**SYNOPSIS**

**ON**

**TITANIC SURVIVAL PREDICTION MODEL**

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE AND ENGINEERING**



**Submitted to:**

**Guide Name: Mr Manoj Kumar**

**Designation: Assistant Professor**

**Submitted by:**

Mohit Mishra - 200240101055

Pradeep Kumar Siware - 200240101068

Mahima Pandey - 200240101046

Ankit Tiwari - 200240101018

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING ROORKEE INSTITUTE OF TECHNOLOGY, ROORKEE**

**November- 2023**

Table of Contents

1. Project Title
2. Domain
3. Problem Statement
4. Project Description
5. Scope of the Work
6. Project Modules
7. Implementation Methodology
8. Technologies to be used
9. Software Platform
10. Hardware Platform:
11. Tools
12. Advantages of this Project
13. Future Scope and further enhancement of the Project
14. Team Details
15. Summary
16. References

**TITANIC SURVIVAL PREDICTION MODEL**

**A machine learning model can be used to predict the survival of passengers aboard the RMS Titanic**. Titanic Survival Prediction

The model takes into account various factors such as economic status, sex, age, etc., and compares and finds relations amongst various features to predict whether a passenger would survive on the Titanic.2 Exploratory data analysis, acquisition of training and testing data, modeling, and solving the problem are necessary to create the model.0 However, if we were asked to make a prediction about any passenger aboard the RMS Titanic whom we knew nothing about, the best prediction we could make would be that they did not survive. It is assumed that a majority of the passengers did not survive the ship sinking.

**DATA SCIENCE**

Data science is a multidisciplinary field that uses scientific methods, algorithms, processes, and systems to extract knowledge and insights from structured and unstructured data. It combines various aspects of statistics, computer science, domain expertise, and data analysis to understand, interpret, and utilize data to solve complex problems, make informed decisions, and support data-driven strategies.

Key components of data science include:

**Data Collection:** Gathering and acquiring data from various sources, which can be structured (e.g., databases) or unstructured (e.g., text, images, and videos).

**Data Cleaning and Preprocessing:** This involves organizing, cleaning, and transforming raw data to make it suitable for analysis. Data scientists often deal with missing values, outliers, and other data quality issues.

**Exploratory Data Analysis (EDA):** Exploring the data through visualization and summary statistics to gain initial insights and identify patterns or trends.

**Feature Engineering:** Selecting and creating relevant features (variables) that can improve the performance of machine learning models.

**Machine Learning:** Using algorithms and statistical models to build predictive or descriptive models from the data. This may include tasks such as classification, regression, clustering, and more.

**Data Visualization:** Presenting the results of data analysis in a visually understandable and informative way using charts, graphs, and other visualization techniques.

**Data Interpretation:** Translating the results into actionable insights and recommendations for decision-makers.

**Communication:** Effectively conveying findings and insights to non-technical stakeholders, such as executives or business leaders.

**Domain Expertise:** Understanding the specific industry or field in which the data is being analyzed to ensure the results are meaningful and relevant.

**PROBLEM STATEMENT**

The primary goal of this project is to create a predictive model that can determine the likelihood of a passenger's survival based on various attributes such as age, gender, class, and more. The problem statement can be summarized as follows:

"Given a dataset of Titanic passenger information, including demographic and ticket-related features, build a machine learning model that can predict whether a passenger survived (1) or did not survive (0) the sinking of the Titanic."

**PROJECT DESCRIPTION**

The "Titanic Survival Prediction" project is a data science and machine learning endeavor aimed at building a predictive model to determine the likelihood of survival for passengers on the ill-fated RMS Titanic. This project is inspired by the historic sinking of the Titanic in 1912, which resulted in a tragic loss of life, with many passengers and crew members perishing in the disaster.

**1. Data Collection:**

Gather a comprehensive dataset containing information about Titanic passengers, including their personal details, ticket class, cabin, and whether they survived or not.

**2. Data Preprocessing:**

Clean, explore, and preprocess the dataset to handle missing values, outliers, and feature engineering. This step involves ensuring the data is ready for model building.

**3. Exploratory Data Analysis (EDA):**

Perform data visualization and statistical analysis to gain insights into the relationships and patterns within the dataset. Explore factors that may have influenced passenger survival rates.

**4. Feature Engineering:**

Identify the most relevant features (independent variables) that contribute to the prediction of passenger survival. This step helps reduce model complexity and improves prediction accuracy.

**5. Model Building:**

Develop and train machine learning models using various algorithms such as logistic regression, decision trees, random forests, support vector machines, or neural networks. Evaluate and compare the performance of these models.

**6. Model Evaluation:**

Assess the predictive accuracy of the models using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Utilize techniques like cross-validation to ensure robust model performance.

**7.**  **Hyperparameter Tuning:**

Fine-tune the model hyperparameters to optimize model performance. This step involves using techniques like grid search or random search.

**8. Model Interpretation:**

Interpret the model's predictions and feature importance to understand which factors had the most influence on passenger survival.

**9. Deployment:**

Create a user-friendly interface or application to allow users to input passenger details and receive predictions about their likelihood of survival if they were on the Titanic.

**SCOPE OF THE WORK**

**1.Data Collection and Sources:** Gather the Titanic passenger data, including information about passengers such as age, gender, class, ticket fare, cabin, port of embarkation, and whether they survived or not. This data can be obtained from various sources, including datasets available online.

**2.Data Preprocessing:** Clean and preprocess the data to handle missing values, outliers, and format inconsistencies. This may include imputing missing values, converting categorical variables into numerical representations (e.g., one-hot encoding), and scaling or normalizing numerical features.

**3.Feature Selection and Engineering:** Create new features or transform existing ones to improve the predictive power of the model. This may involve extracting information from text fields, creating interaction terms, or engineering features that capture important relationships.

**4.Model Selection:** Choose appropriate machine learning algorithms for the task, such as logistic regression, decision trees, random forests, support vector machines, or neural networks. Selecting the right model may involve experimentation and comparison to determine which one performs best.

**5.Model Training:** Split the data into training and validation sets to train and evaluate the selected models. Hyperparameter tuning may be necessary to optimize model performance.

**6.Model Deployment:** *Specify how the* trained model will be deployed in a production environment, such Titanic Survival Prediction. Consider integration with existing systems and processes.

**7.Documentation and Reporting:** Outline the documentation requirements, including the creation of comprehensive reports that summarize the project, its findings, and model performance.

**8.Maintenance and Updates:** Define the scope for ongoing model monitoring, maintenance, and updates. This includes addressing changing lending conditions, regulatory requirements, and model performance issues over time.

**9.Ethical Considerations:** Incorporate ethical considerations into the project scope, including data privacy, informed consent, and the responsible use of AI in lending.

**10.** **Reporting:** Ensure Prepare a report or presentation summarizing the project's findings, insights, and the model's performance. Visualizations and explanations can make the results more accessible to a wider audience.

**11.Future Enhancements**: Identify potential improvements to the model, such as feature engineering, using more advanced machine learning techniques, or incorporating additional data sources for better predictions.

**14.Ethical Considerations:** Consider the ethical implications of the project, including issues related to bias, fairness, and privacy. Ensure that the model's predictions do not disproportionately affect specific groups unfairly.

The scope of a Titanic survival prediction project can vary based on its specific objectives and the depth of analysis desired. It is a good practice to define the project's goals, timelines, and resources at the outset to ensure its successful execution.

**PROJECT MODULES**

A " Titanic Survival Prediction " project typically consists of several modules or components, each of which performs a specific task or set of related tasks. These modules work together to create a comprehensive Titanic Survival Prediction system. Below, we outline the common modules for a Titanic Survival Prediction project:

**1.Data Collection Module:**

\*Collect and aggregate historical Titanic Survival Prediction application data from various sources.

\*Store and organize the data in a suitable database or data storage system.

**2.Data Preprocessing Module:**

\*Clean and prepare the data for analysis, including handling missing values, duplicates, and outliers.

\*Perform data imputation, encoding of categorical variables, and feature scaling.

**3.Exploratory Data Analysis (EDA) Module:**

\*Conduct data exploration to gain insights into the dataset.

\*Create summary statistics, data visualizations, and identify correlations between features.

**4.Feature Engineering Module:**

\*Create new features or transform existing ones to improve the model's predictive power.

\*Calculate financial metrics, generate feature interactions, and perform dimensionality reduction if necessary.

**5.Model Building Module:**

\*Select machine learning algorithms suitable for the Titanic Survival Prediction prediction task.

\*Develop and train predictive models using the prepared dataset.

**6.Model Evaluation Module:**

\*Split the dataset into training and testing sets.

\*Evaluate model performance using various metrics (e.g., accuracy, precision, recall, F1-score, ROC AUC).

\*Implement cross-validation techniques for robust assessment.

**IMPLEMENTATION METHODOLOGY**

**1.Define the Problem:**

Clearly define the problem you want to solve. In this case, it's predicting whether Titanic Survival Prediction will default or not.

**2.Data Collection:**

Gather historical data related to Titanic Survival Prediction applications.

**3.Data Preprocessing:**

Prepare the data for analysis and modeling. This includes handling missing values, encoding categorical variables, and scaling numerical features. Common

***preprocessing steps include:***

***Data Cleaning***: Removing duplicates, handling missing values.

***Feature Engineering:*** Creating new features or transforming existing ones.

***Encoding Categorical Variables***: Converting categorical data into numerical form (e.g., one-hot encoding or label encoding).

***Scaling:*** Standardizing or normalizing numerical features.

**4.Exploratory Data Analysis (EDA):**

Visualize and analyze the data to gain insights into the relationships between features and the target variable. EDA helps you understand the data and identify potential patterns.

**5.Data Splitting:**

Split the data into training and testing sets. The training set is used to train the model, and the testing set is used to evaluate its performance.

**6.Model Selection:**

Choose a machine learning algorithm or a combination of algorithms that are suitable for the problem. Common algorithms for binary classification problems like Titanic Survival Prediction include logistic regression, decision trees, random forests, and support vector machines.

**7.Model Training:**

Train the selected model(s) on the training data. Use cross-validation techniques to fine-tune hyperparameters and ensure the model's generalization ability.

**8.Model Evaluation:**

Evaluate the model's performance on the testing data using appropriate evaluation metrics like accuracy, precision, recall, F1 score, and ROC AUC. Choose the metrics that are most relevant to the problem.

**TECHNOLOGIES TO BE USED**

## Jupyter Notebook:

Jupyter Notebook is a free and open-source online software that lets us create and share documents with live code, equations, visualisations, and narrative prose. It is frequently used in data science, scientific computing, and other sectors that need interactive computing and data exploration.

Jupyter Notebook is an interactive computing environment that allows you to create and run code in a variety of computer languages, including Python, R, Julia, and others. It combines code execution, text, equations, and visualisations into a single document, making it a powerful data analysis, experimentation, and documentation tool.

A cell is the primary building unit of a Jupyter Notebook. A cell may hold both code and markdown text. Markdown cells enable you to write explanatory text, insert equations, and format the information using Markdown syntax, whereas code cells allow you to write and run code.

Jupyter Notebook's ability to execute code cells interactively is one of its most important features. Running a code cell and seeing the output right in the notebook makes it simple to experiment, iterate, and visualise outcomes. Jupyter Notebook is a good tool for data exploration, analysis, and dissemination of research findings due to its interactivity and ability to blend code with explanatory prose and visualisations.

## MS-EXCEL:

Microsoft Excel is a strong spreadsheet application. It is part of the Microsoft Office family of tools and is frequently used for tabular data organization, analysis, and manipulation. Excel has a grid-like interface with rows and columns that allows users to enter, modify, and format data.

Excel features and functionality include mathematical and statistical operations, data visualisation tools, and the ability to generate charts, graphs, and tables. It accepts a variety of data kinds, including numbers, text, dates, and formulae, and allows users to make computations and automate processes.

## POWER BI:

Power BI is a business intelligence and data visualization tool developed by Microsoft. It is designed to help individuals and organizations transform raw data into meaningful insights, interactive reports, and visually appealing dashboards. Power BI is widely used for data analysis, reporting, and decision-making across various industries.

**SOFTWARE PLATFORM**

## Python Language

The coding technique on developing the Virtual Mouse application will be the Python with the aid of the integrated development environment (IDE) that are used for developing computer programs, known as the PyCharm. A Python library provides various operators, covering basic arithmetic, bit manipulation, indirection, comparisons, logical operations and others.

* **Numpy:**

NumPy, short for "Numerical Python," is a fundamental Python library for numerical and scientific computing. It provides support for large, multi-dimensional arrays and matrices, along with a variety of high-level mathematical functions to operate on these arrays. NumPy is the foundation for many other libraries used in data science, machine learning, and scientific research

* **Pandas:**

Pandas is a popular Python library for data manipulation and analysis. It provides data structures and functions necessary to work with structured data, making it an essential tool for data scientists, analysts, and developers dealing with tabular or labeled data.

* **Matplotlib:**

Matplotlib is a popular Python library for creating static, animated, and interactive visualizations. It is highly customizable and provides a wide range of tools for generating various types of plots and charts. Matplotlib is often used in data analysis, scientific research, and data visualization tasks.

* **Seaborn:**

Seaborn is a popular Python data visualization library built on top of Matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn is particularly useful for creating aesthetically pleasing visualizations with minimal code.

## Sklearn Library:

Scikit-learn, sometimes known as sklearn, is a popular open-source Python machine learning package. It includes a variety of tools and techniques for data preparation, feature selection, model training, and model assessment, among other things. Scikit-learn is based on other scientific Python libraries like NumPy, SciPy, and Matplotlib, and it interacts nicely with the greater Python environment.

## MS-Excel:

Microsoft Excel is a strong spreadsheet application. It is part of the Microsoft Office family of tools and is frequently used for tabular data organization, analysis, and manipulation. Excel has a grid-like interface with rows and columns that allows users to enter, modify, and format data.

Software will be using:

**OS:** Windows 7

**Language:** Python

**Tool Used:** Power BI and MS Excel

**Editor:** Jupyter notebook

**HARDWARE PLATFORM**

## Computer Desktop or Laptop

The computer desktop or a laptop will be utilized to run the visual software in order to display what webcam had captured. A notebook which is a small, lightweight and inexpensive laptop computer being proposed to increase mobility.

**Processor:** Core2Duo

**System type:** 32-bit OS, x32-based processor

**Main Memory:** 4GB RAM

**SSD:** 256GB

**Display:** 15.2" Monitor

**TOOLS**

## Jupyter Notebook:

Jupyter Notebook is a free and open-source online software that lets us create and share documents with live code, equations, visualisations, and narrative prose. It is frequently used in data science, scientific computing, and other sectors that need interactive computing and data exploration.

Jupyter Notebook is an interactive computing environment that allows you to create and run code in a variety of computer languages, including Python, R, Julia, and others. It combines code execution, text, equations, and visualisations into a single document, making it a powerful data analysis, experimentation, and documentation tool.

A cell is the primary building unit of a Jupyter Notebook. A cell may hold both code and markdown text. Markdown cells enable you to write explanatory text, insert equations, and format the information using Markdown syntax, whereas code cells allow you to write and run code.

Jupyter Notebook's ability to execute code cells interactively is one of its most important features. Running a code cell and seeing the output right in the notebook makes it simple to experiment, iterate, and visualise outcomes. Jupyter Notebook is a good tool for data exploration, analysis, and dissemination of research findings due to its interactivity and ability to blend code with explanatory prose and visualisations.

**Power BI:**

Power BI is a business intelligence and data visualization tool developed by Microsoft. It is designed to help individuals and organizations transform raw data into meaningful insights, interactive reports, and visually appealing dashboards. Power BI is widely used for data analysis, reporting, and decision-making across various industries. Here are some key features and components of Power BI:

**Power BI Desktop:**

Power BI Desktop is a Windows application that allows users to create interactive reports and data visualizations. It provides a user-friendly interface for importing, transforming, and modeling data from various sources.

**Power Query:**

Power Query is a data transformation tool integrated into Power BI. It enables users to connect to different data sources, perform data cleansing, combine data from multiple sources, and apply various data transformations.

**Power Pivot:**

Power Pivot is an in-memory data modeling tool that allows users to create data models, relationships, and calculations within Power BI Desktop. It's particularly useful for managing large datasets and creating complex calculations.

**Power View and Power Map:**

Power View and Power Map are tools for creating interactive reports and geographical visualizations within Power BI. These components enable users to explore data using dynamic charts and maps.

**DAX (Data Analysis Expressions):**

DAX is a formula language used in Power BI for creating custom calculations and measures. It's similar to Excel formulas but tailored for data modeling.

**Git hub:**

GitHub is a web-based platform for version control and collaborative software development. It is widely used by developers and teams for managing, sharing, and collaborating on code projects.

**ADVANTAGES OF THIS PROJECT**

**Historical Significance:**

The Titanic dataset is historically significant, making it a compelling project for data science and machine learning enthusiasts. Analyzing and predicting survival on the Titanic allows individuals to engage with a well-known historical event in a meaningful way.

**Educational Tool:**

This project is an excellent learning opportunity for individuals who are new to data science and machine learning. It provides hands-on experience with data preprocessing, feature engineering, model selection, and evaluation.

**Interdisciplinary Skills:**

The project requires skills from multiple disciplines, including statistics, data analysis, programming, and machine learning. As a result, it helps participants develop a broad skill set.

**Real Data Application:**

The Titanic dataset is real data, which makes the project more relevant and relatable. It demonstrates the practical application of data science techniques to solve real-world problems.

**Feature Engineering:**

Feature engineering is a crucial aspect of the project. It encourages creative thinking and problem-solving to extract meaningful information from the available data.

**Interpretability:**

Understanding how different factors influence survival predictions provides valuable insights. Participants can learn how to interpret model results and understand feature importance.

**Ethical Considerations:**

The project encourages participants to think about and address ethical concerns, such as bias and fairness in machine learning. It's an opportunity to practice responsible AI.

**Communication Skill:**

Preparing a report or presentation about the project helps develop communication skills. Clear explanations and visualizations can make complex results more accessible to a broader audience.

**Portfolio Enhancement**:

For students or job seekers, a well-executed Titanic survival prediction project can be a valuable addition to a data science or machine learning portfolio, demonstrating practical skills and knowledge.

**Community and Collaboration**:

Many online communities and forums are dedicated to Titanic datasets. Engaging in these communities can foster collaboration, learning, and knowledge sharing.

**Industry Relevance**:

The skills and experience gained from this project are transferable to various industries and applications, making it a practical starting point for a career in data science or machine learning.

**Future Scope and Further Enhancement of Project**

**Advanced Modeling Techniques:**

Experiment with more advanced machine learning techniques, such as gradient boosting, deep learning, and ensemble methods, to improve prediction accuracy.

**Explainable AI (XAI):**

Implement techniques for model interpretability and explanation to understand the factors contributing to survival predictions. This is crucial for gaining insights and ensuring transparency.

**Bias and Fairness Analysis:**

Investigate and address potential bias and fairness issues in the model. Consider fairness-aware machine learning techniques to ensure that predictions are not unfairly skewed against certain groups.

**Feature Engineering:**

Explore more complex feature engineering techniques, such as creating interaction features, time-based features, or engineering features based on domain knowledge.

**Dynamic Predictions:**

Extend the project to create a system that can provide dynamic predictions, updating survival predictions as new data becomes available or as additional information is provided.

**Survival Clustering:**

Explore clustering techniques to group passengers based on similarities in their survival predictions. This can help identify passenger segments with varying survival rates.

**Natural Language Processing (NLP):**

Analyze and extract insights from textual information, such as passenger names, ticket descriptions, or cabin information using NLP techniques.

**Collaboration and Competition:**

Participate in data science competitions or collaborate with others to compare your model's performance with those of other data scientists. Platforms like Kaggle often host Titanic prediction competitions.

**Extend to Other Datasets**:

Apply the skills and knowledge gained from the Titanic dataset to similar historical events or datasets. This can provide a broader perspective on data analysis and modeling.

**Team Details**

| **Group#** | **Course Name** | **Student ID** | **Student Name** | **Role** | **Signature** |
| --- | --- | --- | --- | --- | --- |
| TITANIC SURVIVAL PREDICTION MODEL | B.Tech | 200240101055 | Mohit Mishra | Data Collection |  |
| B.Tech | 200240101068 | Pradeep Kumar  Siware | ML Model training and testing |  |
| B.Tech | 200240101046 | Mahima Pandey | Data Preprocessing |  |
| B.Tech | 200240101018 | Ankit Tiwari | EDA |  |

**SUMMARY**

The Titanic survival prediction project is a valuable endeavor for learning and applying data science and machine learning techniques. It offers historical significance, educational benefits, and opportunities for skill development. To enhance the project's future scope, you can explore advanced modeling techniques, feature engineering, deployment, bias mitigation, and more. Additionally, you can expand the project to encompass related datasets, create educational materials, and continue to address ethical considerations in your work. The Titanic project serves as a foundation for more advanced data science and machine learning projects while allowing you to gain valuable insights and knowledge in the field.

**REFERENCES**

For Jupyter Notebook Installation: <https://www.anaconda.com/download>

For Power BI: <https://powerbi.microsoft.com/en-in/downloads/>

Dataset: <https://www.kaggle.com/>

Git hub link for Titanic Survival Prediction: <https://github.com/MohitMishra/Titanic-Survival-Prediction>

Git hub link for Power BI file: <https://github.com/MohitMishra/Titanic-Survival-Prediction-Dashboard>