**DATA, INFERENCE**

**&**

**APPLIED MACHINE LEARNING**

**(COURSE 18-785)**

**ASSIGNMENT 6**

**Mark Iraguha**

**(miraguha)**

Table of Contents

[Libraries Used: 4](#_Toc181656736)

[Introduction: 4](#_Toc181656737)

[Question 1 Report: 5](#_Toc181656738)

[1.1 Four or more steps to implementing a rule based approach to decision making 5](#_Toc181656739)

[1.2 Overfitting and its downsides in statistical learning 7](#_Toc181656740)

[1.3 Used approaches to avoid overfitting: 9](#_Toc181656741)

[1.4 Examples of metrics used to evaluate model performance 10](#_Toc181656742)

[1.5 Uses of benchmarks with examples 12](#_Toc181656743)

[Question 2 Report: 14](#_Toc181656744)

[2.1Machine Learning, its evolution and reasons for its popularity 14](#_Toc181656745)

[2.2 Examples of machine learning techniques that can be viewed as either supervised or unsupervised approaches: 15](#_Toc181656746)

[2.3 Difference between classification and regression: 16](#_Toc181656747)

[2.4 Difference between supervised and unsupervised learning: 18](#_Toc181656748)

[2.5 Examples of successful applications of machine learning, explaining the technique appropriate and type of learning involved: 20](#_Toc181656749)

[Question 3 Report: 22](#_Toc181656750)

[3.1 Relationship between the variables. 22](#_Toc181656751)

[3.2 Collinearity 23](#_Toc181656752)

[3.3 Multivariate linear model and Variable Significance 24](#_Toc181656753)

[3.4 Forward vs Backward Selection 25](#_Toc181656754)

[3.5 Stepwise Approach 25](#_Toc181656755)

[3.6 Implementation Overview 26](#_Toc181656756)

[3.7 Findings 27](#_Toc181656757)

[3.8 Discussion 27](#_Toc181656758)

[3.9 Conclusions 28](#_Toc181656759)

[Question 4 Report: 29](#_Toc181656760)

[4.1 Difference between logistic regression and linear regression 29](#_Toc181656761)

[4.2 Probability of survival for a passenger on the titanic 29](#_Toc181656762)

[4.3 Survival Probabilities broken down by Pclass, Gender, Survival Probability (%) 30](#_Toc181656763)

[4.4 Logistic Regression and Statistical Significance 31](#_Toc181656764)

[4.5 Model Performance 33](#_Toc181656765)

[4.6 Analysis and Insights 36](#_Toc181656766)

[4.7 Conclusion 36](#_Toc181656767)

[4.8 Overall Insights 36](#_Toc181656768)

[4.9 Overall Conclusion 37](#_Toc181656769)

[References: 38](#_Toc181656770)

# Libraries Used:

Matplotlib – a python plotting library used to create animated, interactive and static visualizations.[1]

Pandas – another Python library used that provides data structures and functions used to carry out data analysis.[2]

Numpy – a simple yet powerful data structure provided in python.[3]

Tabulate – a python library that tabulates data to an output[4].

Statsmodel – a python library that provides a wide range of statistical models and tools for analyzing data[5].

Scikit-Learn - a free machine learning library for Python that supports both supervised and unsupervised machine learning, providing diverse algorithms for classification, regression, clustering, and dimensionality reduction[6].

# Introduction:

This report details the completion of Assignment 6. Assignment 6 requests answers to 4 critical thinking and data analytical questions.

Objectives include:

* Understanding and dealing with nonlinearity.
* Fitting classification models.
* Choosing optimal model parameters.
* Performing cross-validation.
* Evaluating and comparing model performance.

# Question 1 Report:

## 1.1 Nonlinearity explained and its necessity in considering nonlinear relationships between variables

Nonlinearity is the relationship between variables that’s not a straight line. This could involve complex dependencies and interactions that can’t be handled by a linear equation.

Reasons for considering nonlinear relationships between variables:

## Nonlinear models can capture complex patterns, dependencies and interactions that linear models might miss.

* Nonlinear relationships are normally used in many real-word systems and problems.
* Knowledge of nonlinear relationships between models is key in constructing more accurate and realistic predictive models.

## 1.2 Nonlinear Model Equation and Example:

Mathematical equation for a nonlinear model:

* + The general form of a nonlinear model can be represented as:
  + Here, (y) is the dependent variable, (x₁, x₂, ..., x­n) are the independent variables, and (f) is a nonlinear function.

Example of a nonlinear model application:

* + Logistic Regression: Used in classification problems where the relationship between the input variables and the probability of a certain class is nonlinear.  
      
    where (z=b0 + b1x1 + ... + bnxn)

## 1.3 Parsimonious Nonlinear Models:

Parsimony in Linear vs. Nonlinear Models:

* + A model is considered parsimonious if it provides a good fit to the data with a relatively small number of parameters.
  + Linear models are often considered more parsimonious due to their simplicity, but nonlinear models can also be parsimonious.

Mathematical comparison:

* + Linear Model:
  + Nonlinear Model:
  + The nonlinear model may have fewer parameters than the linear model if the function ( f ) is carefully chosen to capture the underlying relationships efficiently.

## 1.4 Surrogate Data for Nonlinearity Testing:

Characteristics preserved in surrogates:

* + Surrogate data generation techniques aim to preserve certain characteristics of the original data while removing the nonlinear relationships.
  + Typically, the marginal distributions of the variables and the linear correlations between them are preserved.

Surrogate techniques:

* + Fourier Transform Surrogates: Involves transforming the data into the frequency domain, randomizing the phases, and then transforming back to the time domain.
  + Amplitude-Adjusted Fourier Transform (AAFT) Surrogates: Similar to Fourier Transform, but adjusts the amplitudes to match the original data's distribution.

## 1.5 Information Theory Concepts and Applications:

Definitions:

* + Information: A measure of the amount of uncertainty reduced by observing an event.
  + Entropy: A measure of uncertainty or randomness in a system.
  + Mutual Information: A measure of the mutual dependence between two variables.

Mathematical formulas:

* + Entropy (H):
  + Mutual Information (I):

Entropy for measuring regularity:

* + Entropy can be used to construct features that quantify the complexity or regularity of a time series. For example, approximate entropy (ApEn) measures the likelihood that similar patterns of observations will remain similar in the next time step.
  + Application: Detecting heart rate variability and analyzing physiological time series.

Mutual Information for feature selection:

* + Mutual information can be used for feature selection by identifying the most informative features for predicting a target variable.
  + Advantage over correlation: Mutual information captures nonlinear relationships and provides a more comprehensive measure of dependence compared to linear correlation.

# Question 2 Report:

## 2.1 Decision trees

Components of a Decision Tree:

* + Nodes: A decision tree consists of nodes, which can be either decision nodes or leaf nodes. Decision nodes represent attributes or features used for splitting the data, while leaf nodes represent the outcomes or decisions.
  + Branches: These are the connections between nodes, representing the possible values or conditions of the attributes. Each branch leads to a subsequent node based on the attribute value.

Pruning:

* + Pruning is the process of removing unnecessary nodes and branches from a decision tree to simplify the model and reduce overfitting. It helps in improving the generalization ability of the tree.

Necessity of Pruning:

* + Overly complex trees with many branches may fit the training data perfectly but struggle to generalize well on unseen data. Pruning helps in reducing the tree's complexity and improving its performance on new data.

Advantages of Decision Trees:

* + Interpretability: Decision trees are easy to understand and interpret, making them attractive for practical applications where explainability is crucial.
  + Non-parametric: They make no assumptions about the underlying data distribution, making them suitable for various types of data.
  + Handling Nonlinear Relationships: Decision trees can capture complex nonlinear relationships between variables, making them versatile for classification tasks.
  + Feature Importance: They provide insights into the relative importance of different features, aiding in feature selection and understanding the data.

## 2.2 Steps to improve the existing rule-based classifier:

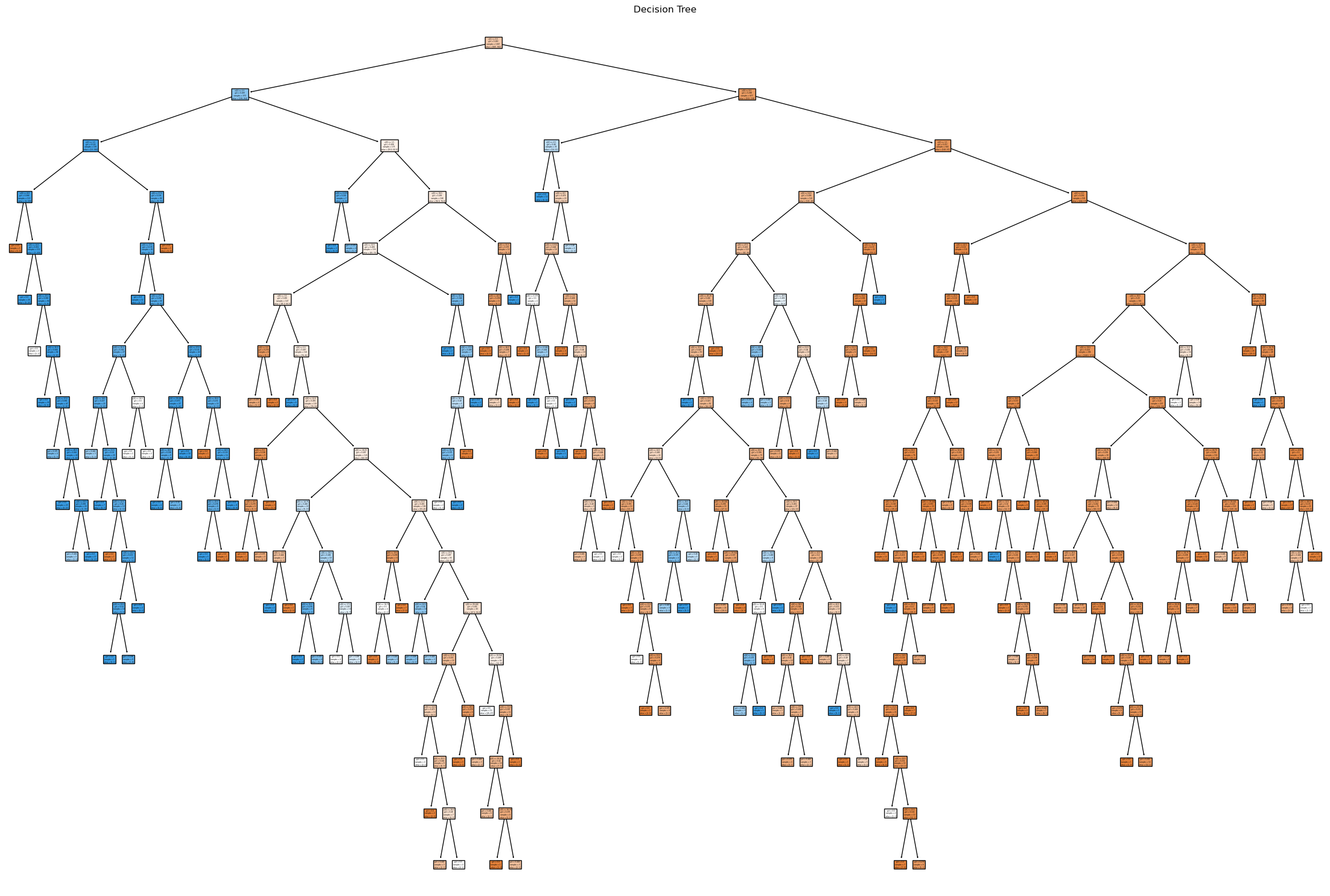
Creating a data-driven classifier:

* 1. Data Collection and Preprocessing: Ensure availability of a large and representative dataset. Clean and preprocess the data to handle missing values, outliers, and encode categorical variables.
  2. Feature Engineering: Analyze the existing rules and domain knowledge to identify relevant features. Create new features or transform existing ones to capture meaningful patterns in the data.
  3. Model Selection: Choose an appropriate data-driven classification algorithm, such as Decision Trees, Random Forests, or Gradient Boosting.
  4. Model Training: Split the data into training and validation sets. Train the selected model on the training set and optimize hyperparameters using techniques like cross-validation.
  5. Model Evaluation: Evaluate the performance of the new model on the validation set using appropriate metrics (e.g., accuracy, precision, recall). Compare it with the existing rule-based classifier.

Testing the new model's validity:

* 1. Hold-out Testing: Split the data into training, validation, and testing sets. Train and optimize the model on the training set, validate on the validation set, and finally test on the unseen testing set.
  2. Cross-Validation: Use techniques like k-fold cross-validation to assess the model's performance across different subsets of the data, providing a more robust evaluation.
  3. Statistical Significance Testing: Perform hypothesis testing (e.g., t-test, chi-square test) to compare the performance of the new model with the existing one, ensuring the improvements are statistically significant.

## 2.3 Decision Tree for Titanic dataset:



## 2.4 Tree Performance Evaluation (before and after pruning):

## 2.5 Performance comparison:

Advantages and Disadvantages:

* Decision Tree:
  + Pros: Interpretability handles nonlinear relationships, feature importance.
  + Cons: May overfit, sensitive to small changes in data, can be complex to interpret for large trees.
* Logistic Regression:
  + Pros: Simple, well-understood, provides probability estimates.
  + Cons: Assumes linear relationships, may not capture complex patterns.

Best for competing in the Kaggle competition is Logistic Regression

# Question 3 Report:

## 3.1 Parsimonious Models using Local Modeling

Concept:

* + Local modeling is an approach that focuses on building models for small regions or neighborhoods of the state space, rather than a single global model for the entire dataset.
  + The idea is to capture local patterns and relationships that might be obscured in a global model, leading to more parsimonious and accurate representations.

Step-by-step procedure:

* 1. Divide the State Space: Partition the input space into smaller regions or neighborhoods based on the input variables. This can be done using techniques like k-means clustering or grid-based approaches.
  2. Local Model Construction: For each neighborhood, construct a separate model using the data points within that region. This can be a simple linear model or a more complex nonlinear model, depending on the problem.
  3. Model Complexity: Keep the models parsimonious by limiting the number of parameters or features used in each local model. This helps to avoid overfitting and improves interpretability.
  4. Prediction: For a new data point, determine its neighborhood based on its input values and use the corresponding local model to make predictions.
  5. Ensemble Prediction (optional): Combine the predictions from multiple local models to make a final prediction, using techniques like weighted averaging or voting.

## 3.2 Collinearity

Collinearity occurs when independent variables are linearly dependent. In other words, one variable can be expressed as a linear combination of others[18].

Collinearity affects predictor variables in a few ways such as:

* Due to inflated standard errors, the p-values associated with the coefficients may become larger, leading to higher likelihood of failing to reject the null hypothesis[19].
* The coefficient estimates become highly sensitive to small changes in the model, such as adding or removing a predictor variable.
* When collinearity exists, interpreting the coefficients becomes problematic because it is difficult to ascertain how much of the change in the dependent variable is attributable to each independent variable independently.
* Collinearity increases the variance of the estimated coefficients. When predictor variables are highly correlated, it becomes difficult to isolate the individual effect of each variable on the dependent variable.

## 3.4 Sensitivity to Feature Types

Different distance metrics in KNN are sensitive to the types of features used. For example:

* + Euclidean distance is suitable for continuous numerical features but may not work well with categorical variables.
  + Manhattan distance is less sensitive to outliers and can be more appropriate for high-dimensional data.
  + Cosine distance is commonly used for text data or when features are normalized.

## 3.5 KNN vs Logistic Regression

Advantages and Disadvantages:

* KNN:
  + Pros: Simple, non-parametric, handles nonlinear relationships, no assumptions about data distribution.
  + Cons: Sensitive to the choice of distance metric, can be computationally expensive for large datasets.
* Logistic Regression:
  + Pros: Interpretable, provides probability estimates, handles linear relationships well.
  + Cons: Assumes linearity, may not capture complex patterns, requires feature scaling.

Kaggle Competition Choice:

* Both models have their strengths and weaknesses. KNN's ability to handle nonlinear relationships might be advantageous for the Titanic dataset.
* However, the choice depends on the specific dataset and problem. Logistic regression's simplicity and interpretability could also be beneficial.
* Consider ensemble methods or feature engineering to further improve performance and increase the chances of success in the Kaggle competition

# Question 4 Report:

## 4.1 Difference between logistic regression and linear regression

**Purpose:**

Logistic regression is used for binary classification problems where the dependent variable is categorical, typically with two possible outcomes (e.g. yes/no). The output is a probability that maps to a binary outcome, often interpreted as the likelihood of an event occurring[20].

Linear Regression is used to predict a continuous dependent variable. For example, it can be used to forecast stock prices based on various independent variables like ADJ Close, Open Price. The output is a numeric value that can take any within a range[20].

Output Type:

Linear Regression: This produces continuous output values such as predicting temperature.

Logistic Regression: This produces probabilities that can be converted into binary outcomes such as if the probability is greater than 0.5, classify as 1, otherwise, classify as 0.

Loss Function:

Linear Regression minimizes the sum of squared differences between observed and predicted values (OLS).

Logistic Regression minimizes logistic loss, which assess, how well the predicted probabilities match the actual outcomes.

## 4.2 Probability of survival for a passenger on the titanic

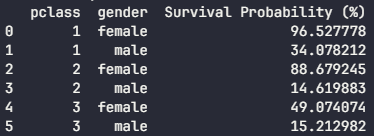
**Steps followed:**

1. Sum function and Probability formula:
   1. Utilized the sum() function and shape[index] attribute to get the total number of survivors and passengers.
   2. Dividing the two variables to get the probability.
2. Result probability is 0.381970.



## 4.3 Survival Probabilities broken down by Pclass, Gender, Survival Probability (%)

This table illustrates how different groups may have fared based on the dataset:

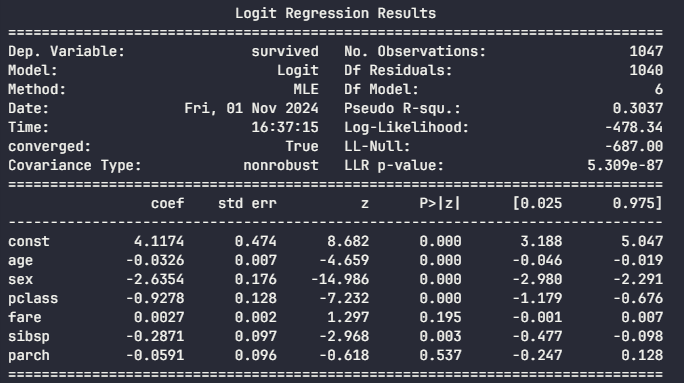


Steps followed:

1. **Creating the Sample Dataset**:
   * A DataFrame is created with columns for passenger class (Pclass), gender (Sex), age (Age), and whether they survived (Survived).
2. **Calculating Survival Probabilities**:
   * The dataset is grouped by Pclass and Sex.
   * The total number of passengers and the number of survivors are aggregated using .agg().
   * The survival probability is calculated as the ratio of survivors to total passengers multiplied by 100 to get a percentage.
3. **Formatting the Resulting Table**:
   * The resulting DataFrame is filtered to show only relevant columns.
   * The Sex column is renamed to Gender for clarity.

## 4.4 Logistic Regression and Statistical Significance

Implemented logistic regression model using both scikit-learn and statsmodels by fitting models using passenger class, gender (as "male" or "female"), and age as predictors.

****

Parameter Estimates and Statistical Significance

In the logistic regression model summary from statsmodels, indication of how to interpret the coefficients for each predictor variable (e.g., Pclass, Sex, Age) is shown and their statistical significance through p-values is checked.

Using an alpha level of 0.05 to assess each predictor's significance:

* Sex:
  + pvalue: 0.000
  + Since 0.000 < 0.05, reject the null hypothesis; thus, this predictor is statistically significant.
  + Coefficient indicating lower survival odds for males compared to females.
* Fare:
  + pvalue: 0.0027
  + Since 0.195 > 0.05, fail to reject the null hypothesis; thus, it is not statistically significant.
* Age:
  + pvalue: 0.000
  + Since 0.000 < 0.05, reject the null hypothesis; thus, this predictor is statistically significant.
  + Coefficient indicating that older passengers had lower odds of survival.
* Pclass:
  + pvalue: 0.000
  + Since 0.000 < 0.05, reject the null hypothesis; thus, this predictor is statistically significant.
  + Coefficient indicates higher survival odds for higher classes.
* Sibsp:
  + pvalue: -0.003
  + Since -0.003 < 0.05, reject the null hypothesis; thus, this predictor is statistically significant.
* Parch:
  + pvalue: -0.537
  + Since -0.537 > 0.05, fail to reject the null hypothesis; thus, this predictor is not statistically significant.

Conclusion:

Based on an alpha level of **0.05**, the conclusion is that:

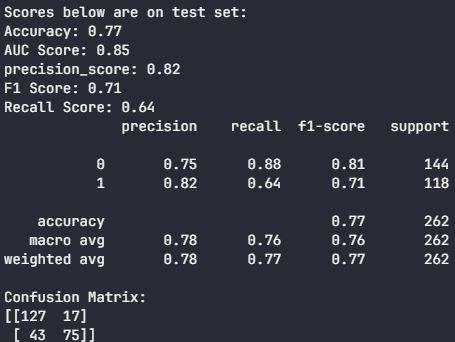
* The variables **Pclass**, **Sex**, **Age, Sibsp** and the intercept are statistically significant predictors of survival.
* The variable **Fare, Parch** are not statistically significant in predicting survival in this model.

## 4.5 Model Performance

1. Classification Accuracy:

To measure the performance of a logistic regression model, utilization of classification accuracy is necessary, which is calculated as the ratio of correctly predicted instances to the total number of instances.

Since the logistic regression model has been created from the previous stages and predictions have been made, a classification accuracy gives us the value of 0.77 or 77%.



This means that the model correctly predicted survival for **77%** of the instances in the test set.

2. Confusion Matrix:

To further evaluate the performance of the logistic regression model, a confusion matrix is utilized.

A confusion matrix provides insight into how many instances were correctly or incorrectly classified.

A structure for the Confusion Matrix would look like this:

|  |  |  |
| --- | --- | --- |
|  | Predicted Survived (1) | Predicted Not Survived (0) |
| Actual Survived (1) | TP | FN |
| Actual Not Survived (0) | FP | TN |

Where:

* **TN** (True Negatives): Correctly predicted not survived.
* **FP** (False Positives): Incorrectly predicted survived.
* **FN** (False Negatives): Incorrectly predicted not survived.
* **TP** (True Positives): Correctly predicted survived.

Additional metrics can be derived such as:

* **Precision**: TP / (TP + FP)
* **Recall (Sensitivity)**: TP / (TP + FN)
* **F1 Score**: Harmonic mean of precision and recall

The resultant confusion matrix values:

TP = 127

FN = 17

FP = 43

TN = 75



Steps followed:

1. **Prepare Data**: Clean and organize your dataset to make it ready for modeling.
2. **Train Model**: Split your dataset into training and testing sets, then train your logistic regression model on the training set.
3. **Make Predictions**: Use the trained model to predict outcomes on the test set.
4. **Calculate Accuracy**: Find out how many predictions were correct compared to the total number of cases.
5. **Create Confusion Matrix**: Build a confusion matrix using actual results versus predicted results from your test set.
6. **Calculate Additional Metrics**: Use the confusion matrix values to find precision, recall, and F1 score.

## 4.6 Analysis and Insights

The confusion matrix shows that while the model has good accuracy, there are areas for improvement:

* The number of false positives (43) indicates that some people were incorrectly predicted as survivors when they were not. This suggests that an improvement to feature selection or adjust model settings might be required.
* The false negatives (17) show that some survivors were missed by the model. This is important in situations where identifying survivors is crucial.

## 4.7 Conclusion

The logistic regression model has a solid classification accuracy of 77%. However, looking at the confusion matrix reveals opportunities for improvement in precision and recall. By addressing false positives and false negatives, future versions of this model can perform better in real-world situations. Ongoing evaluation and adjustments will be key to making this model more effective.

## 4.8 Overall Insights

The analysis of Titanic survival data illustrates the key differences between logistic and linear regression. Logistic regression is appropriate for binary classification tasks, such as predicting whether a passenger survived, while linear regression is used for predicting continuous outcomes. The overall survival rate for passengers was around 38.2%, showing notable differences based on factors like passenger class (Pclass) and gender, with higher-class passengers and females having better chances of survival. The logistic regression model identified several important predictors, including sex, Pclass, age, and the number of siblings or spouses on board (Sibsp), all of which were statistically significant. In contrast, fare and the number of parents or children aboard (Parch) did not show significant effects.

## 4.9 Overall Conclusion

The model achieved a classification accuracy of 77%, with 127 true positives, 17 false negatives, 43 false positives, and 75 true negatives. While the model performs reasonably well, there are areas for improvement, particularly in reducing false positives and false negatives. By refining feature selection and adjusting the model parameters, it may be possible to enhance its effectiveness in accurately identifying survivors in similar situations.

# References:

[1] “Matplotlib,” *Wikipedia*. Aug. 30, 2024. Accessed: Sep. 01, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Matplotlib&oldid=1243075914

[2] “pandas (software),” *Wikipedia*. Jul. 15, 2024. Accessed: Sep. 01, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Pandas\_(software)&oldid=1234683004

[3] R. Python, “NumPy Tutorial: Your First Steps Into Data Science in Python – Real Python.” Accessed: Sep. 02, 2024. [Online]. Available: https://realpython.com/numpy-tutorial/

[4] *tabulate: Pretty-print tabular data*. Python. Accessed: Sep. 15, 2024. [OS Independent]. Available: https://github.com/astanin/python-tabulate

[5] P. Mania, “What Is Statsmodels in Python? The Ultimate Guide,” Python Mania. Accessed: Sep. 30, 2024. [Online]. Available: https://pythonmania.org/what-is-statsmodels-in-python-the-ultimate-guide/

[6] Nik, “Introduction to Scikit-Learn (sklearn) in Python • datagy,” datagy. Accessed: Oct. 16, 2024. [Online]. Available: http://datagy.io/python-scikit-learn-introduction/

[7] “Mastering Rule-Based Systems: Implementation, Benefits, and Best Practices,” Mastering Rule-Based Systems: Implementation, Benefits, and Best Practices. Accessed: Nov. 03, 2024. [Online]. Available: https://www.xaqt.com/blog/mastering-rule-based-systems/

[8] “The Role Of Domain Knowledge In Decision Making,” FasterCapital. Accessed: Nov. 03, 2024. [Online]. Available: https://fastercapital.com/keyword/the-role-of-domain-knowledge-in-decision-making.html

[9] “What is Overfitting? - Overfitting in Machine Learning Explained - AWS,” Amazon Web Services, Inc. Accessed: Nov. 03, 2024. [Online]. Available: https://aws.amazon.com/what-is/overfitting/

[10] D. C.-E. Lin, “8 Simple Techniques to Prevent Overfitting,” Medium. Accessed: Nov. 03, 2024. [Online]. Available: https://towardsdatascience.com/8-simple-techniques-to-prevent-overfitting-4d443da2ef7d

[11] “Overfitting,” *Wikipedia*. Oct. 02, 2024. Accessed: Nov. 03, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Overfitting&oldid=1248918418

[12] “Mean squared error,” *Wikipedia*. Jun. 11, 2024. Accessed: Nov. 03, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Mean\_squared\_error&oldid=1228454019

[13] “Evaluating Deep Learning Models: The Confusion Matrix, Accuracy, Precision, and Recall | DigitalOcean.” Accessed: Nov. 03, 2024. [Online]. Available: https://www.digitalocean.com/community/tutorials/deep-learning-metrics-precision-recall-accuracy

[14] “Accuracy.” Accessed: Nov. 03, 2024. [Online]. Available: https://graphite-note.com/the-importance-of-accuracy-in-machine-learning/

[15] zuzanna, “What is a benchmark and why do you need it?,” MIM Solutions - We make artificial intelligence work for you. Accessed: Nov. 03, 2024. [Online]. Available: https://www.mim.ai/what-is-a-benchmark-and-why-do-you-need-it/

[16] “History and Evolution of Machine Learning: A Timeline,” WhatIs. Accessed: Nov. 03, 2024. [Online]. Available: https://www.techtarget.com/whatis/feature/History-and-evolution-of-machine-learning-A-timeline

[17] “Decision Tree vs. Neural Network - What’s the Difference?,” This vs. That. Accessed: Nov. 03, 2024. [Online]. Available: https://thisvsthat.io/decision-tree-vs-neural-network

[18] “Collinearity Definition & Examples,” Quickonomics. Accessed: Oct. 29, 2024. [Online]. Available: https://quickonomics.com/terms/collinearity/

[19] “A Beginner’s Guide to Collinearity: What it is and How it affects our regression model.” Accessed: Oct. 29, 2024. [Online]. Available: https://www.stratascratch.com/blog/a-beginner-s-guide-to-collinearity-what-it-is-and-how-it-affects-our-regression-model/

[20] “Linear vs Logistic Regression: How to Choose the Right Regression Model for Your Data,” freeCodeCamp.org. Accessed: Nov. 03, 2024. [Online]. Available: https://www.freecodecamp.org/news/linear-regression-vs-logistic-regression/