EXPERIMENT 1

Aim of the Experiment

To study the concept of regression and to implement Linear Regression and Logistic Regression models on a real-world Resume Matching Dataset for prediction and classification purposes.

Theory

What is Regression?

Regression is a supervised machine learning technique used to predict output values based on input features.

There are two main types used in this experiment:

- Linear Regression → Predicts continuous values
- Logistic Regression → Predicts binary outcomes (0 or 1)

Introduction to Regression

Regression is a type of supervised machine learning technique where a model is trained using labeled data to predict output values from given input features. It establishes a relationship between independent variables and a dependent variable.

Based on the nature of the output, regression techniques are broadly classified into:

- Linear Regression, which predicts continuous numerical values
- Logistic Regression, which predicts categorical or binary outcomes

---

Linear Regression

Overview

Linear Regression attempts to model the relationship between input variables and a continuous output by fitting a straight line that best represents the data points.

In this experiment, Linear Regression is used to predict the similarity score between resume skills and job required skills.

Mathematical Representation

For multiple input variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Where:

- y represents the predicted output
- x represents the input features
- β represents the learned weights
- β_0 represents the intercept

Logistic Regression

Overview

Logistic Regression is a classification algorithm used when the output variable has only two possible outcomes.

Instead of predicting exact values, it predicts the probability of a data point belonging to a specific class.

In this experiment, Logistic Regression is used to classify candidates as:

- High Match (1)
- Low Match (0)

Sigmoid Function

The predicted values are passed through the sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

The output lies between 0 and 1 and is interpreted as probability.

Dataset Description

Dataset Source

- Dataset Name: Resume Matching Dataset
- File Name: resume_data.csv
- Domain: Recruitment and HR Analytics
- Data Format: Structured and tabular
- Approximate Records: ~9500 resumes

Attributes Used

Independent Variables

- skills – Candidate resume skills
- skills_required – Skills required for job position

These text features were converted into numerical format using TF-IDF vectorization.

Dependent Variable (For Linear Regression)

- matched_score – Numerical similarity score between resume and job description

Since matched_score is continuous, it is suitable for Linear Regression.

Dependent Variable (For Logistic Regression)

matched_score was converted into binary class:

- Score ≥ 0.5 → High Match (1)
- Score < 0.5 → Low Match (0)

Linear Regression Model Formulation

For multiple predictors, Linear Regression is expressed as:

$$Matched\_Score = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

Where:

- β_0 is the intercept
- $\beta_1, \beta_2, \ldots \beta_n$ are regression coefficients

---

Limitations of Linear Regression

- Works only when the relationship is approximately linear
- Sensitive to outliers
- Cannot capture complex non-linear patterns
- Requires numerical input features

Experimental Procedure

1. The resume dataset was loaded into Google Colab using Pandas.
2. Relevant columns (skills, skills_required, matched_score) were selected.
3. Missing values were removed to ensure data consistency.
4. Text features were combined and converted into numerical features using TF-IDF Vectorization.
5. The dataset was divided into input features and output labels.
6. Data was split into training and testing sets in an 80:20 ratio.
7. Linear Regression was trained to predict matched_score.

8. Logistic Regression was trained to classify candidates as High Match or Low Match.

9. Model predictions were evaluated using suitable performance metrics.

10. Hyperparameter tuning was performed using GridSearchCV.

Performance Evaluation

Linear Regression Metrics

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score

A high R^2 value close to 1 indicates strong predictive capability of the model.

Logistic Regression Metrics

- Accuracy
- Confusion Matrix
- Precision
- Recall
- F1-Score

Hyperparameter Optimization

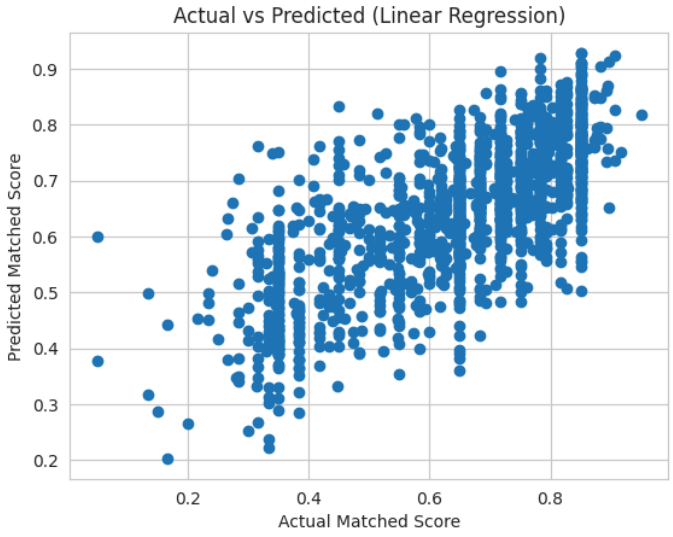
For Logistic Regression, GridSearchCV was used to:

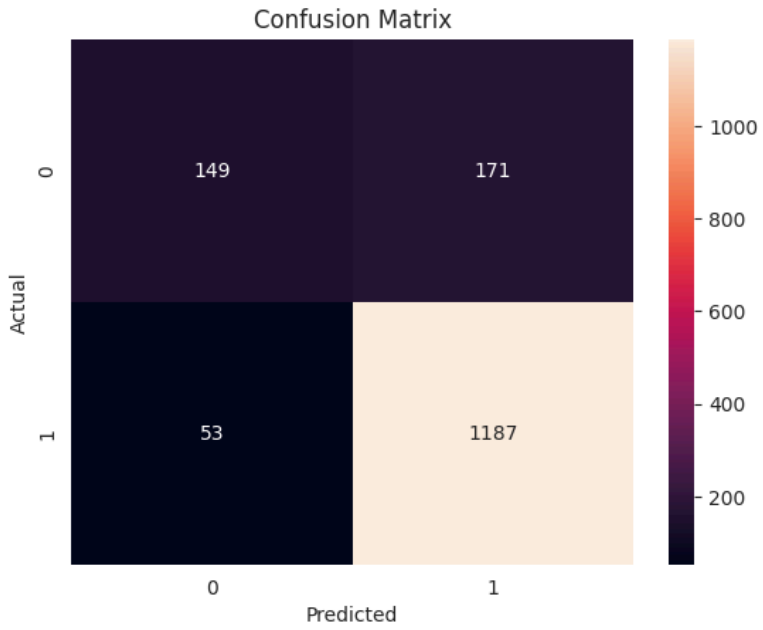
- Test multiple values of regularization parameter C
- Experiment with different solver algorithms
- Select the best combination using cross-validation

Tools and Libraries Used

- NumPy – Numerical operations
- Pandas – Data handling and preprocessing
- Matplotlib – Data visualization
- Seaborn – Statistical visualization
- Scikit-learn – Model training, evaluation, and tuning

Result:





The Linear Regression model successfully predicted resume-job similarity scores with satisfactory accuracy based on MAE, MSE, RMSE, and R^2 score.

The Logistic Regression model effectively classified candidates into High Match and Low Match categories with good accuracy. The confusion matrix showed proper classification of most candidates. Hyperparameter tuning further improved the classification performance.

Conclusion / Observations

Comparison: Linear Regression vs Logistic Regression

Aspect	**Linear Regression**	**Logistic Regression**
Type of Problem	Regression	Classification
Output	Continuous values	Binary values (0/1)
Target Variable	matched_score	High_Match
Model Function	Linear equation	Sigmoid function
Evaluation Metrics	MSE, RMSE, R^2 Score	Accuracy, Confusion Matrix
Use Case in Experiment	Predict similarity score	Classify candidate suitability

---

In this experiment, Linear Regression and Logistic Regression were implemented on the Resume Matching Dataset to analyze candidate-job similarity.

Linear Regression was used to predict the similarity score as a continuous value and showed satisfactory performance.

Logistic Regression was used for classification and successfully identified high-match candidates.

This experiment demonstrates that Linear Regression is suitable for continuous prediction tasks, while Logistic Regression is effective for classification problems in recruitment analytics.