A Capstone Project report submitted

in partial fulfillment of requirement for the award of degree

**BACHELOR OF TECHNOLOGY**

in

**SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE**

by

# 2203A52192 T.HEMANTH KUMAR

Under the guidance of

**Dr.Ramesh Dadi**

Assistant Professor, School of CS&AI.



SR University, Ananthsagar,Warangal,Telagnana-506371

i

# CONTENTS

**S.NO. TITLE PAGE NO.**

1. DATASET 1
2. METHODOLOGY 2- 3
3. RESULTS 4 - 6

**CHAPTER 1**

# DATASET

## Project -1

The Study Hours dataset contains data on the number of hours students studied and their corresponding exam scores. It is commonly used to demonstrate linear regression, helping analyze how study time influences academic performance. This dataset is ideal for basic predictive modeling and understanding the relationship between effort and results.

## Project – 2

The Supermarket Sales dataset contains records of customer transactions from a retail supermarket. It includes details such as branch location, customer type, gender, product category, quantity, price, payment method, and customer rating. This data helps analyze sales trends, customer preferences, and branch performance.

## Project – 3

This dataset provides comprehensive metadata for Netflix titles, including information such as title names, types (Movie or TV Show), genres, directors, cast members, countries of origin, release years, ratings, and durations. Analyzing this dataset allows us to explore various patterns, such as the growth of Netflix’s content over time, the geographical distribution of its titles, popular genres, and the influence of key directors or actors. These insights can aid in understanding content strategies, viewer preferences, and potential market trends within the streaming industry.

**CHAPTER 2**

**METHODOLOGY**

# Project – 1

Data Collection:

The dataset consists of two variables: hours studied and corresponding exam scores. The data is assumed to be clean, with no missing or invalid values.

Data Visualization:

A scatter plot was used to visualize the relationship between study hours and scores, helping identify the linear trend.

Model Building:

A simple linear regression model was applied, where the number of study hours is the independent variable and the exam score is the dependent variable.

Training and Testing:

The dataset was split into training and testing sets to evaluate the model’s performance on unseen data.

Model Evaluation:

The model’s accuracy was assessed using performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R² Score.

# Project -2

Data Import and Cleaning:

The dataset was imported using pandas. Null values and duplicates were checked and handled to ensure data quality.

Exploratory Data Analysis (EDA):

Descriptive statistics and visualizations (bar charts, pie charts, histograms) were used to understand patterns in customer demographics, sales by branch, product line performance, and payment methods.

Time-Based Analysis:

Date and time columns were used to identify trends over time, such as peak shopping hours and busiest days.

Sales and Revenue Insights:

Total sales, taxes, and quantity sold were analyzed to assess the overall financial performance across branches and product categories.

Customer Behavior Analysis:

Customer type, gender, and ratings were analyzed to understand preferences and satisfaction levels.

# Project – 3

Data Loading and Cleaning:

The dataset was imported and inspected for missing or duplicate values. Missing data in key columns like director, cast, and country were handled appropriately.

Exploratory Data Analysis (EDA):

Visualizations and summary statistics were used to understand content distribution by type (Movie/TV Show), genre, country, release year, and rating.

Time-Based Analysis:

The 'date\_added' and 'release\_year' columns were analyzed to identify trends in Netflix content over time.

Content-Based Insights:

The most frequent directors, actors, and genres were identified to understand Netflix’s content focus and diversity.

Visualization:

Graphs such as bar plots, pie charts, and time series plots were used to present insights in a clear and interpretable manner.

**CHAPTER 3**

# RESULTS

## Project – 1

RangeIndex: 25 entries, 0 to 24

Data columns (total 2 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Hours 25 non-null float64

1 Scores 25 non-null int64

dtypes: float64(1), int64(1)

memory usage: 532.0 bytes

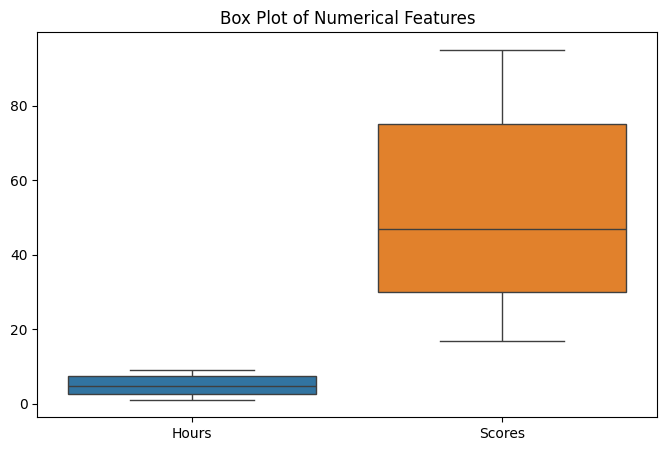
None

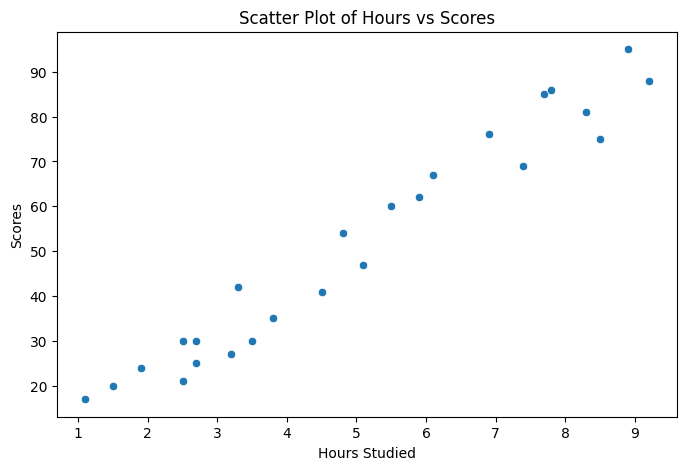
Skewness:

Hours 0.190338

Scores 0.233965

dtype: float64





Model Performance:

Linear Regression:

MAE: 3.9208

MSE: 18.9432

R2 Score: 0.9678

Decision Tree:

MAE: 5.4000

MSE: 31.7000

R2 Score: 0.9461

Random Forest:

MAE: 3.1400

MSE: 14.9402

R2 Score: 0.9746

# Project-2

P(A) = 0.5010

P(B) = 0.3110

P(A ∩ B) = 0.1720

P(A ∪ B) = 0.6400

P(A | B) = 0.5531

P(Health & Beauty) = 0.1520

P(Quantity > 5 | Ewallet) = 0.4870

P(Cash | Yangon) = 0.3235

P(Member & Rating > 8) = 0.1570

Permutations of 5 transactions from 20: 1860480

Combinations of 3 product lines from 6: 20

Factorial of branch count: 6

Permutations of 3 payment methods from 3: 6

Combinations of 5 employees from 15: 3003

Unique invoice codes (6 letters from 26): 165765600

# Project – 3

**Epoch 1/5**

**/usr/local/lib/python3.11/dist-packages/keras/src/layers/core/embedding.py:90: UserWarning: Argument `input\_length` is deprecated. Just remove it.**

**warnings.warn(**

**111/111 ━━━━━━━━━━━━━━━━━━━━ 11s 77ms/step - accuracy: 0.4939 - loss: 0.6946 - val\_accuracy: 0.5011 - val\_loss: 0.6933**

**Epoch 2/5**

**111/111 ━━━━━━━━━━━━━━━━━━━━ 10s 73ms/step - accuracy: 0.5018 - loss: 0.6935 - val\_accuracy: 0.5011 - val\_loss: 0.6931**

**Epoch 3/5**

**111/111 ━━━━━━━━━━━━━━━━━━━━ 11s 78ms/step - accuracy: 0.5047 - loss: 0.6936 - val\_accuracy: 0.5011 - val\_loss: 0.6932**

**Epoch 4/5**

**111/111 ━━━━━━━━━━━━━━━━━━━━ 9s 66ms/step - accuracy: 0.5003 - loss: 0.6938 - val\_accuracy: 0.4989 - val\_loss: 0.6932**

**Epoch 5/5**

**111/111 ━━━━━━━━━━━━━━━━━━━━ 10s 66ms/step - accuracy: 0.4999 - loss: 0.6933 - val\_accuracy: 0.4989 - val\_loss: 0.6934**

