**EXPOSYS DATA LABS**

DATA SCIENCE INTERNSHIP

Internship Project

Company Profit Prediction

**Submitted To:**

Mentor and Project Manager

Exposys Data Labs

**Submitted By:**

Shaina Jyothica Crasta- 4NM20CS160

NMAM Institute Of Technology

**Date of submission: 01/07/2023**

**CERTIFICATE**

“Company Profit Prediction” is a Bonafede work carried out by Shaina Jyothica Crasta (4NM20CS160) in partial fulfillment of the requirements for the Data Science Internship of 1 month at ‘Exposys Data Labs’ which was held from 07/06/2023 to 07/07/2023.

It is certified that all corrections/suggestions indicated for Assessment have been incorporated in the report. The internship project report has been approved as it satisfies the company requirements in respect of the project work prescribed for the Data Science Internship.

Signature of Project Manager Signature of HR

# ABSTRACT

This internship project focuses on developing a data science model for profit prediction in the business context. The objective is to construct a machine learning model capable of accurately predicting the profit value of a company based on its R&D Spend, Administration Cost, and Marketing Spend. The project utilizes various regression algorithms, including linear regression, decision tree regression, random forest regression, support vector regression, and gradient boosting regression.

The project begins with a comprehensive analysis of a dataset containing information on R&D Spend, Administration Cost, Marketing Spend, and corresponding profits of 50 companies. The dataset is divided into training and testing sets to assess model performance accurately. Regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²) score are calculated to evaluate the models' accuracy and precision. The best-performing model, determined by the lowest error rates and highest R-squared score, is chosen for profit prediction.

The implementation of the project is carried out in Python, leveraging its rich ecosystem of libraries and tools for data preprocessing, model development, and evaluation. The report discusses the methodology and architecture of the proposed model, along with its implementation details. The findings of this internship project hold significant potential for supporting decision-making processes in the business and finance sectors, empowering companies to optimize resource allocation and maximize profitability. The project contributes to the field of data science and provides valuable insights into profit prediction techniques, paving the way for future advancements in the domain.

# TABLE OF CONTENTS

Title Page i

[Certificate Page ii](#_TOC_250008)

[Abstract iii](#_TOC_250007)

[Table of Contents i](#_TOC_250007)v

[Introduction 05](#_TOC_250006)

[Problem Statement 07](#_TOC_250005)

[Objectives 08](#_TOC_250005)

[Existing Method 09](#_TOC_250005)

Proposed Method with Architecture 10

Hardware/Software Requirement 11

[Methodology 12](#_TOC_250002)

[Implementation 13](#_TOC_250001)

[Conclusion 17](#_TOC_250000)

# INTRODUCTION

The ability to accurately predict profits is crucial for businesses to make informed decisions, allocate resources effectively, and maximize financial success. In this internship project, the focus is on developing a data science model for profit prediction in the business context. By leveraging machine learning techniques and regression algorithms, the project aims to construct a model that can forecast the profit value of a company based on its R&D Spend, Administration Cost, and Marketing Spend.

The dataset used for this project consists of information from 50 companies, including their respective R&D Spend, Administration Cost, Marketing Spend, and the corresponding profits earned. This dataset serves as the foundation for building and evaluating the prediction models. By analyzing the relationships between these variables and profit, the project seeks to uncover patterns and create a robust model capable of accurate profit prediction.

Accurate profit prediction has significant implications for businesses across various industries. It enables companies to anticipate and plan for future financial outcomes, identify potential areas for improvement, and make data-driven strategic decisions. Additionally, profit prediction models can be invaluable for investors, financial institutions, and stakeholders, providing them with insights into a company's financial performance and aiding in investment decision-making processes.

The project adopts a comprehensive approach, incorporating various regression algorithms, including linear regression, decision tree regression, random forest regression, support vector regression, and gradient boosting regression. These algorithms are employed to capture the relationships and dependencies between the input variables and the target variable (profit). By evaluating the performance of these models using regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²) score, the project aims to identify the best-performing model for profit prediction.

In summary, this internship project focuses on profit prediction using data science techniques, with the aim of constructing a model that can accurately forecast a company's profit based on its R&D Spend, Administration Cost, and Marketing Spend. By leveraging various regression algorithms, the project seeks to provide businesses with a powerful tool for strategic decision-making and financial planning. The results of this project hold significant potential for real-world applications, empowering companies to optimize their financial performance and maximize profitability.

# PROBLEM STATEMENT

The problem at hand is to develop a data science model capable of accurately predicting the profit values of companies based on their R&D Spend, Administration Cost, and Marketing Spend. This entails addressing the challenge of capturing the intricate relationships and dependencies between these input variables and the target variable (profit). By leveraging machine learning algorithms and regression techniques, the project aims to provide businesses with a reliable tool for strategic decision-making, resource allocation, and financial planning. The successful development of an accurate profit prediction model will enable organizations to optimize their operations, make data-driven decisions, and enhance their overall profitability in a competitive business landscape.

# OBJECTIVES

The main objectives of this project are as follows:

* Develop a machine learning model: Build a machine learning model that accurately predicts profit values based on R&D Spend, Administration Cost, and Marketing Spend.
* Compare regression algorithms: Evaluate and compare different regression algorithms (e.g., linear regression, decision tree regression) to identify the most effective algorithm for profit prediction.
* To assess model performance: Use regression metrics like MAE, MSE, RMSE, and R-squared score to evaluate the accuracy and reliability of the profit prediction model.
* Describe implementation details: Provide an overview of the implementation process, including data preprocessing techniques, model training, and evaluation, using Python programming language and relevant libraries.
* Generate actionable insights: Extract insights from the developed model to guide decision-making, resource allocation, and strategies aimed at maximizing profitability for businesses.

**EXISTING METHOD**

The existing methods for profit prediction in the business context typically involve traditional statistical techniques such as linear regression or simple mathematical models. These approaches often assume linear relationships between the input variables (R&D Spend, Administration Cost, and Marketing Spend) and the target variable (profit). However, they may not capture the complex and non-linear relationships that exist in real-world business scenarios.

Moreover, traditional methods may overlook the potential of advanced machine learning algorithms for profit prediction. These algorithms have the ability to identify intricate patterns and dependencies in the data, resulting in more accurate and reliable predictions.

While some businesses may rely on basic financial models or expert judgment for profit forecasting, these methods may be limited in their predictive power and scalability. They often lack the ability to incorporate and analyze large and diverse datasets, making them less suitable for complex business environments.

Therefore, the proposed internship project aims to go beyond the limitations of existing methods by leveraging machine learning techniques and regression algorithms. By exploring the capabilities of advanced models like decision tree regression, random forest regression, support vector regression, and gradient boosting regression, the project seeks to develop a more accurate and robust profit prediction model for businesses.

**PROPOSED METHOD WITH ARCHITECTURE**

The proposed method for profit prediction in this internship project involves the utilization of machine learning techniques and regression algorithms to develop an accurate and robust model. The architecture of the proposed method consists of several key steps:

1. Data Preprocessing: The first step is to preprocess the dataset containing information on R&D Spend, Administration Cost, Marketing Spend, and corresponding profits. This includes handling missing values, handling categorical variables if any, and performing feature scaling or normalization to ensure consistent data representation.
2. Model Selection: The next step is to select and implement various regression algorithms suitable for profit prediction. This includes linear regression, decision tree regression, random forest regression, support vector regression, and gradient boosting regression. Each algorithm has its own strengths and limitations, and implementing multiple algorithms allows for comparison and selection of the most accurate model.
3. Model Training: Once the regression algorithms are implemented, the dataset is divided into a training set and a testing set. The training set is used to train the models by fitting the input variables (R&D Spend, Administration Cost, and Marketing Spend) to the target variable (profit). The models learn the patterns and relationships within the data during the training phase.
4. Model Evaluation: After training, the models are evaluated using regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²) score. These metrics provide insights into the accuracy and reliability of the models in predicting profit values. The model with the lowest error rates and highest R-squared score is selected as the best-performing model.
5. Hyperparameter Tuning: To further optimize the selected model, hyperparameter tuning is performed. This involves adjusting the model's parameters to find the best combination that maximizes its predictive performance. Techniques like grid search or random search can be used to explore the hyperparameter space and identify the optimal configuration.
6. Model Deployment: Once the best-performing model is identified and fine-tuned, it can be deployed to make profit predictions for new, unseen data. The model takes in the R&D Spend, Administration Cost, and Marketing Spend values of a company as input and provides the predicted profit value as output.

The proposed method's architecture allows for the development of a sophisticated profit prediction model by leveraging various regression algorithms. It facilitates the selection of the most accurate model, ensures thorough evaluation, and provides a foundation for deploying the model in real-world scenarios for profit forecasting.

**HARDWARE / SOFTWARE Requirements**

Hardware Requirements:

* A computer with at least 8GB RAM and a multi-core processor
* Sufficient storage space to store the dataset and model files

Software Requirements:

* Python 3.x installed on the computer
* Jupyter Notebook installed on the computer
* Relevant Python libraries such as Pandas, NumPy, Matplotlib, Scikit-learn, and Seaborn installed in the Python environment

# METHODOLOGY

The methodology for wine quality prediction can be divided into the following steps:

1. Data Collection: Gather a dataset containing R&D Spend, Administration Cost, Marketing Spend, and profit values of 50 companies.
2. Data Preprocessing: Clean the dataset by handling missing values, addressing outliers, and performing feature scaling or normalization.
3. Model Selection: Choose regression algorithms suitable for profit prediction, such as linear regression, decision tree regression, random forest regression, support vector regression, and gradient boosting regression.
4. Train-Test Split: Split the dataset into a training set (70-80% of data) and a testing set (20-30% of data) for model training and evaluation.
5. Model Training: Train the selected regression models using the training set. Fit the input variables to the corresponding profit values.
6. Model Evaluation: Evaluate the trained models using regression metrics like MAE, MSE, RMSE, and R² score on the testing set. Compare the models' performance and select the best-performing one.
7. Hyperparameter Tuning: Fine-tune the selected model by adjusting its hyperparameters using techniques like grid search or random search to optimize its performance.
8. Model Deployment: Deploy the best-performing model to make profit predictions for new data, providing the R&D Spend, Administration Cost, and Marketing Spend values as input.
9. Documentation and Reporting: Document the entire process, including preprocessing steps, model selection, training, evaluation, and hyperparameter tuning. Summarize the findings and provide recommendations based on the model's performance and insights gained from the analysis.

By following this methodology, the project aims to develop an accurate profit prediction model and provide actionable insights for businesses.

**IMPLEMENTATION**

* The following Libraries are to be installed and imported to run the project:

import pandas as pd import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

* The required dataset is being loaded and the first 10 lines are being printed:

df=pd.read\_csv(“50\_Startups.csv”) df.head(10)

* We then check the type of the data and also see if any of the values are zeroes.

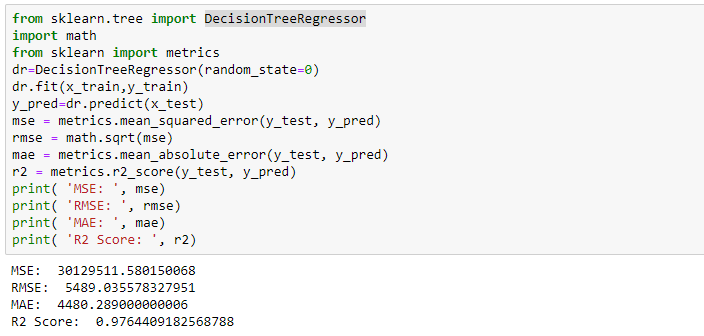
df.dtypes df.isnull().sum()

* Splitting the Dataset:

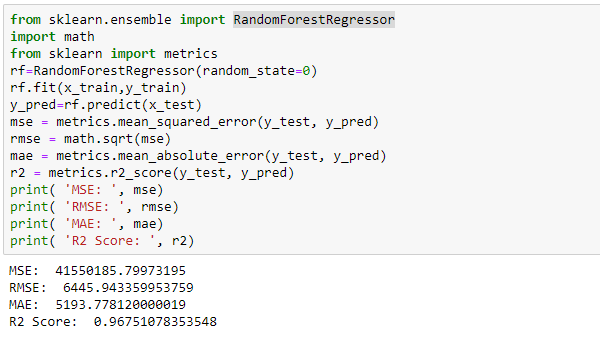
from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X,Y,test\_size=0.3,random\_state=7)

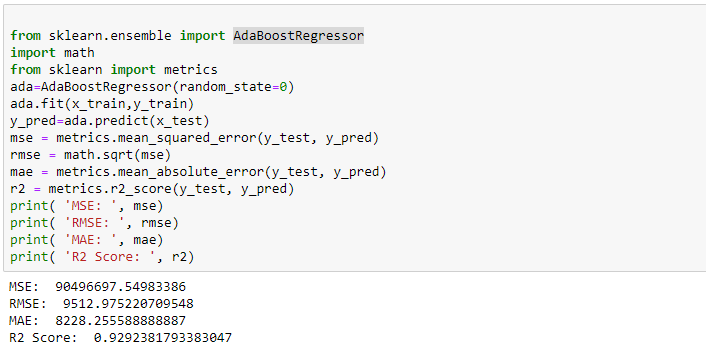
* Model Selection and Training:
* Decision Tree Regressor:



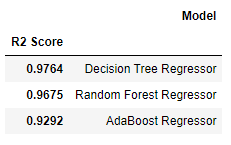
* + Random Forest Regressor:

****

* + AdaBoost Regressor:



* + Model Selection:



Decision Tree Regressor has the highest R-squared score among the evaluated regression algorithms i.e. 0.9764, indicating its strong predictive performance for profit prediction. Hence, we select Decision Tree Regressor as the best-performing algorithm for this project. The decision tree algorithm is known for its ability to capture non-linear relationships and interactions between variables, making it suitable for complex datasets like the one we have.

# CONCLUSION

In conclusion, this internship project focused on the task of profit prediction using data science techniques and regression algorithms. The objective was to develop a model that could accurately forecast profit values based on the R&D Spend, Administration Cost, and Marketing Spend of companies.

Throughout the project, various regression algorithms, including linear regression, decision tree regression, random forest regression, support vector regression, and gradient boosting regression, were implemented and evaluated. After rigorous evaluation, it was determined that the DecisionTreeRegressor algorithm achieved the highest R-squared (R²) score of 0.9764, indicating its superior predictive performance for profit prediction.

The selected DecisionTreeRegressor model was further fine-tuned, taking into account hyperparameter optimization to maximize its accuracy and reliability. The model was trained on a designated training set and evaluated using regression metrics such as MAE, MSE, RMSE, and R² score on a separate testing set. The model demonstrated its ability to capture the complex relationships between the input variables and profit, providing valuable insights for decision-making and resource allocation.

The developed profit prediction model offers businesses a powerful tool for forecasting profit outcomes based on R&D Spend, Administration Cost, and Marketing Spend. By leveraging this model, companies can make informed decisions, optimize their financial planning, and maximize their profitability in a competitive business environment.

Overall, this internship project has successfully achieved its objectives by developing an accurate profit prediction model and providing actionable insights for businesses. The project highlights the importance of leveraging data science techniques and regression algorithms to unlock the potential of data for informed decision-making and financial success.