**ENGLISH PREMIER LEAGUE**

**MATCH PREDICTION**

**A MINI PROJECT REPORT**

***Submitted by***

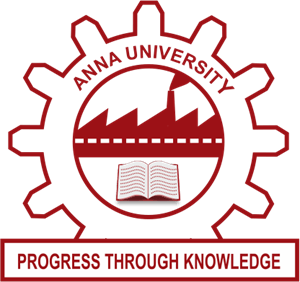
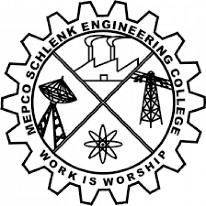
**Jayavelan S (Reg. No. 9517202309)**

**Nandhagopal V (Reg. No. 9517202309)**

**Sanjay K N (Reg. No. 9517202309101)**

***in partial fulfillment for the requirement of the course***

**23AD453 – MINI PROJECT II: DATA ANALYTICS**

*** ***

**DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**MEPCO SCHLENK ENGINEERING COLLEGE (AUTONOMOUS)**

**SIVAKASI**

**APRIL 2025**

**BONAFIDE CERTIFICATE**



This is to certify that it is the bonafide work done by Jayavelan S (9517202309), Nandhagopal V (9517202309), Sanjay K N (9517202309101) for the course **23AD453 - Mini Project – II: DATA ANALYTICS** work entitled **“English Premier League Match Prediction”** at Department of Artificial Intelligence & Data Science, Mepco Schlenk Engineering College, Sivakasi during the year 2024 – 2025.

………………………… ……………………………

***Faculty In-Charge Head of the Department***

Submitted for the Practical Examination held at Mepco Schlenk Engineering College, Sivakasi on ……………………….

………………………… ……………………………

***Internal Examiner - I Internal Examiner - II***

**ACKNOWLEDGEMENT**

First and foremost, we express our deepest gratitude to the **LORD ALMIGHTY** for His abundant blessings, which have guided us through our past, present, and future endeavors, making this project a successful reality.

We extend our sincere thanks to our management, esteemed Principal, **Dr. S. Arivazhagan, M.E., Ph.D.,** for providing us with a conducive environment, including advanced systems and well-equipped library facilities, which have been instrumental in the completion of this project.

We are profoundly grateful to our Head of the Artificial Intelligence and Data Science Department, **Dr. J. Angela Jennifa Sujana, M.Tech., Ph.D.,** for granting us this golden opportunity to work on a project of this significance and for her invaluable guidance and encouragement throughout.

Our heartfelt appreciation goes to our project guides, **Dr. R. Nirmalan, M.Tech., Ph.D.,** Assistant Professor, and **Mrs.A. Revatheeswari, M.Tech., Ph.D.,** Assistant Professor, Department of Artificial Intelligence and Data Science for their unwavering support, insightful suggestions, and expert guidance. Their experience and encouragement have been vital in shaping our project and helping us bring forth our best.

We also extend our sincere gratitude to all our faculty members, lab technicians, and our beloved family and friends, whose timely assistance and moral support played a pivotal role in making this mini-project a success.

**ABSTRACT**

**List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Title of Table** | **Page No.** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title of Figure** | **Page No.** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**List of Symbols and Abbreviations**

|  |  |
| --- | --- |
| **Symbol / Abbreviation** | **Description** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

**SYNOPSIS**

**Module 1: Problem Statement Identification**

This module is dedicated to understanding the problem at hand and structuring data effectively for analysis. It emphasizes data warehouse schemas, which form the backbone of organizing and managing large datasets. By leveraging the Star, Snowflake, and Constellation schemas, students can structure their data efficiently. Additionally, the module covers the analytical operations of Online Analytical Processing (OLAP), which facilitates multidimensional data analysis, aiding in problem scoping and decision-making.

**Module 2: Data Acquisition**

This module focuses on gathering data from various sources and preparing it for analysis. Students will explore different dataset types, including structured (tabular data), semi-structured (JSON, XML), and unstructured (text, images, audio). The module ensures proper categorization of attributes into numerical, categorical, ordinal, and nominal data. Further, students will learn to identify independent (features) and dependent (target) attributes, derive new attributes when necessary, and perform feature selection to retain only the most relevant attributes for analysis. The ability to source and organize high-quality data is fundamental to ensuring meaningful insights in later stages.

**Module 3: Preprocessing of Datasets**

Before applying machine learning algorithms, raw data must undergo preprocessing to enhance its quality and usability. This module teaches students to clean data by handling missing values, detecting and removing outliers, and eliminating duplicate records. Using Weka, students will perform data normalization to scale features within a standard range, standardization to achieve a consistent mean and variance, and transformation techniques such as one-hot encoding and discretization. Additionally, dimensionality reduction techniques, like Principal Component Analysis (PCA), will be explored to optimize model performance while reducing computational complexity.

**Module 4: Model Generation**

Building predictive models is the core objective of this module. It covers different machine learning techniques, including association rule mining, regression, classification, and clustering models. Students will work with algorithms like Apriori and FP Growth for association rule mining, various regression models (linear, multiple linear, polynomial, logistic), and classification models such as Decision Trees (ID3) and Naïve Bayes. For clustering, methods like k-means, Agglomerative Clustering, and DBSCAN will be applied. This module provides a hands-on approach to selecting and implementing suitable models for diverse data-driven problems.

**Module 5: Testing of Model**

Once a model is built, evaluating its performance is crucial to ensure its reliability and accuracy. This module introduces various evaluation metrics tailored for different models. For classification models, key metrics include accuracy, precision, recall, F1-score, and ROC-AUC. Regression models will be evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared. Clustering models will be assessed using the Silhouette Score, Dunn Index, and Davies-Bouldin Index, while association rule mining will involve Support, Confidence, and Lift. Additionally, visualization techniques such as confusion matrices, ROC curves, scatter plots, and heatmaps will help interpret model results effectively.

**Module 6: Deployment**

The final step in the data science workflow is deploying the developed models for practical use. This module focuses on finalizing the trained models, integrating them into a user-friendly interface, and ensuring their seamless deployment. Students will learn how to prepare the final model, develop a functional UI for interaction, and host the model on appropriate platforms. This stage bridges the gap between theoretical learning and real-world implementation, allowing users to interact with the model in a meaningful way.

**Table of Contents**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chapter No.** | **Contents** | | | | **Page No.** |
|  | Abstract | | | |  |
|  | List of Tables | | | |  |
|  | List of Figures | | | |  |
|  | List of Symbols and Abbreviations | | | |  |
|  | Synopsis | | | |  |
| **1** | **INTRODUCTION** | | | |  |
|  | 1.1 | Overview of the Project | | |  |
|  | 1.2 | Objectives | | |  |
|  | 1.3 | Scope of the Project | | |  |
|  | 1.4 | Tools & Programming Languages Used | | |  |
| **2** | **PROBLEM STATEMENT IDENTIFICATION** | | | |  |
|  | 2.1 | Understanding the Problem | | |  |
|  | 2.2 | Data Warehouse Design | | |  |
|  | 2.2.1 | | Star Schema | |  |
|  | 2.2.2 | | Snowflake Schema | |  |
|  | 2.2.3 | | Constellation Schema | |  |
|  | 2.3 | OLAP and Analytical Operations | | |  |
| **3** | **DATA ACQUISITION** | | | |  |
|  | 3.1 | Data Collection Methods | | |  |
|  | 3.2 | Types of Data | | |  |
|  | 3.3 | Attribute Categorization and Classification | | |  |
|  | 3.3.1 | | Independent vs. Dependent Attributes | |  |
|  | 3.3.2 | | Feature Selection and Extraction | |  |
| **4** | **DATA PREPROCESSING** | | | |  |
|  | 4.1 | Data Cleaning | | |  |
|  | 4.1.1 | | Handling Missing Values | |  |
|  | 4.1.2 | | Removing Outliers and Duplicates | |  |
|  | 4.2 | Data Normalization and Standardization | | |  |
|  | 4.3 | Data Transformation | | |  |
|  | 4.3.1 | | Encoding Categorical Data | |  |
|  | 4.3.2 | | Feature Selection Techniques | |  |
| **5** | **MODEL GENERATION** | | | |  |
|  | 5.1 | Association Rule Mining | | |  |
|  | 5.1.1 | | Apriori Algorithm | |  |
|  | 5.1.2 | | FP-Growth Algorithm | |  |
|  | 5.2 | Regression Models | | |  |
|  | 5.2.1 | | Linear Regression | |  |
|  | 5.2.2 | | Multiple Linear Regression | |  |
|  | 5.2.3 | | Logistic Regression | |  |
|  | 5.3 | Classification Models | | |  |
|  | 5.3.1 | | | Decision Tree Classifier |  |
|  | 5.3.2 | | | Naïve Bayes Algorithm |  |
|  | 5.3.3 | | | K-Nearest Neighbor (KNN) Classification |  |
|  | 5.4 | Clustering Models | | |  |
|  | 5.4.1 | | | k-means Algorithm |  |
|  | 5.4.2 | | | Agglomerative Clustering |  |
|  | 5.4.3 | | | DBSCAN Clustering |  |
| **6** | **TESTING OF MODEL** | | | |  |
|  | 6.1 | Evaluation Metrics | | |  |
|  | 6.1.1 | | | Association Rule Mining Evaluation |  |
|  | 6.1.2 | | | Regression Model Evaluation |  |
|  | 6.1.3 | | | Classification Model Evaluation |  |
|  | 6.1.4 | | | Clustering Model Evaluation |  |
|  | 6.2 | Results Visualization | | |  |
|  | 6.2.1 | | | ROC Curve |  |
|  | 6.2.2 | | | Confusion Matrix |  |
|  | 6.2.3 | | | Scatter Plots and Heatmaps |  |
| **7** | **DEPLOYMENT** | | | |  |
|  | 7.1 | Finalizing the Model | | |  |
|  | 7.2 | User Interface (UI) Integration | | |  |
|  | 7.2.1 | | | Tools Used |  |
|  | 7.2.2 | | | Deployment |  |
| **8** | **CONCLUSION** | | | |  |
|  | 8.1 | Summary of Findings | | |  |
|  | 8.2 | Challenges Faced | | |  |
|  | 8.3 | Future Enhancements | | |  |
| **9** | **REFERENCES** | | | |  |
| **10** | **APPENDICES** | | | |  |

**CHAPTER 1**

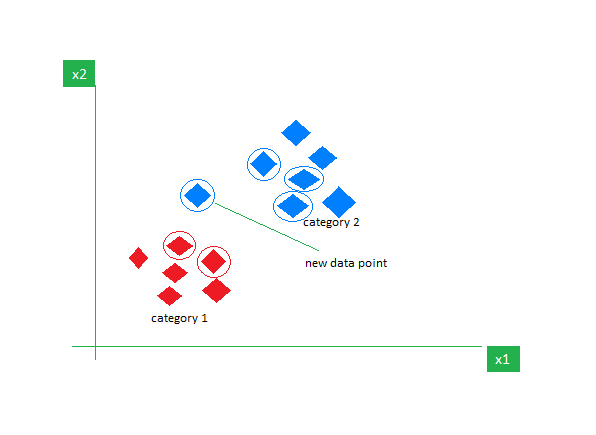
**INTRODUCTION**

K-Nearest Neighbors (KNN) is a simple way to classify things by looking at what’s nearby. Imagine a streaming service wants to predict if a new user is likely to cancel their subscription (churn) based on their age. They checks the ages of its existing users and whether they churned or stayed. If most of the “K” closest users in age of new user canceled their subscription KNN will predict the new user might churn too. The key idea is that users with similar ages tend to have similar behaviors and KNN uses this closeness to make decisions.

**1.1 Getting Started with K-Nearest Neighbors**

K-Nearest Neighbors is also called as a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification it performs an action on the dataset.

As an example, consider the following table of data points containing two features:



**Fig.1.1 KNN Algorithm Visualization**

**CHAPTER 9**

**REFERENCES**

*(FORMAT: [1]. Author, "Title of paper," Title of Journal, vol. 1, no. 1, pp. 1-10, Month Year.)*

[1] I.A. Glover and P.M. Grant, Digital Communications, 3rd ed. Harlow: Prentice Hall, 2009.

[2] C. W. Li and G. J. Wang, "MEMS manufacturing techniques for tissue scaffolding devices," in Mems for Biomedical Applications, S. Bhansali and A. Vasudev, Eds. Cambridge: Woodhead, 2012, pp. 192-217.