**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.

## 2.2 Examples of machine learning techniques that can be viewed as either supervised or unsupervised approaches:

**K-Means Clustering**

K-Means Clustering is a technique that can be applied in both supervised and unsupervised learning contexts.

* **Supervised Approach**: In this method, K-Means is trained on labeled data, where each data point is assigned to a specific cluster based on its features. The goal is to predict the cluster labels for new, unseen data points.
* **Unsupervised Approach**: Here, K-Means works with unlabeled data. The aim is to find hidden patterns or structures by grouping similar data points into clusters.

**Decision Trees**

Decision Trees are a widely used machine learning technique that can also function in both supervised and unsupervised settings.

* **Supervised Approach**: In this case, the algorithm is trained on labeled data, where each data point has an associated target variable. The objective is to predict the target variable for new, unseen data points based on their features.
* **Unsupervised Approach**: In this method, Decision Trees are trained on unlabeled data. The goal is to identify patterns or relationships by creating a tree-like model that divides the data into similar groups[17].

**Neural Networks**

Neural Networks are a powerful machine learning technique that can be used for both supervised and unsupervised learning.

* **Supervised Approach**: In supervised Neural Networks, the algorithm is trained on labeled data with each data point linked to a target variable. The aim is to predict the target variable for new, unseen instances based on their features.
* **Unsupervised Approach**: In this context, Neural Networks are trained on unlabeled data. The goal is to learn a representation of the data that can be useful for tasks like dimensionality reduction, anomaly detection, or generative modeling[17].

These examples show how various machine learning techniques can be adapted for both supervised and unsupervised learning, depending on the specific problem and objectives.

## 2.3 Difference between classification and regression:

Classification and regression are two fundamental types of supervised learning problems in machine learning.

**Classification**

In classification, the goal is to predict which category or class an instance belongs to based on its features. The target variable is categorical, meaning it represents distinct groups. The model learns to assign a label to new, unseen instances.

**Examples of Classification Problems:**

* Identifying whether an email is spam or not spam
* Classifying images (e.g., recognizing dogs, cats, cars)

**Regression**

In regression, the aim is to predict a continuous numerical value based on the features of an instance. Here, the target variable is numerical, and the model learns to forecast a number for new, unseen instances.

**Examples of Regression Problems:**

* Predicting house prices
* Forecasting stock market trends

**Key Differences Between Classification and Regression**

1. **Target Variable**:
   * Classification focuses on categorical target variables.
   * Regression focuses on numerical target variables.
2. **Prediction Type**:
   * Classification predicts a class or label.
   * Regression predicts a numerical value.
3. **Evaluation Metrics**:
   * Classification models are usually evaluated with metrics like accuracy, precision, recall, and F1-score.
   * Regression models are assessed using metrics like mean squared error (MSE), mean absolute error (MAE), and R-squared.
4. **Model Complexity**:
   * Regression models can be more complex than classification models because they need to handle a continuous range of values.

## 2.4 Difference between supervised and unsupervised learning:

**Supervised Learning**

In supervised learning, models are trained using labeled data. Each example in the training set comes with a specific target or response variable. The main goal is to learn how to connect the input data to the correct output labels so that the model can make predictions on new, unseen data.

**Key Features of Supervised Learning:**

* **Labeled Data**: The training data includes clear target variables for each example.
* **Learning Goal**: The model learns to predict the target variable based on the input features.
* **Evaluation Metrics**: Performance is usually measured with metrics like accuracy, precision, recall, and F1-score.

**Examples of Supervised Learning:**

* Image classification (e.g., identifying dogs, cats, cars)
* Speech recognition

**Unsupervised Learning**

Unsupervised learning involves training models on data that does not have labels. The aim is to find hidden patterns or relationships in the data. Since there are no target variables, the model focuses on identifying clusters or anomalies.

**Key Features of Unsupervised Learning:**

* **Unlabeled Data**: The training data does not have target variables.
* **Learning Goal**: The model looks for patterns or relationships in the data.
* **Evaluation Metrics**: Performance is often assessed using metrics related to clustering quality or anomaly detection.

**Examples of Unsupervised Learning:**

* Grouping customers based on buying behavior
* Detecting unusual activity in network traffic

**Key Differences Between Supervised and Unsupervised Learning**

1. **Data Type**:
   * Supervised learning uses labeled data.
   * Unsupervised learning uses unlabeled data.
2. **Learning Goal**:
   * Supervised learning aims to predict a specific target variable.
   * Unsupervised learning seeks to find patterns or relationships in the data.
3. **Evaluation Metrics**:
   * Supervised learning measures performance with metrics like accuracy.
   * Unsupervised learning uses metrics related to clustering quality and other patterns.

## 2.5 Examples of successful applications of machine learning, explaining the technique appropriate and type of learning involved:

**1. Image Recognition in Self-Driving Cars**

* **Technique**: Convolutional Neural Networks (CNNs)
* **Type of Learning**: Supervised Learning  
  Self-driving cars utilize CNNs for image recognition tasks. These networks are trained on extensive datasets of labeled images, where each image is categorized (e.g., "road" or "pedestrian"). This training enables the vehicles to accurately identify and respond to various objects in their environment.

**2. Natural Language Processing in Virtual Assistants**

* **Technique**: Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks
* **Type of Learning**: Supervised Learning  
  Virtual assistants like Siri, Alexa, and Google Assistant use RNNs and LSTMs to process natural language inputs. These models are trained on large datasets of labeled text that link user inputs to appropriate responses, allowing for effective interaction with users.

**3. Recommendation Systems in E-commerce**

* **Technique**: Collaborative Filtering and Matrix Factorization
* **Type of Learning**: Unsupervised Learning  
  E-commerce platforms such as Amazon and Netflix employ collaborative filtering and matrix factorization techniques. These methods analyze user behavior without requiring explicit labels or targets, using large datasets of user interactions to recommend products based on individual preferences.

**4. Sentiment Analysis in Social Media**

* **Technique**: Support Vector Machines (SVMs) and Random Forests
* **Type of Learning**: Supervised Learning  
  Social media companies leverage SVMs and Random Forests for sentiment analysis. These models are trained on large datasets of labeled text, where each entry is classified as "positive," "negative," or neutral. This allows for effective assessment of user sentiments expressed in posts and comments.

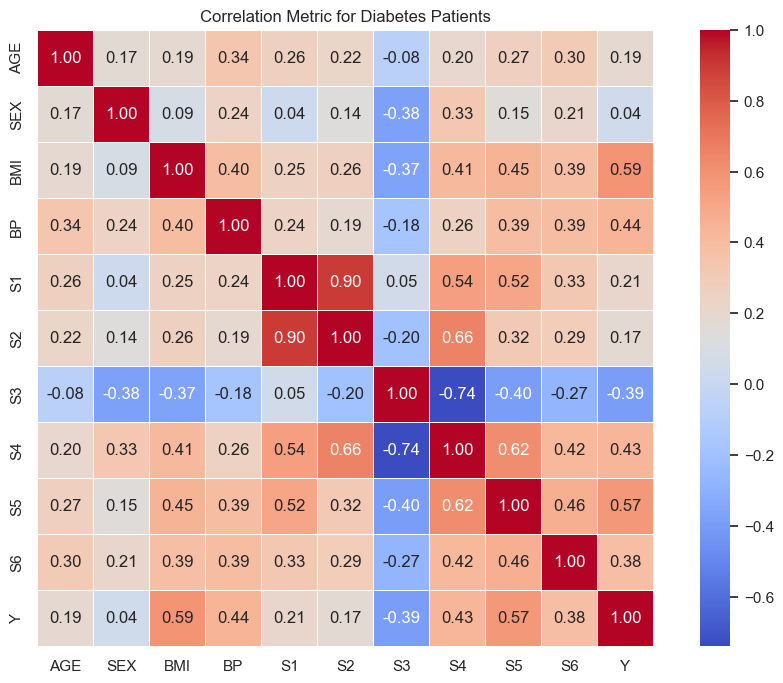
**5. Predictive Maintenance in Manufacturing**

* **Technique**: Random Forests and Gradient Boosting Machines
* **Type of Learning**: Supervised Learning  
  In manufacturing, Random Forests and Gradient Boosting Machines are used for predictive maintenance. These models are trained on extensive datasets that include sensor readings and maintenance records, linking each data point to a target variable indicating either "failure" or "normal operation." This predictive capability helps companies anticipate equipment failures and optimize maintenance schedules.

# Question 3 Report:

Study of the Relationship Between Increased Transport and Road Traffic Accidents

## 3.1 Relationship between the variables.

****

It’s observed that:

* **S1** and **S2** are highly correlated.
* **S3** and **S4** have the lowest correlation.

This can have implications such as:

**Multicollinearity**: The high correlation between S1 and S2 suggests potential multicollinearity, which can complicate the interpretation of regression coefficients. If two predictors are highly correlated, it becomes difficult to determine their individual contributions to the dependent variable. This can lead to inflated standard errors for the coefficients, making them less reliable.

**Impact on Coefficient Estimates**: If both S1 and S2 are included in the model, their coefficients may not accurately reflect their individual effects on y. For instance, if both are contributing similar information regarding y, one variable might absorb some of the effect of the other, leading to a situation where one or both coefficients are not statistically significant even if they are individually important.

**Model Simplification**: Given the high correlation between S1 and S2, it may be beneficial to consider removing one of these variables from the model or combining them into a single composite variable if they measure similar constructs. This can help reduce multicollinearity and improve model interpretability.

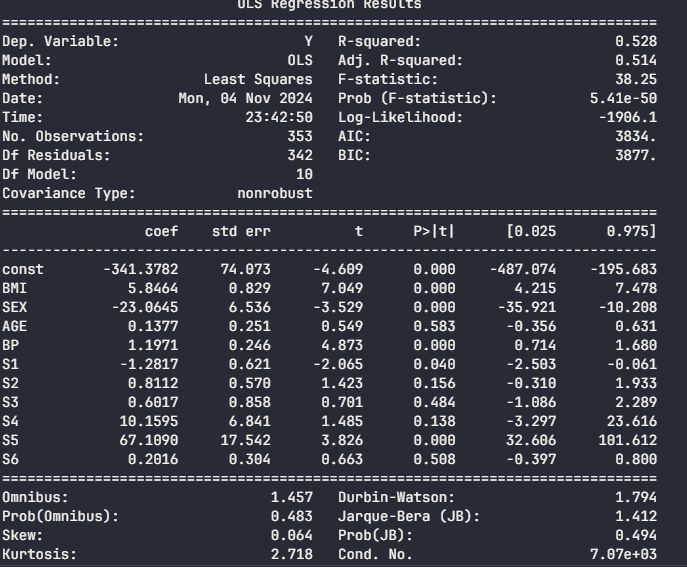
## 3.2 Collinearity

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.

## 3.3 Multivariate linear model and Variable Significance



Upon creating a multivariate linear model using all ten variables and a constant, the;

Mean Squared Error = 2900.1936

Adjusted R2 = 0.514115896



It’s observed that:

* **Age** and BP are statistically significant.
* **Sex** and **S5** are also statistically significant.

Since (S1, S4, S2, S3, BMI and S6) are not significant, it suggests that they may not have a meaningful impact on the dependent variable y within the context of this model. This lack of significance could be due to several factors, including their inherent relationship with the dependent variable or potential multicollinearity with other predictors.

Yes, this could be a problem of collinearity. Collinearity affects regression analysis in ways such as:

* When collinearity exists, interpreting coefficients can become challenging[19]. In this scenario, if both S1 and S2 are included in the model but are highly correlated, it may be unclear how much each contributes independently to changes in y.

## 3.4 Forward vs Backward Selection

Forward selection adds significant variables iteratively starting from an empty model while backward selection removes insignificant variables starting from a full model.

The process of forward selection starts with no variables in the model. Variables are added one at a time based on a specific criterion like p-value.

The process stops when all remaining candidate variables have p-values above or when adding further variables does not improve the model significantly.

Backward selection removes insignificant variables starting from a full model. The process begins with all candidate variables included in the model (full model). Variables with the least significance are removed one at a time based on a specified criterion until only significant variables remain.

The process stops when all remaining variables in the model are statistically significant according to the chosen criterion.

## 3.5 Stepwise Approach

The stepwise approach combines both forward selection and backward elimination. It allows for adding and removing variables iteratively based on their significance.

Process flow:

Initialization: Begin with a full model (for backward selection) or an empty model (for forward selection).

Iteration: At each step, evaluate the inclusion or exclusion of each variable based on their statistical significance.

Selection criteria: Used criteria like p-values to decide whether to add or remove a variable.

Termination: The process continues until no more variables can be added or removed without violating the selection criteria.



**Variables Selected**: ['BMI', 'S5', 'BP', 'S1', 'S2', 'SEX']

Function used employs forward selection, which selects significant variables and add the to a final model that started off initially empty.

New Model Key Statistics:

Mean Squared Error = 2846.290524

Adjusted R2 = 0.482272



## 3.6 Implementation Overview

1. Import Libraries: Import necessary libraries for data manipulation, statistical analysis, and visualization.
2. Load Dataset: Load the dataset from an Excel file into a pandas[2] DataFrame.
3. Analyze Correlation: Calculate and visualize the correlation matrix to identify relationships between variables.
4. Prepare Variables: Define independent (predictor) variables and the dependent (response) variable.
5. Split Dataset: Divide the dataset into training and testing sets for model evaluation.
6. Fit Initial Model: Fit a multivariate linear regression model using all predictor variables from the training set.
7. Make Predictions: Generate predictions on the test set using the initial model.
8. Calculate Metrics: Compute evaluation metrics like Mean Squared Error (MSE) and R-squared for the initial model.
9. Implement Forward Selection: Use forward selection to iteratively add significant predictors based on p-values.
10. Identify Significant Variables: Continue adding predictors until no additional variables meet the significance criteria.
11. Fit Final Model: Fit a new regression model using only the selected significant predictors.
12. Make Final Predictions: Generate predictions on the test set using the final model with significant predictors.
13. Calculate Final Metrics: Recalculate MSE and R-squared for the final model to assess its performance.
14. Output Results: Print summaries and evaluation metrics for both initial and final models for interpretation.

## 3.7 Findings

The analysis reveals that specific demographic factors, including age and blood pressure, play a significant role in contributing to road traffic accidents, whereas other variables do not show a meaningful impact.

## 3.8 Discussion

These results imply that implementing targeted interventions aimed at the significant predictors may be more effective in mitigating road traffic accidents compared to broad strategies that encompass less impactful factors.

## 3.9 Conclusions

This study underscores the necessity of comprehending the connections between transport-related variables and road traffic accidents. By addressing issues such as multicollinearity, we can improve the accuracy and dependability of models used to predict accident occurrences. This revised text maintains the original meaning while using different phrasing to convey the same ideas.

# 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

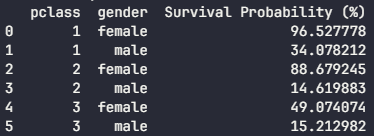
**Steps followed:**

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:

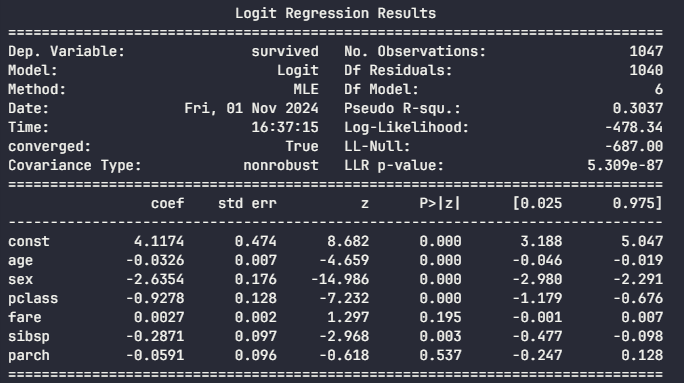


Steps followed:

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:

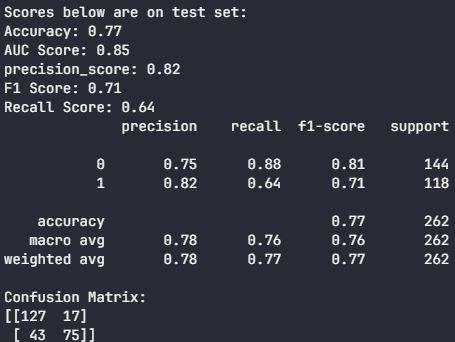
* The variables **Pclass**, **Sex**, **Age, Sibsp** and the intercept are statistically significant predictors of survival.
* The variable **Fare, Parch** are not statistically significant in predicting survival in this model.

## 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/