CLASSIFICATION AND REGRESSION

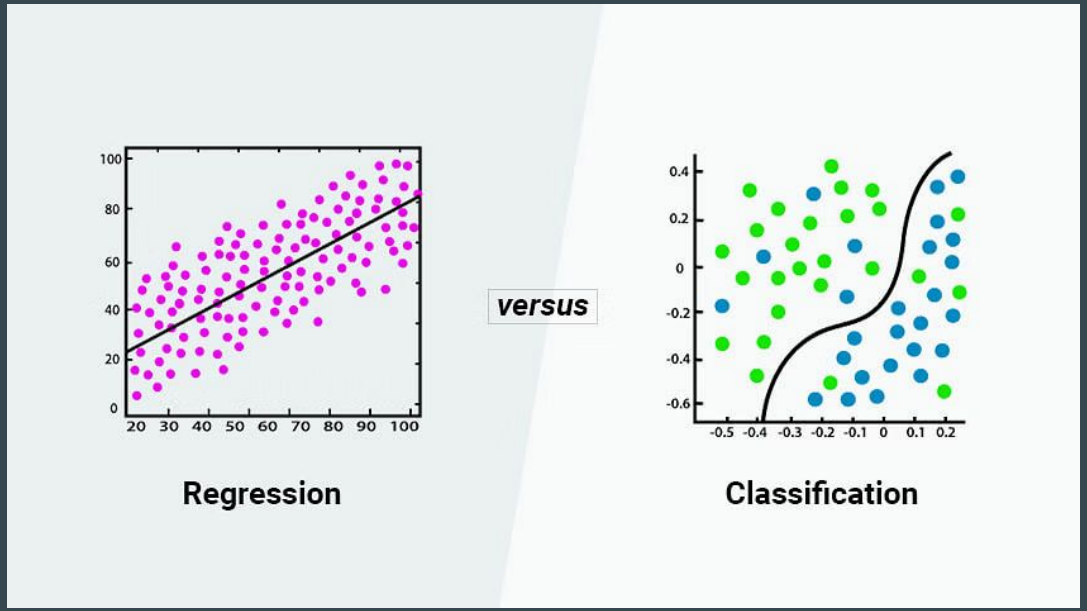
ROLL NUMBER	NAME
13000121058	ARKAPRATIM GHOSH

CA 1 : MACHINE LEARNING (PEC-CS701E)

CSE : SEMESTER 7

CONTENT

1. INTRODUCTION
2. REGRESSION AND TYPES
1. CLASSIFICATION AND TYPES
1. DIFFERENCES
2. CONCLUSION
3. REFERENCES



INTRODUCTION

Regression and Classification algorithms are Supervised Learning algorithms. Both the algorithms are used for prediction in Machine learning and work with the labeled datasets. But the difference between both is how they are used for different machine learning problems.

The main difference between Regression and Classification algorithms that Regression algorithms are used to **predict the continuous** values such as price, salary, age, etc. and Classification algorithms are used to **predict/Classify the discrete values** such as Male or Female, True or False, Spam or Not Spam, etc.

REGRESSION

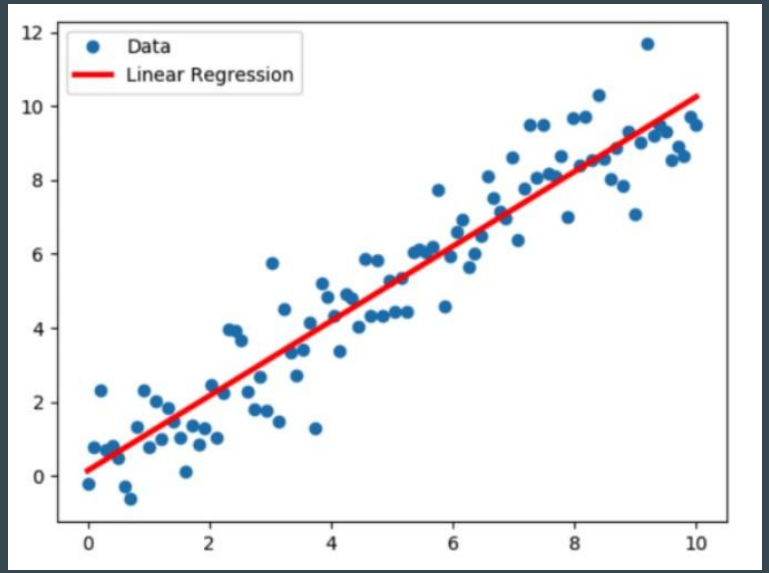
- **Definition**: Regression is a statistical method used to model and analyze the relationships between a dependent variable and one or more independent variables.
- **Purpose**: To predict continuous outcomes (e.g., prices, temperatures).
- **Types**: Linear regression, multiple regression, polynomial regression.
- **Example**: Predicting house prices based on features like size, location, and number of rooms.
- **Output**: Produces a continuous value.
- **Evaluation Metrics**: Mean squared error (MSE), root mean squared error (RMSE), R-squared.

TYPES OF REGRESSION

1. LINEAR REGRESSION

The most extensively used modelling technique is linear regression, which assumes a linear connection between a dependent variable (Y) and an independent variable (X). It employs a regression line, also known as a best-fit line. The linear connection is defined as Y = c+m*X + e, where 'c' denotes the intercept, 'm' denotes the slope of the line, and 'e' is the error term.

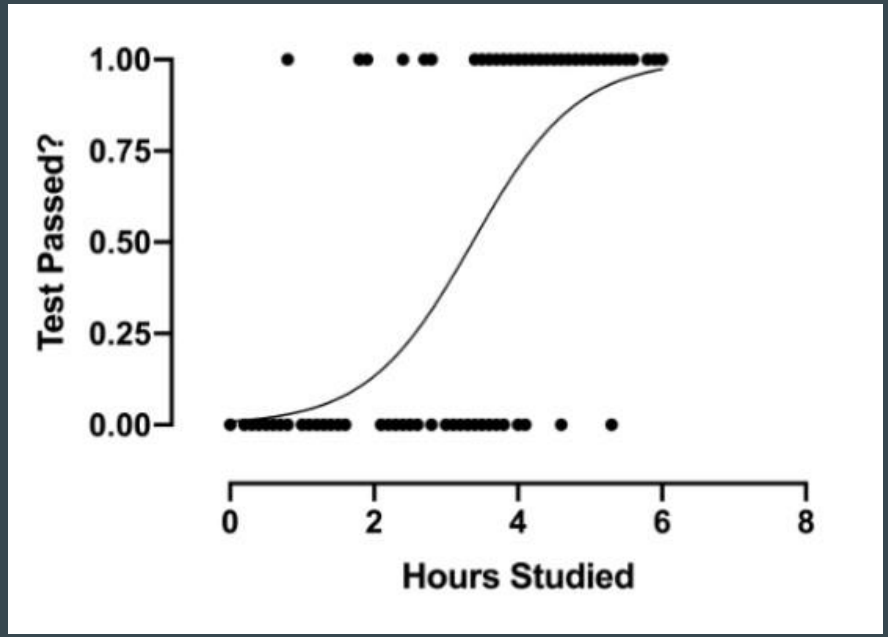
The linear regression model can be simple (with only one dependent and one independent variable) or complex (with numerous dependent and independent variables) (with one dependent variable and more than one independent variable).



TYPES OF REGRESSION

2. LOGISTIC REGRESSION

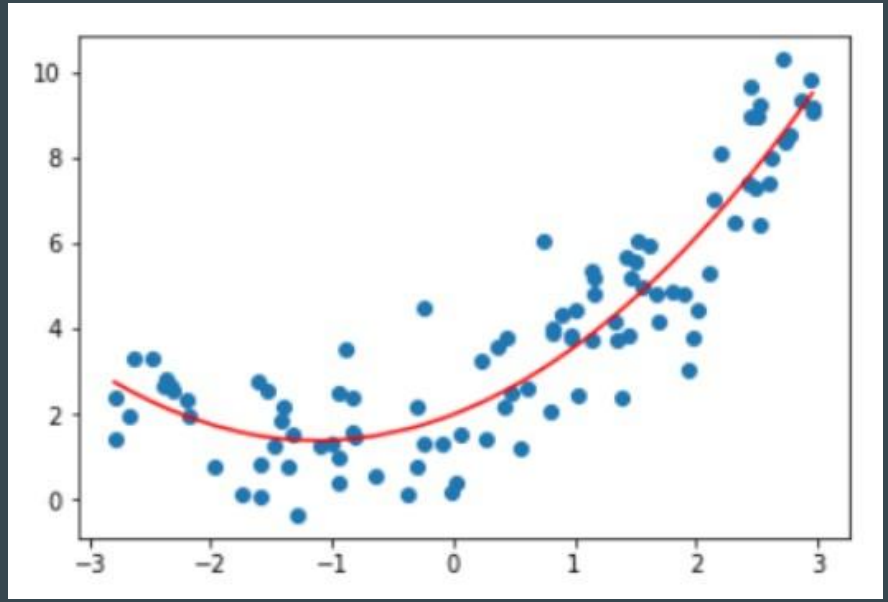
When the dependent variable is discrete, the logistic regression technique is applicable. In other words, this technique is used to compute the probability of mutually exclusive occurrences such as pass/fail, true/false, 0/1, and so forth. Thus, the target variable can take on only one of two values, and a sigmoid curve represents its connection to the independent variable, and probability has a value between 0 and 1.



TYPES OF REGRESSION

3. POLYNOMIAL REGRESSION

Polynomial regression analysis represents a nonlinear relationship between dependent and independent variables. This technique is a variant of the multiple linear regression model, but the best fit line is curved rather than straight.

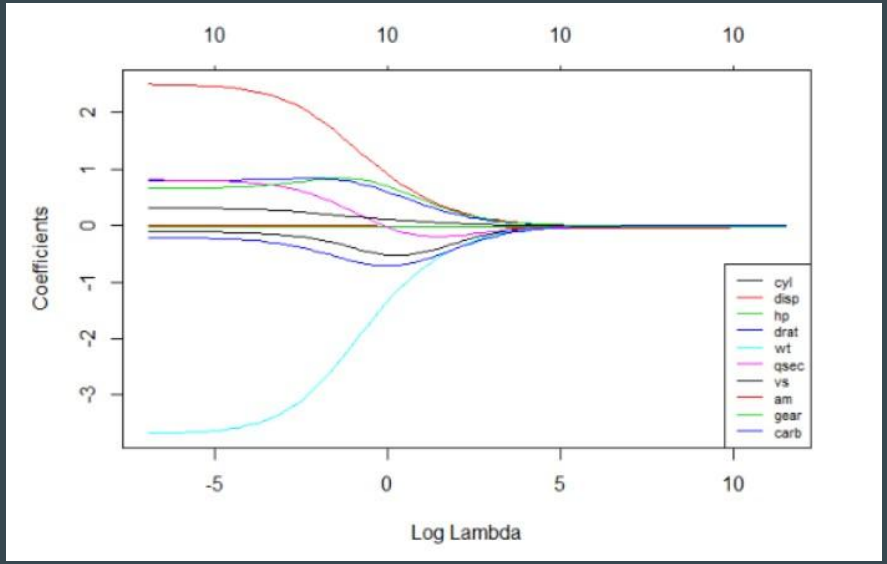


TYPES OF REGRESSION

4. RIDGE REGRESSION

When data exhibits multicollinearity, that is, the ridge regression technique is applied when the independent variables are highly correlated. While least squares estimates are unbiased in multicollinearity, their variances are significant enough to cause the observed value to diverge from the actual value. Ridge regression reduces standard errors by biassing the regression estimates.

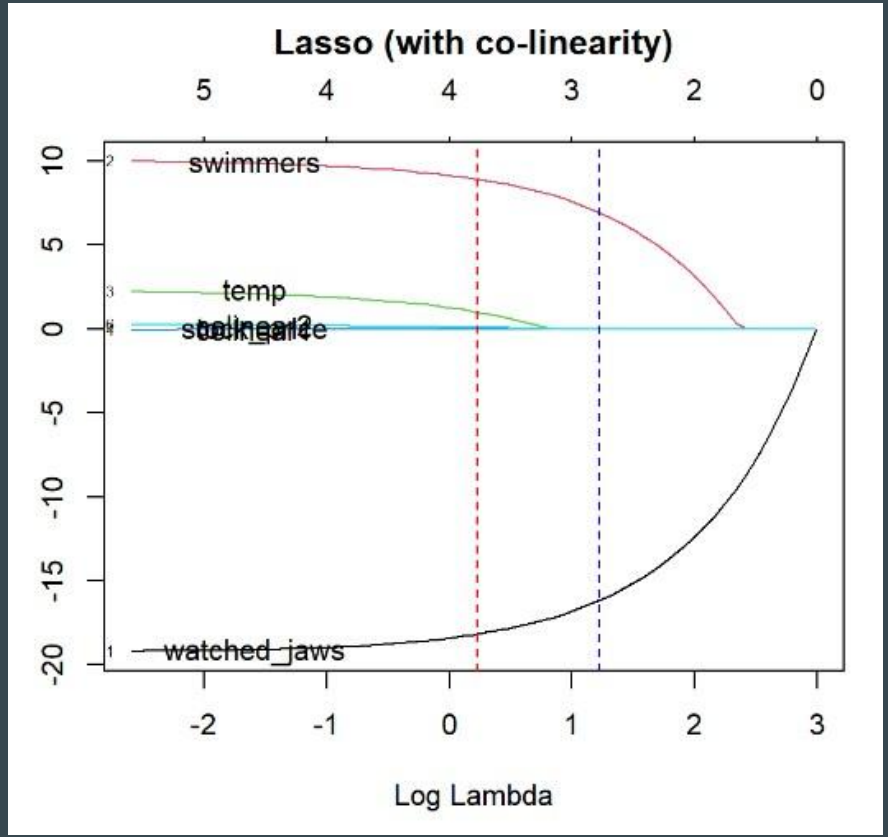
The lambda (λ) variable in the ridge regression equation resolves the multicollinearity problem.



TYPES OF REGRESSION

5. LASSO REGRESSION

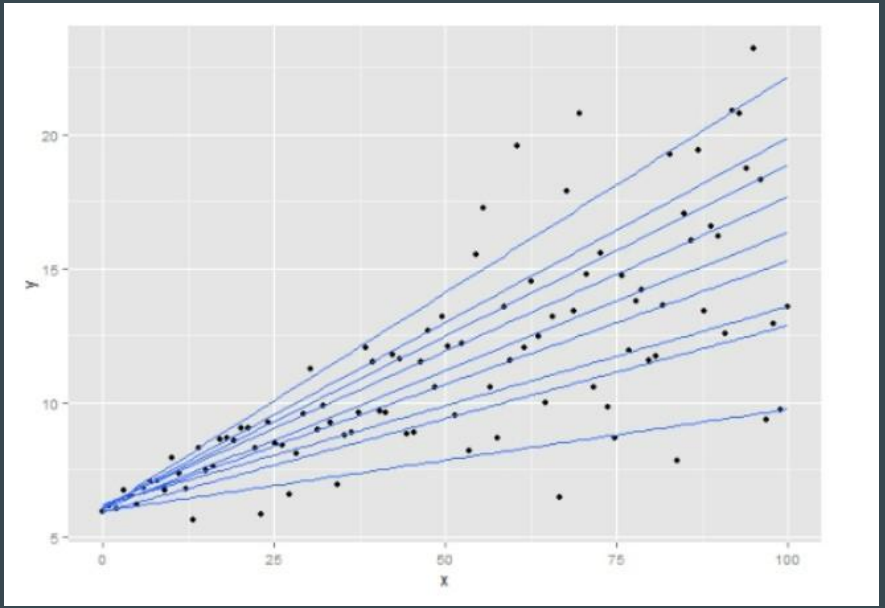
As with ridge regression, the lasso (Least Absolute Shrinkage and Selection Operator) technique penalizes the absolute magnitude of the regression coefficient. Additionally, the lasso regression technique employs variable selection, which leads to the shrinkage of coefficient values to absolute zero.



TYPES OF REGRESSION

6. QUANTILE REGRESSION

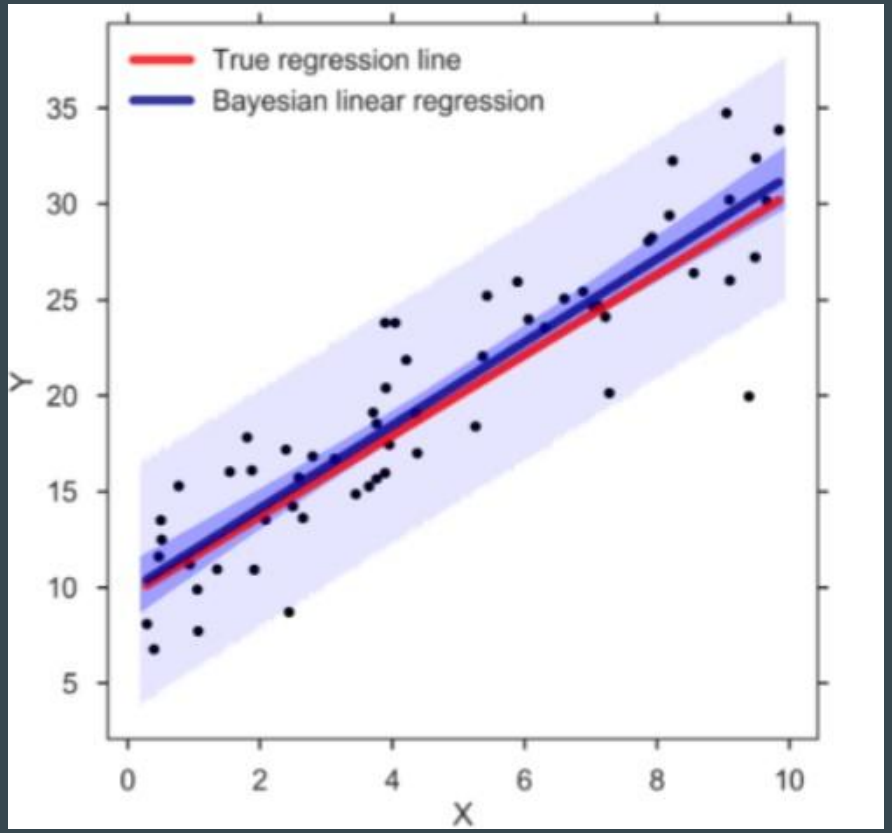
The quantile regression approach is a subset of the linear regression technique. Statisticians and econometricians employ quantile regression when linear regression requirements are not met or when the data contains outliers.



TYPES OF REGRESSION

7. BAYESIAN LINEAR REGRESSION

Machine learning utilizes Bayesian linear regression, a form of regression analysis, to calculate the values of regression coefficients using Bayes' theorem. Rather than determining the least-squares, this technique determines the features' posterior distribution. As a result, the approach outperforms ordinary linear regression in terms of stability.

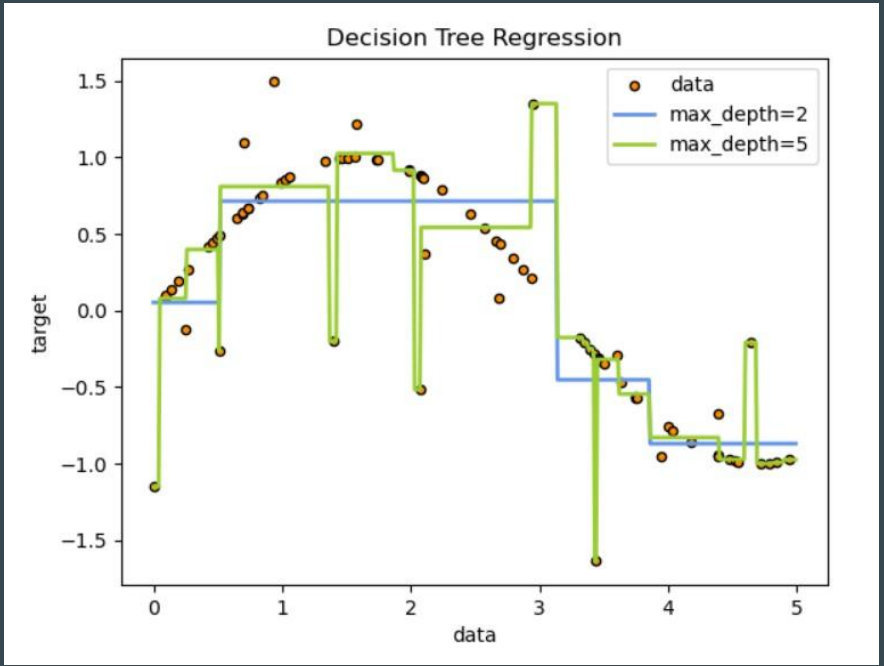


TYPES OF REGRESSION

8. DECISION TREE REGRESSION

The decision trees is used to fit a sine curve with addition noisy observation. As a result, it learns local linear regressions approximating the sine curve.

We can see that if the maximum depth of the tree (controlled by the max_depth parameter) is set too high, the decision trees learn too fine details of the training data and learn from the noise, i.e. they overfit.

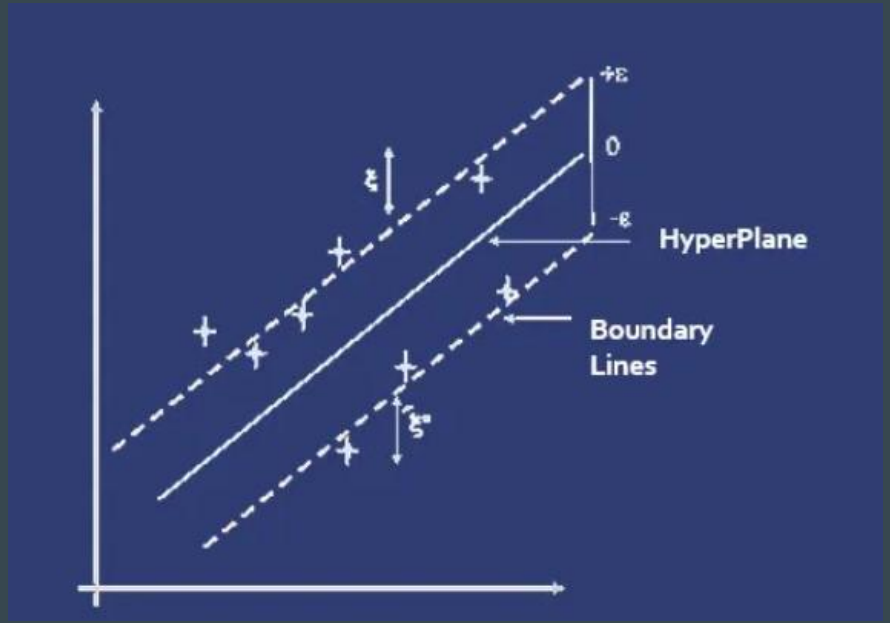


TYPES OF REGRESSION

9. SUPPORT VECTOR REGRESSION

Support vector regression (SVR) is a type of support vector machine (SVM) that is used for regression tasks. It tries to find a function that best predicts the continuous output value for a given input value.

SVR can use both linear and non-linear kernels. A linear kernel is a simple dot product between two input vectors, while a non-linear kernel is a more complex function that can capture more intricate patterns in the data. The choice of kernel depends on the data's characteristics and the task's complexity.



EVALUATING REGRESSION MODELS

R SQUARED TECHNIQUE

- R-squared is a statistical measure that indicates how much of the variation of a dependent variable is explained by an independent variable in a regression model.
- In investing, R-squared is generally interpreted as the percentage of a fund's or security's price movements that can be explained by movements in a benchmark index.
- An R-squared of 100% means that all movements of a security (or other dependent variable) are completely explained by movements in the index (or whatever independent variable you are interested in).

Formula for R-Squared

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

EVALUATING REGRESSION MODELS

ADJUSTED R SQUARED TECHNIQUE

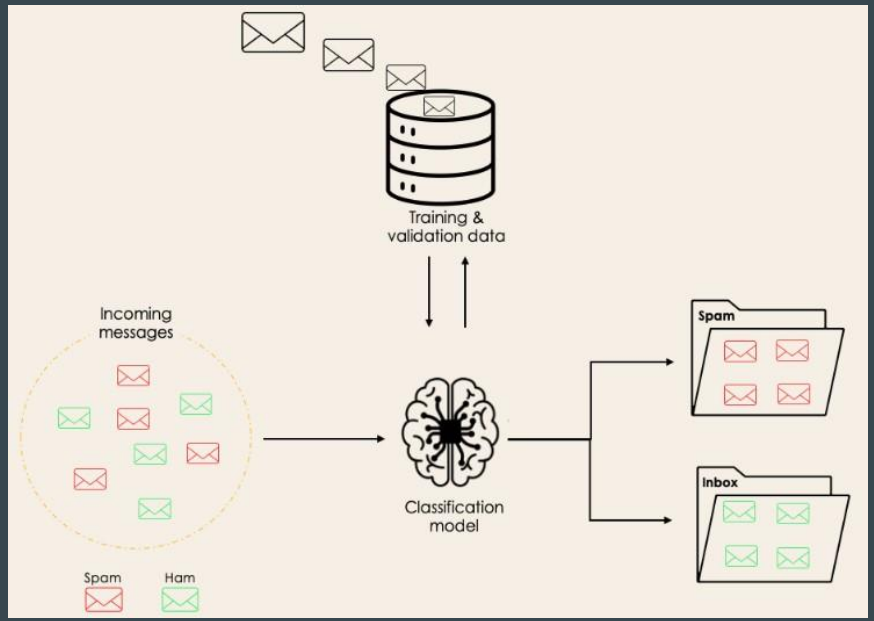
The adjusted R-squared compares the descriptive power of regression models that include diverse numbers of predictors. This is often assessed using measures like R-squared to evaluate the goodness of fit. Every predictor added to a model increases R-squared and never decreases it. Thus, a model with more terms may seem to have a better fit just for the fact that it has more terms, while the adjusted R-squared compensates for the addition of variables; it only increases if the new term enhances the model above what would be obtained by probability and decreases when a predictor enhances the model less than what is predicted by chance.

$$Adj\ R^2 = 1 - (1 - R^2) \times \frac{n-1}{n-k-1}$$

k – number of independent variables
n – sample size

CLASSIFICATION

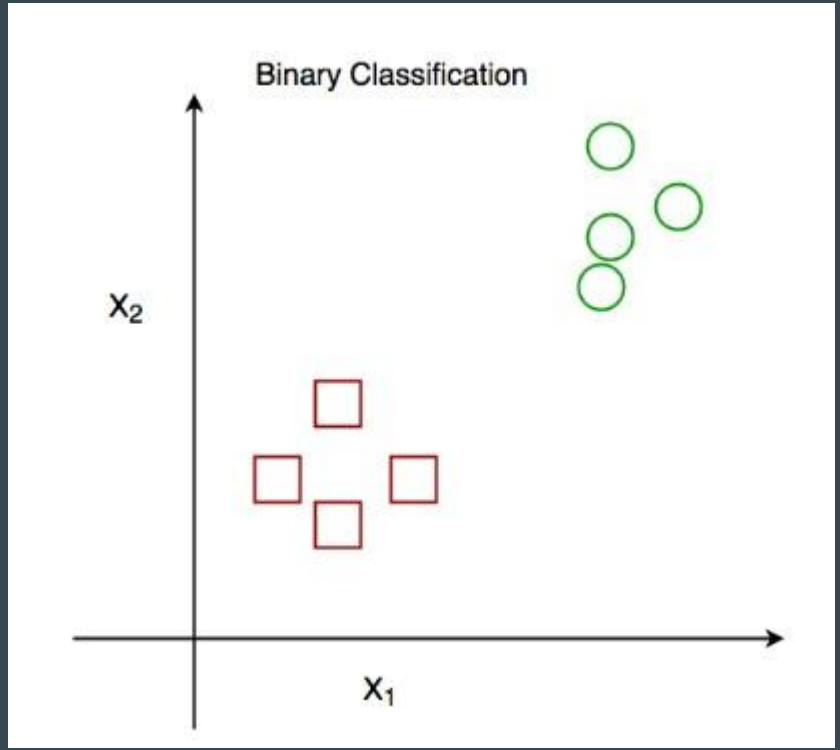
Classification is a supervised machine learning method where the model tries to predict the correct label of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data.



TYPES OF CLASSIFICATION

1. BINARY CLASSIFICATION

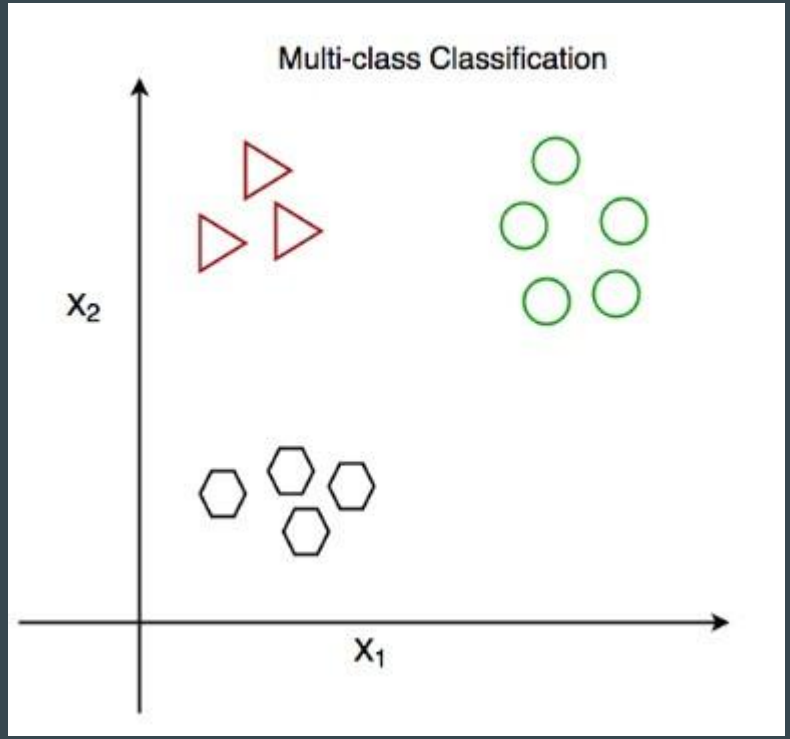
In a binary classification task, the goal is to classify the input data into two mutually exclusive categories. The training data in such a situation is labeled in a binary format: true and false; positive and negative; 0 and 1; spam and not spam, etc. depending on the problem being tackled. For instance, we might want to detect whether a given image is a truck or a boat.



TYPES OF CLASSIFICATION

2. MULTI CLASS CLASSIFICATION

The multi-class classification, on the other hand, has at least two mutually exclusive class labels, where the goal is to predict to which class a given input example belongs to. In the following case, the model correctly classified the image to be a plane.

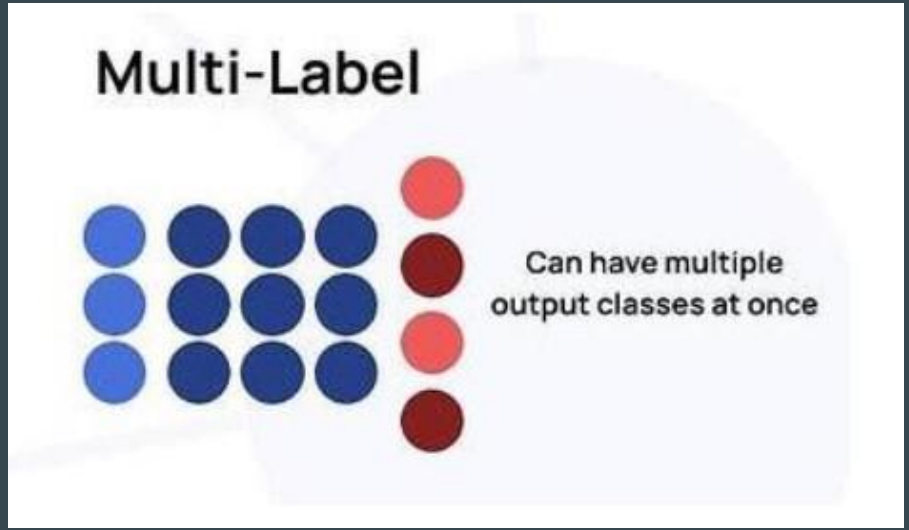


TYPES OF CLASSIFICATION

3. MULTI LABEL CLASSIFICATION

In multi-label classification tasks, we try to predict 0 or more classes for each input example. In this case, there is no mutual exclusion because the input example can have more than one label.

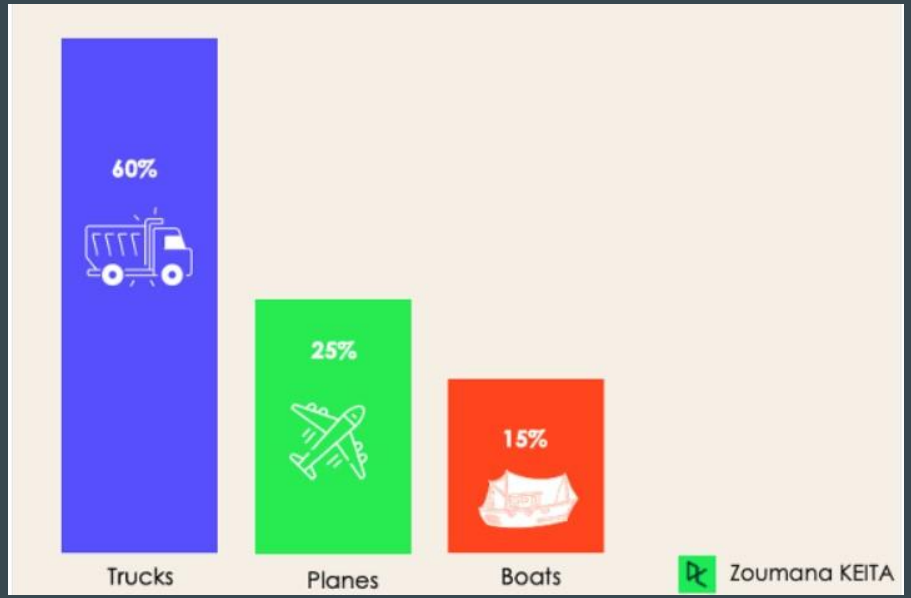
Such a scenario can be observed in different domains, such as auto-tagging in Natural Language Processing, where a given text can contain multiple topics. Similarly to computer vision, an image can contain multiple objects, as illustrated below: the model predicted that the image contains: a plane, a boat, a truck, and a dog.



TYPES OF CLASSIFICATION

4. IMBALANCED CLASSIFICATION

For the imbalanced classification, the number of examples is unevenly distributed in each class, meaning that we can have more of one class than the others in the training data. Let’s consider the following 3-class classification scenario where the training data contains: 60% of trucks, 25% of planes, and 15% of boats.



CLASSIFICATION ALGORITHMS

1. **Decision Tree**
   - **Description**: A model that splits the data into subsets based on feature values, creating a tree-like structure.
   - **Advantage**: Easy to understand and interpret.
2. **Random Forest Classifier**
   - **Description**: An ensemble method using multiple decision trees to improve accuracy and reduce overfitting.
   - **Technique**: Utilizes bagging (bootstrap aggregating).
3. **K-Nearest Neighbors (KNN)**
   - **Description**: Classifies data points based on the majority class of the nearest K neighbors.
   - **Advantage**: Simple and effective for small datasets.
4. **Support Vector Machine (SVM)**
   - **Description**: Finds the hyperplane that best separates data points of different classes.
   - **Advantage**: Effective in high-dimensional spaces.

DIFFERENCES

Regression Algorithm	Classification Algorithm
In Regression, the output variable must be of continuous nature or real value.	In Classification, the output variable must be a discrete value.
The task of the regression algorithm is to map the input value (x) with the continuous output variable(y).	The task of the classification algorithm is to map the input value(x) with the discrete output variable(y).
Regression Algorithms are used with continuous data.	Classification Algorithms are used with discrete data.
In Regression, we try to find the best fit line, which can predict the output more accurately.	In Classification, we try to find the decision boundary, which can divide the dataset into different classes.
Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.
The regression Algorithm can be further divided into Linear and Non-linear Regression.	The Classification algorithms can be divided into Binary Classifier and Multi-class Classifier.

CONCLUSION

Regression and classification are fundamental techniques in supervised machine learning. **Regression** predicts continuous outcomes, while **classification** assigns discrete labels. Both are essential for analyzing and making predictions based on data, with various algorithms available to optimize accuracy and performance.

REFERENCES

- https://www.datacamp.com/blog/classification-machine-learning
- https://www.geeksforgeeks.org/ml-classification-vs-regression/
- https://www.javatpoint.com/regression-vs-classification-in-machine-learning
- https://www.investopedia.com/terms/r/r-squared.asp
- https://www.analyticsvidhya.com/blog/2021/07/demystifying-the-difference-between-multi-class-and-multi-label-classification-problem-statements-in-deep-learning/