Internship Project Report

Intern Details:

Name: Harshan Attar

College: Ajeenkya DY Patil School of Engineering

Course: B.E. (AIDS)

Class: TE - A

Roll.No: A - 12

Intern at: iGurus Consultancy Services LLP

Position: Research and Analytics Intern

Project 1

Title:

Predicting Students Performance using Simple Linear Regression

Problem Statement:

Develop a model to predict students' marks based on the number of hours they study. This project aims to provide a straightforward tool for estimating academic performance, helping students and educators understand the relationship between study habits and achievement.

Objectives:

1. **Model Development**: Develop a predictive model to estimate students' marks based on their study hours.
2. **Insight Generation**: Generate insights into the impact of study habits on academic achievement.
3. **Tool Development:** Create a user-friendly tool for students and educators to estimate academic performance based on study hours.

Code File:

https://github.com/Harshan-Attar/Data-Science/tree/800edec29f1bafc0d66a39ffdb6968e675b8e89b/Projects/Project%201

Project 2

Title:

Salary Prediction: Polynomial Regression Analysis for Job Positions

Problem Statement:

Utilizing polynomial regression on the provided dataset, which includes positions and corresponding salaries, to predict salary levels for different positions. This project aims to explore the relationship between job positions and salaries, providing insights into salary trends and assisting in salary negotiation strategies.

Objectives:

1. **Dataset Exploration:** Explore the dataset to understand the distribution of positions and corresponding salaries.
2. **Polynomial Regression:** Implement polynomial regression to model the relationship between positions and salaries.
3. **Insight Generation:** Generate insights into salary trends based on the polynomial regression model.
4. **Salary Negotiation Strategies**: Provide recommendations for salary negotiation strategies based on the model's predictions.

Code File:

https://github.com/Harshan-Attar/Data-Science/tree/800edec29f1bafc0d66a39ffdb6968e675b8e89b/Projects/Project%202

Project 3

Title:

Profit Prediction for Startups Using Multiple Linear Regression.

Problem Statement:

Develop a predictive model to estimate the profit of startups based on their R&D spending, administration spending, marketing spending, and location. The goal is to accurately forecast the profit of startups using multiple linear regression, leveraging the 50 startups dataset containing information from various companies in New York, California, and Florida. This model will assist stakeholders in making informed decisions regarding resource allocation and market strategies.

Objectives:

1. **Data Exploration**: Explore the dataset to understand the distributions and relationships between administration, research and development, marketing spend, and profits.
2. **Multiple Linear Regression:** Implement multiple linear regression to model the relationship between operational aspects and profits.
3. **Impact Analysis:** Analyze the impact of administration, research and development, and marketing spend on startup profitability.
4. **Strategic Decision-making:** Provide insights for strategic decision-making and resource allocation for aspiring entrepreneurs based on the model's predictions.

Code File:

https://github.com/Harshan-Attar/Data-Science/tree/800edec29f1bafc0d66a39ffdb6968e675b8e89b/Projects/Project%203

Project 4

Title:

Diabetes Prediction: Decision Tree Classifier Evaluation

Problem Statement:

Implementing a decision tree classifier on the provided dataset containing information related to diabetes, such as glucose levels, blood pressure, and other health indicators, to predict the likelihood of an individual having diabetes. This project aims to develop a predictive model for early detection of diabetes, assisting healthcare professionals in identifying individuals at risk and facilitating timely intervention and management strategies.

Objectives:

1. **Data Exploration:** Explore the dataset to understand the distributions and relationships between health indicators and diabetes.
2. **Decision Tree Classifier:** Implement a decision tree classifier to predict the likelihood of an individual having diabetes based on health indicators.
3. **Performance Evaluation:** Evaluate the performance of the decision tree classifier in predicting diabetes likelihood.
4. **Early Detection:** Develop a predictive model for early detection of diabetes.

Code File:

https://github.com/Harshan-Attar/Data-Science/tree/800edec29f1bafc0d66a39ffdb6968e675b8e89b/Projects/Project%204

Project 5

Title:

House Price Prediction: Multiple Linear Regression Analysis

Problem Statement:

Utilizing multiple linear regression on the provided housing dataset, which includes features such as area, number of bedrooms, bathrooms, etc., to predict the prices of houses. This project aims to develop a model that can accurately estimate housing prices based on various attributes, assisting both buyers and sellers in making informed decisions in the real estate market.

Objectives:

1. **Data Exploration:** Explore the housing dataset to understand the distributions and relationships between different features and house prices.
2. **Multiple Linear Regression**: Implement multiple linear regression to model the relationship between housing features and prices.
3. **Informed Decision-making:** Develop a model that can accurately estimate housing prices, assisting both buyers and sellers in making informed decisions in the real estate market.

Code File:

https://github.com/Harshan-Attar/Data-Science/tree/800edec29f1bafc0d66a39ffdb6968e675b8e89b/Projects/Project%205

Project 6

Title:

Flickr Image Scraper: Automating Image Collection

Problem Statement:

Developing a web scraping tool to extract images from Flickr, a popular image-sharing platform, for various purposes such as research, content creation, or personal use. This project aims to streamline the process of gathering images from Flickr, providing users with a convenient method to acquire diverse image datasets for their specific needs.

Objectives:

1. **Image Extraction**: Extract images from Flickr based on user-defined criteria (e.g., tags, categories).
2. **Image Dataset Creation:** Create diverse image datasets for various purposes such as research, content creation, or personal use.
3. **User Convenience:** Provide users with a convenient method to acquire images from Flickr for their specific needs.

Code File:

https://github.com/Harshan-Attar/Data-Science/tree/800edec29f1bafc0d66a39ffdb6968e675b8e89b/Projects/Project%206

Project 7

Title:

Indeed Job Data Extractor: Web Scraping Job Positions, Locations, and Company Names

Problem Statement:

This project focuses on developing a web scraping tool to extract job position titles, locations, and company names from Indeed, a popular job search platform. By automating the data extraction process, the tool aims to provide users with valuable insights into job market trends and facilitate job seekers in finding relevant job opportunities.

Objectives:

1. **Data Extraction:** Automate the extraction process to gather job market data efficiently.
2. **Insight Generation:** Provide users with valuable insights into job market trends based on the extracted data.
3. **Job Seeker Assistance:** Facilitate job seekers in finding relevant job opportunities by providing accurate and up-to-date information.

Code File:

https://github.com/Harshan-Attar/Data-Science/tree/800edec29f1bafc0d66a39ffdb6968e675b8e89b/Projects/Project%207

Project 8

Title:

Job Data Extraction from Naukri.com and Jobstreet.com: Web Scraping Job Positions, Locations, and Company Names.

Problem Statement:

This project aims to develop a web scraping tool to extract job position titles, locations, and company names from two prominent job search platforms, Naukri.com and Jobstreet.com. By automating the data extraction process from these platforms, the tool seeks to provide users with comprehensive insights into job market trends and facilitate efficient job searching for candidates and recruiters alike.

Objectives:

1. **Data Extraction:** Automate the extraction process to gather job market data efficiently from both platforms.
2. **Comprehensive Insights:** Provide users with comprehensive insights into job market trends by combining data from two prominent job search platforms.
3. **Efficient Job Searching:** Facilitate efficient job searching for candidates and recruiters by providing accurate and up-to-date information from multiple sources.

Code File:

https://github.com/Harshan-Attar/Data-Science/tree/800edec29f1bafc0d66a39ffdb6968e675b8e89b/Projects/Project%208

Project 9

Title:

Sentiment Analysis of Amazon Product Reviews using NLTK.

Problem Statement:

This project focuses on conducting sentiment analysis on Amazon product reviews using the Natural Language Toolkit (NLTK). By analyzing the sentiment expressed in the reviews from the provided dataset, the goal is to gain insights into customer opinions and sentiments towards the products, facilitating better understanding of customer satisfaction and preferences.

Objectives:

1. **Insight Generation:** Gain insights into customer opinions and sentiments towards the products.
2. **Customer Satisfaction:** Understand customer satisfaction and preferences based on the sentiment analysis of reviews.
3. **Product Improvement:** Identify areas for product improvement based on customer feedback and sentiment analysis.
4. **Decision Making:** Provide valuable information for decision-making regarding product offerings and marketing strategies.

Code File:

https://github.com/Harshan-Attar/Data-Science/tree/800edec29f1bafc0d66a39ffdb6968e675b8e89b/Projects/Project%209

Project 10

Title:

Possum Species Classification Using Random Forest Algorithm

Problem Statement:

This project aims to classify possum species based on physical characteristics using the Random Forest algorithm. By analyzing data on possum physical attributes and gender from the provided dataset, the goal is to develop a robust classification model that accurately identifies possum species. This classification model can aid researchers and conservationists in wildlife monitoring and management efforts.

Objectives:

1. Data Analysis: Analyze data on possum physical attributes and gender to understand the characteristics of different possum species.
2. Wildlife Monitoring: Provide a tool for researchers and conservationists to monitor possum populations and species distribution.
3. Conservation Efforts: Aid in wildlife management and conservation efforts by providing accurate species classification for possums.

Code File:

https://github.com/Harshan-Attar/Data-Science/tree/800edec29f1bafc0d66a39ffdb6968e675b8e89b/Projects/Project%2010

Project 11

Title:

Market Basket Analysis using Apriori Algorithm on Store Dataset.

Problem Statement:

This project focuses on conducting market basket analysis using the Apriori algorithm on a store dataset containing transaction data of various items such as vegetables, green tea, honey, grapes, shrimps, avocados, etc. By analyzing the frequent itemsets and association rules within the dataset, the goal is to uncover patterns of item co-occurrence and customer purchasing behavior. This analysis can provide valuable insights for optimizing product placement, cross-selling strategies, and promotional campaigns in the store.

Objectives:

1. **Pattern Discovery:** Uncover patterns of item co-occurrence and customer purchasing behavior to understand buying patterns.
2. **Insight Generation:** Generate insights for optimizing product placement, cross-selling strategies, and promotional campaigns in the store.
3. **Store Optimization:** Provide recommendations for store optimization based on the analysis of customer purchasing behavior

Code File:

https://github.com/Harshan-Attar/Data-Science/tree/800edec29f1bafc0d66a39ffdb6968e675b8e89b/Projects/Project%2011