/ojce.2017.72013

[44] Johnson, R.A. and Wichern, D.W. (2007) Applied Mutivariate Statistical Analysis. 6th Edition). Pearson, Prentice Hall, New Jersey.

[45] Williams, G.J. (2011) Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Springer, New York.

[46] Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**, 1-22. https://doi.org/10.18637/jss.v033.i01

[47] Abdulhafedh, A. (2017) Identifying Vehicular Crash High Risk Locations along Highways via Spatial Autocorrelation Indices and Kernel Density Estimation. *World Journal of Engineering and Technology*, **5**, 198-215. https://doi.org/10.4236/wjet.2017.52016

[48] Imbens, G. and Rubin, D.B. (2015) Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139025751