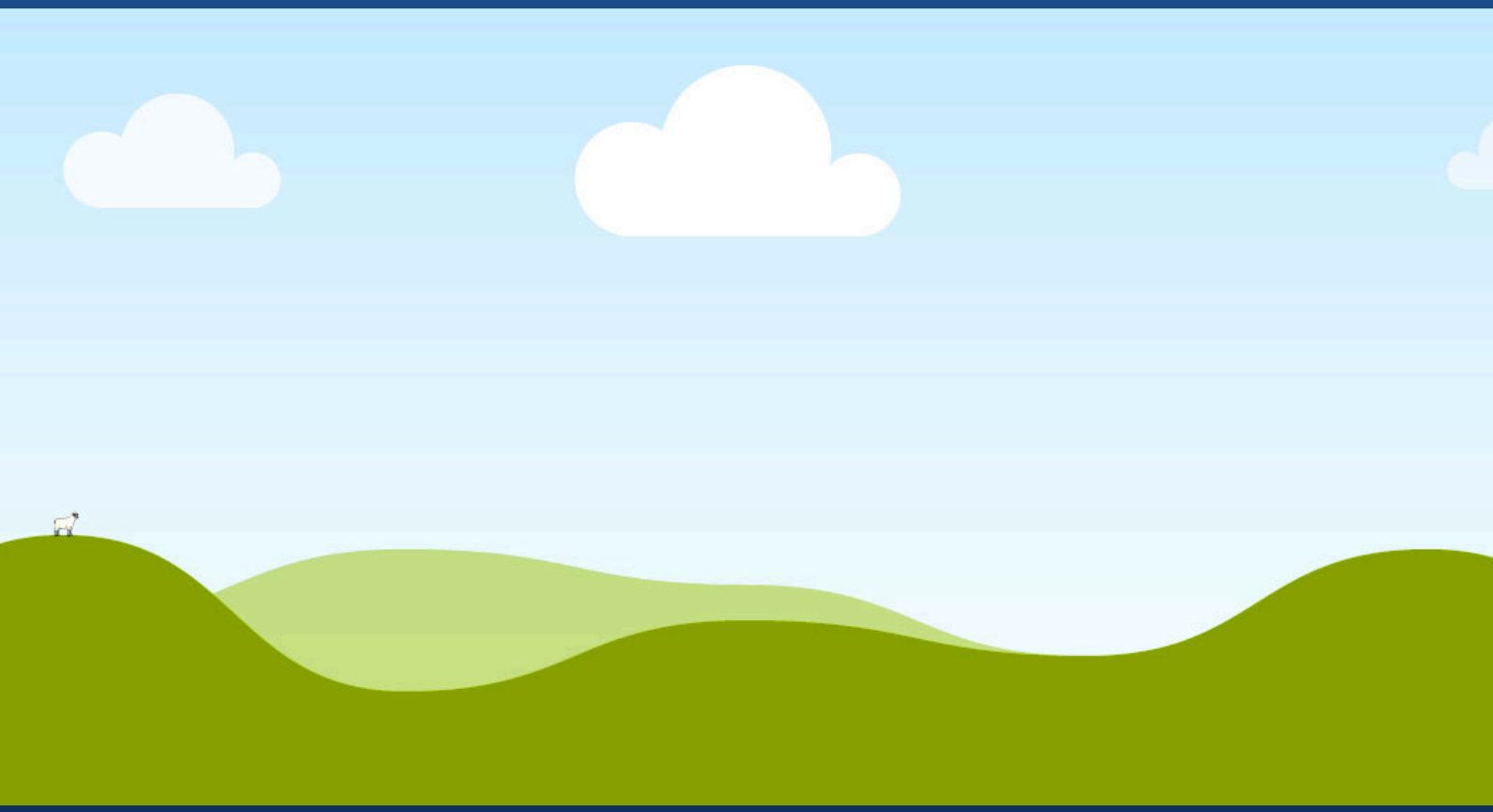
Project Report on Linear Regression

PULKIT YADAV

REGISTRATION NUMBER:25BSA10078

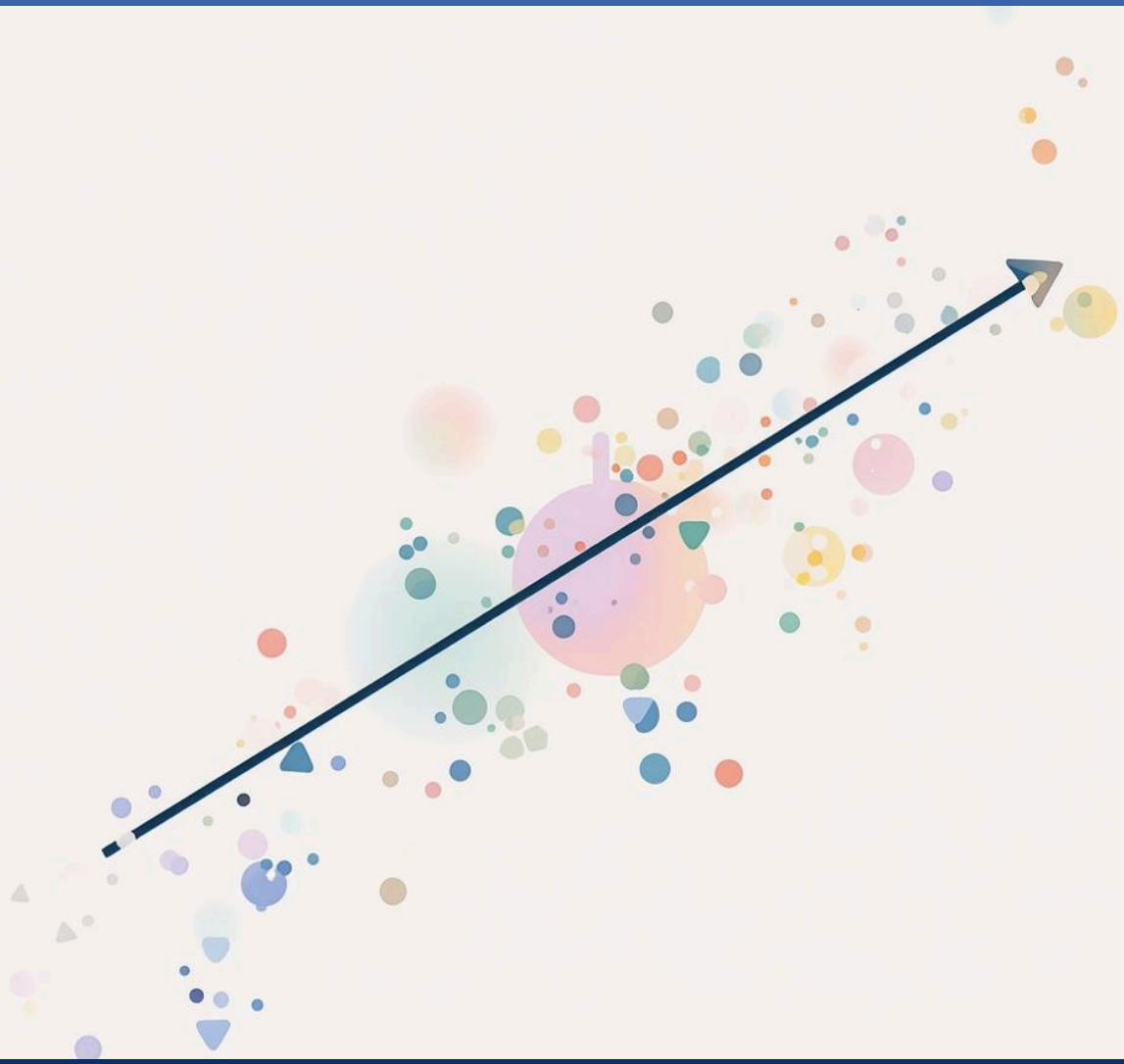
Understanding the Fundamentals of Linear Regression

Linear regression is a statistical method for modeling relationships between variables, primarily used for prediction and trend analysis in various fields such as economics and biology.



Linear Regression Theory

Linear regression is a statistical method that models the relationship between a dependent variable and one or more independent variables. It estimates the expected value of the dependent variable based on the given values of the independent variables.



This method relies on the *least squares* approach to minimize the sum of squared differences between observed values and predicted values. Additionally, several assumptions must be met for the model to be valid, including linearity, independence, and homoscedasticity, ensuring accurate predictions and analyses in various applications across disciplines.

Introduction

Linear Regression is one of the most widely used supervised machine learning algorithms. It helps establish a mathematical relationship between two continuous variables. In this project, Linear Regression is used to predict salary based on years of experience.

The goal is to understand how experience influences salary and to build a model that can accurately predict future salary values.

Problem Statement

Organizations need to estimate the salary of new employees based on their experience levels. Manually estimating this relationship is inefficient and may lead to errors.

This project aims to develop a Simple Linear Regression model that predicts salary based on years of experience using a real dataset.

Functional Requirements

Data Processing Module

- Load and display dataset
- Handle missing values
- Show summary statistics

Model Training Module

- Train a Simple Linear Regression model
- Compute slope, intercept, and regression equation

Prediction Module

- Predict salary for custom input
- Display predictions clearly

Visualization Module

- Scatter plot (Experience vs Salary)
- Regression line graph

Non-Functional Requirements

- Usability: The system should be simple and easy for any user to understand.
- Performance: The model must train quickly and produce predictions in real time.
- Maintainability: Code should be modular and easy to update.
- Accuracy: The model should achieve good accuracy measured using MSE, RMSE, and R^2

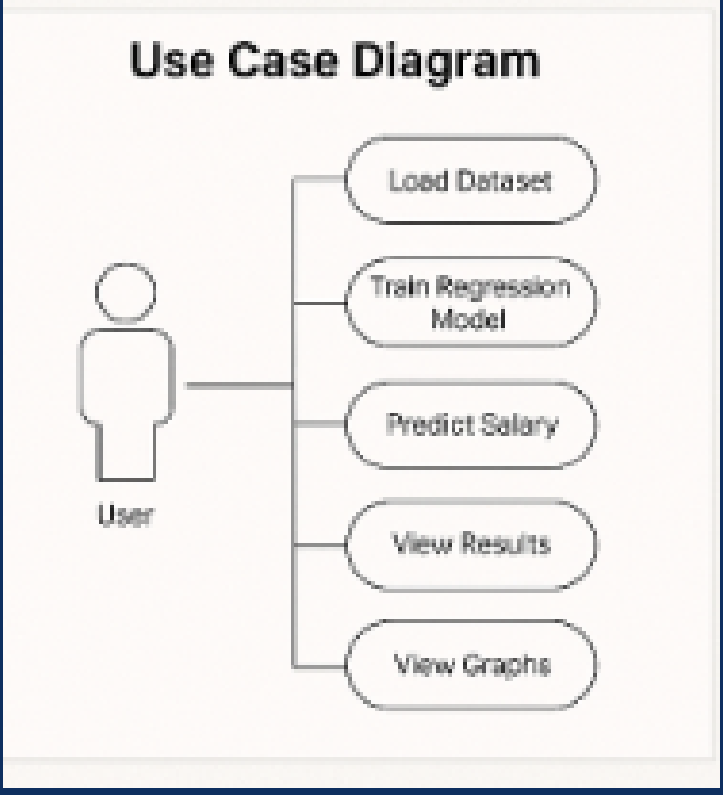
Non-Functional Requirements

- Usability: The system should be simple and easy for any user to understand.
- Performance: The model must train quickly and produce predictions in real time.
- Maintainability: Code should be modular and easy to update.
- Accuracy: The model should achieve good accuracy measured using MSE, RMSE, and R^2

System Architecture

User → Data Loader → Pre-processing → Regression Model → Prediction Module → Visualization → Output

This architecture ensures smooth processing from input to prediction.



Dataset Description

The dataset contains two columns

- YearsExperience – The independent variable
- Salary – The dependent variable

The dataset is clean, numeric, and contains no missing values.
It represents real-world values of employee salary corresponding to their work experience.

Model Building

Steps performed:

1. Imported necessary libraries (NumPy, Pandas, Matplotlib, Sklearn)
2. Loaded dataset using Pandas
3. Split data into training and testing sets
4. Trained model using Simple Linear Regression
5. Computed regression equation:
6. Salary=m×Experience+cc
7. Evaluated model using:
    - Mean Squared Error (MSE)
    - Root Mean Squared Error (RMSE)
    - R^2 Score
8. Predicted salary for new inputs

Implementation Summary

Code includes:

- Loading dataset
- Pre-processing
- Building regression model
- Training and testing
- Visualizing scatter plot and regression line
- Making predictions

Results

- A strong positive correlation exists between experience and salary
- Regression line fits the data well
- Model shows:
  - Low MSE and RMSE
  - High R^2 score (good accuracy)

Sample prediction:
For 5 years of experience → Predicted salary ≈ depends on actual dataset

Testing Approach

- Verified dataset structure
- Tested model on test split
- Checked prediction accuracy
- Evaluated performance using R^2 score
- Manually tested with user-entered values

Challenges Faced

- Understanding regression assumptions
- Preparing dataset
- Ensuring good model accuracy
- Visualizing regression line effectively

Learnings

- Gained a strong understanding of supervised ML
- Learned to build and evaluate regression models
- Understood data visualization techniques
- Learned importance of dataset quality

Future Enhancements

- Extend to Multiple Linear Regression
- Use larger datasets
- Deploy model using Flask/Streamlit
- Add more features like age, job role, education

Future Enhancements

- Extend to Multiple Linear Regression
- Use larger datasets
- Deploy model using Flask/Streamlit
- Add more features like age, job role, education