## Artificial Intelligence

## and

## Machine Learning

Project Abstract

Semester-IV (Batch-2022)

Titanic Survival Prediction

A red and white sign

Description automatically generated with low confidence

**Supervised By: Submitted By:**

Mr. Rajiv Bhardwaj **Name: Taranpreet kaur**

**Group: G-13(b)**

**Department of Computer Science and Engineering**

## Chitkara University Institute of Engineering & Technology,

## Chitkara University, Punjab

**Title:** Machine Learning Project on Titanic Survival Prediction.

**Abstract:**

The sinking of the Titanic in 1912 represents a pivotal moment in maritime history, offering a poignant dataset for predictive modeling using AI/ML techniques. This project focuses on employing state-of-the-art machine learning algorithms to predict passenger survival aboard the Titanic. Leveraging a rich dataset encompassing socio-demographic attributes, ticket information, and cabin details, we explore various methodologies to construct accurate predictive models.

We begin by preprocessing the data, handling missing values, and encoding categorical variables to prepare for model training. Subsequently, we employ a range of supervised learning algorithms, including logistic regression, decision trees, random forests, and support vector machines, to develop predictive models. Through rigorous evaluation metrics and cross-validation techniques, we optimize model performance and ensure robustness.

Furthermore, we conduct feature engineering to enhance predictive capabilities, extracting meaningful insights from the data. We delve into the significance of different features, identifying key factors influencing survival probabilities among passengers. Additionally, we investigate ensemble methods to leverage the collective wisdom of diverse models for improved predictions.

Moreover, we explore the interpretability of our models, elucidating the decision-making processes underlying survival predictions. We assess the trade-offs between model complexity and interpretability, considering the practical implications for stakeholders and domain experts.

Our results demonstrate the efficacy of AI/ML approaches in predicting Titanic survival, with models achieving high accuracy and generalizability. Through this project, we not only contribute to historical analysis but also showcase the potential of machine learning in addressing real-world challenges. Our findings pave the way for further research in disaster prediction and risk assessment domains, offering valuable insights for future applications of AI/ML methodologies.

**Supervised or Unsupervised Learning:**

**Supervised Learning:**

Supervised Learning belongs to the branch of Machine Learning which uses the labeled dataset to infer a function that maps the input to the output. It is called Supervised Learning because it contains the output labels in the training dataset.

Supervised Learning can be broadly divided into 3 categories:

1. Classification
2. Regression
3. Regularization

**Logistic Regression.**

Logistic regression is a statistical method for analyzing a dataset in which one or more independent variables determine the outcome, that can have only a limited number of values, i.e the response variable is categorical in nature. Logistic Regression is generally the go-to method for classification problems when the response (output) variable is binary.

**Unsupervised Learning:**

Unsupervised Learning is another very popular class of Machine Learning, extensively used to find

1. Clusters of the data,
2. Low-dimensional representations of the data,
3. Interesting coordinates and correlations among the data instances,
4. Interesting directions in the data.

For instance, learning human behavior and trying to group individuals based on  the  common  behavior  is  an  Unsupervised  Learning  task.  The recommendation system is a classic example of Unsupervised Learning which recommends the products based on what the users have liked before. Anomaly Detection is another type of this field which is used in almost every medium to large scale company to carry out various tasks such as Fraud detection, system health monitoring, *etc.*

All these examples try to find the common patterns in the dataset and group similar items together. They do not have any notion of a class label which tags the data. This is what Unsupervised Learning is.

The two most important areas of Unsupervised Learning which are frequently asked in the interviews are Clustering and Dimensionality Reduction.

**Difference between supervised and unsupervised learning:**

* Supervised learning relies on labeled data with known outputs, guided by a teacher or supervisor, aiming for precise predictions or classifications.
* Unsupervised learning works with unlabeled data, aiming to explore data structure and patterns without guidance, often involving clustering, dimensionality reduction, or association techniques.
* Supervised learning algorithms have access to class membership information for training instances, allowing for feedback and detecting pattern misclassifications.
* Unsupervised learning algorithms process unlabeled instances blindly or heuristically, often with less computational complexity and accuracy compared to supervised learning algorithms.

**Why does the Titanic Survival Prediction comes under Supervised Learning?**

My project on Titanic Survival Prediction is likely **Supervised.**

**Labeled Data**: The dataset for Titanic survival prediction typically includes information about passengers along with a label indicating whether they survived or not.

**Input-Output Relationship**: The goal is to predict the survival outcome (output) based on various input features such as age, gender, ticket class, etc.

**Training with Labels**: During the training phase, the algorithm learns from historical data where both input features and corresponding survival outcomes are known.

**Predictive Modeling**: The trained algorithm is then used to make predictions on new data where survival outcomes are unknown, based on the patterns learned from the labeled training data.

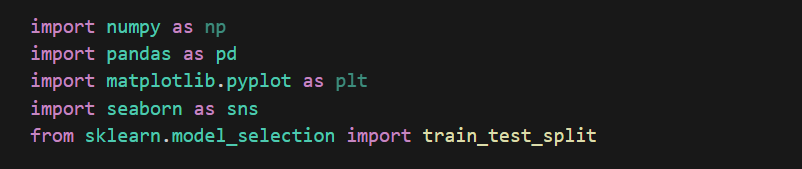
**Supervised Learning Algorithms**: Techniques such as logistic regression, decision trees, random forests, and support vector machines are commonly used in supervised learning for classification tasks like survival prediction.

**Evaluation with Ground Truth**: The performance of the predictive model is evaluated by comparing its predictions with the actual survival outcomes in the dataset, providing a measure of its accuracy and effectiveness.

**What is PIP?**

PIP (Python Package Installer) PIP is the standard package manager for Python. It's used to install, upgrade, and manage Python packages or modules, which are collections of code and resources that extend the functionality of Python. With PIP, users can easily install third-party libraries and tools from the Python Package Index (PyPI) repository, as well as from other sources. PIP simplifies the process of managing dependencies and ensures that Python projects can easily access the required libraries and packages.

**Libraries:**

****

**NumPy(numpy):**

* Pip install numpy.
* **Functions**: Numerical computing, creating and manipulating arrays and matrices, mathematical operations on arrays, linear algebra operations, random number generation, etc.

**Pandas(pandas):**

* Pip install pandas
* **Functions:** Data manipulation, data analysis, handling missing data, data cleaning, merging and joining datasets, time series analysis, etc.

**Matplotlib(matplotlib.pyplot):**

* pip install matplotlib
* **Functions:** Data visualization, creating plots, charts, histograms, scatter plots, line plots, bar plots, etc.

**Seaborn(seaborn):**

* Pip install seaborn
* **Functions:** Seaborn is a Python visualization library based on matplotlib that provides high-level functions for creating informative and attractive statistical graphics.

**Scikit-learn (sklearn.model\_selection):**

* Pip install scikit-learn
* Functions: Cross-validation and dataset splitting functions for generating training and testing sets.

DATA VISULAIZATION:

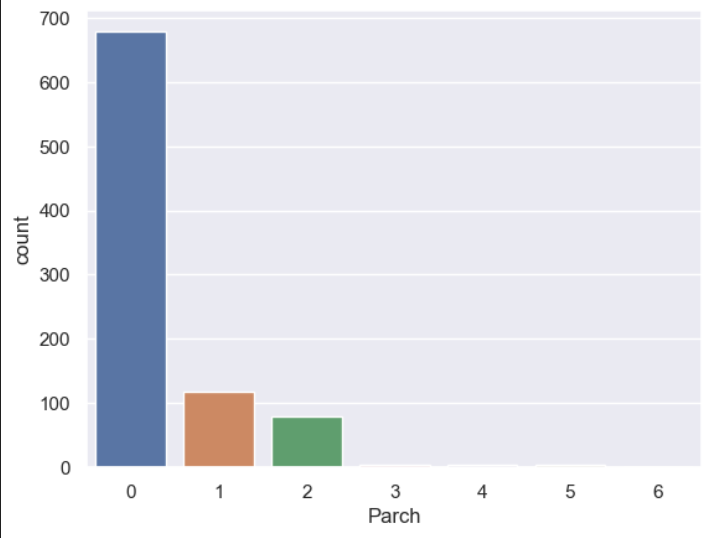
SEABORN PLOTS:

Seaborn is used for data visualization and it is based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

### Countplot:

Countplot shows the count of observations in each categorical bin using

sns.countplot(x='Parch', data=titanic\_data)



Adding hue parameter in countplot

sns.countplot(x='Parch', hue='Survived', data=titanic\_data)

