## **Artificial Intelligence and Machine Learning**

Project Report

Semester-IV (Batch-2022)

SURVIVAL PREDICTION ON TITANIC SHIP

A red and white sign

Description automatically generated with low confidence

**Supervised By: Submitted By:**

Rajeev Thakur Vansh Alawadhi

Vansh Duggal

Roll Number: -2210990933

2210990935

Group - 14

**Department of Computer Science and Engineering,** Chitkara University Institute of Engineering & Technology



* Data Collection:
  + Obtain the Titanic dataset from sources like Kaggle, which provides a CSV file containing information about passengers such as their names, ages, genders, ticket class, fare, cabin number, port of embarkation, and survival status.
* Data Preprocessing:
  + Handle missing values: Identify and handle missing values in the dataset. This might involve techniques like imputation (replacing missing values with a suitable estimate) or removal of rows or columns with too many missing values.
  + Encode categorical variables: Convert categorical variables (such as gender or port of embarkation) into numerical format using techniques like one-hot encoding or label encoding.
  + Feature scaling: Scale numerical features to ensure that they have a similar range and distribution, which can help improve the performance of certain machine learning algorithms.
* Exploratory Data Analysis (EDA):
  + Visualize distributions: Plot histograms and box plots to visualize the distributions of numerical features and identify potential outliers.
  + Analyze correlations: Compute correlation coefficients between features and the target variable (survival status) to identify relationships and dependencies.
  + Explore feature importance: Use techniques like feature importance plots or permutation importance to understand which features are most predictive of survival.
* Feature Engineering:
  + Create new features: Generate new features based on existing ones that might be more informative for predicting survival. For example, combine the "SibSp" and "Parch" features to create a new feature indicating family size.
  + Transform features: Transform skewed distributions or non-linear relationships using techniques like log transformation or polynomial features.



* Model Selection:
  + Choose appropriate algorithms: Select machine learning algorithms suitable for binary classification tasks, considering factors like interpretability, scalability, and performance.
  + Experiment with multiple algorithms: Compare the performance of different algorithms using techniques like cross-validation and select the one that performs best on the validation data.
* Model Training:
  + Split the dataset: Divide the dataset into training and testing sets to train the model on one subset and evaluate its performance on the other.
  + Train the model: Fit the selected machine learning algorithms to the training data to learn the underlying patterns and relationships.
* Model Evaluation:
  + Evaluate performance metrics: Calculate metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score to assess the model's performance on the testing data.
  + Interpret results: Analyze the model's predictions and errors to gain insights into its strengths and weaknesses.
* Hyperparameter Tuning:
  + Fine-tune hyperparameters: Use techniques like grid search, random search, or Bayesian optimization to search for the optimal hyperparameters that maximize the model's performance.
* Model Deployment:
  + Deploy the model: Once satisfied with its performance, deploy the trained model to make predictions on new data, whether it's through a web application, API, or any other system.

**CONCLUSION**

In conclusion, survival prediction on the Titanic dataset involves a systematic process of data collection, preprocessing, exploratory analysis, feature engineering, model selection, training, evaluation, hyperparameter tuning, and deployment. By following these steps, data scientists can build accurate predictive models to determine which passengers were more likely to survive the Titanic disaster. Through careful analysis of the dataset and application of machine learning techniques, insights can be gained into the factors that influenced survival rates, such as passenger demographics, ticket class, and family relationships. This exercise serves as an excellent introductory project in the field of data science, providing hands-on experience with real-world datasets and machine learning algorithms while highlighting the importance of data preprocessing, feature engineering, and model evaluation in predictive modeling tasks. Additionally, the survival prediction model developed using the Titanic dataset can have practical applications in understanding historical events, informing safety measures in transportation systems, and serving as a pedagogical tool for teaching machine learning concepts.

