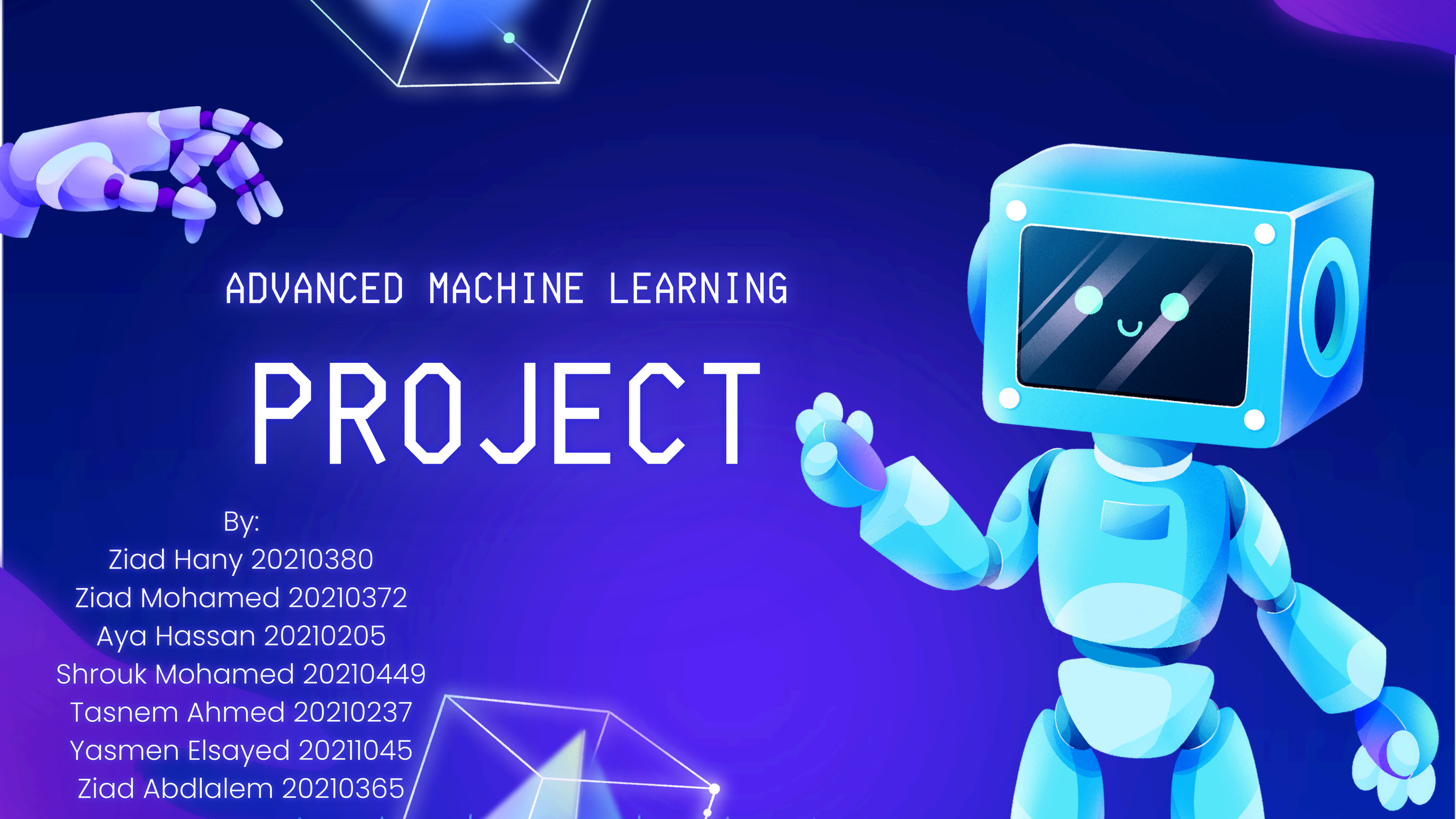


ADVANCED MACHINE LEARNING

PROJECT

By:

Ziad Hany 20210380

Ziad Mohamed 20210372

Aya Hassan 20210205

Shrouk Mohamed 20210449

Tasnem Ahmed 20210237

Yasmen Elsayed 20211045

Ziad Abdlalem 20210365

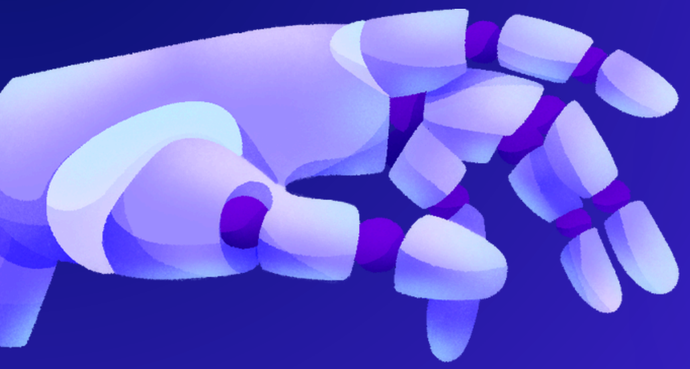
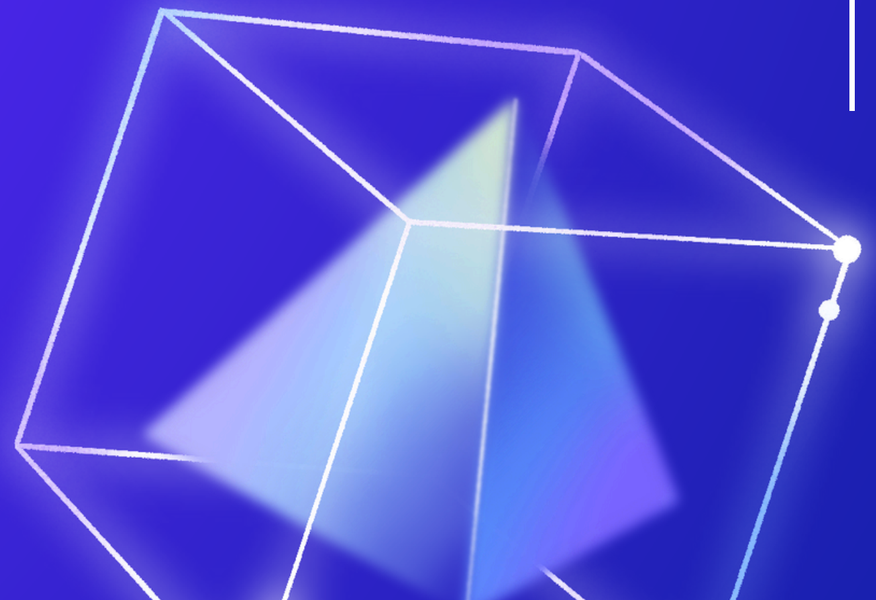


TABLE OF CONTENTS



INTRODUCTION

In today's world, data is abundant and holds immense potential to drive meaningful insights and predictions. Machine learning, a subset of artificial intelligence, plays a crucial role in harnessing this potential. In this project, we aim to apply machine learning techniques to two distinct datasets, utilizing both classification and regression algorithms to extract valuable information.

DATASETS OVERVIEW

FIRST DATASET:

Stroke Prediction Dataset

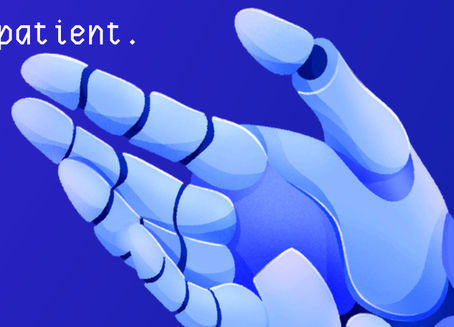
According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally,

responsible for approximately 11% of total deaths.

This dataset is used to predict whether a patient is likely to get stroke based on the

input parameters like gender, age, various diseases, and smoking status.

Each row in the data provides relavant information about the patient.



ABOUT STROKE DATASET:

we used it in classification

in this dataset their is (2) of classes
and their labels :Stroke , NO Stroke

number of samples used in : 5109

- split of samples :
- training:70% of the dataset(3576 samples)
- testing:30% of the dataset (1533 samples)

APPLIED MODELS ON IT:

01

- Artificial neural network (ANN) :

in this project first we trained the ann with 100 epochs

_ Before scaling number of features :6

_After scaling number of features : 6

Features Names:[age, hypertension, heart_disease, avg_glucose_level, bmi, stroke]

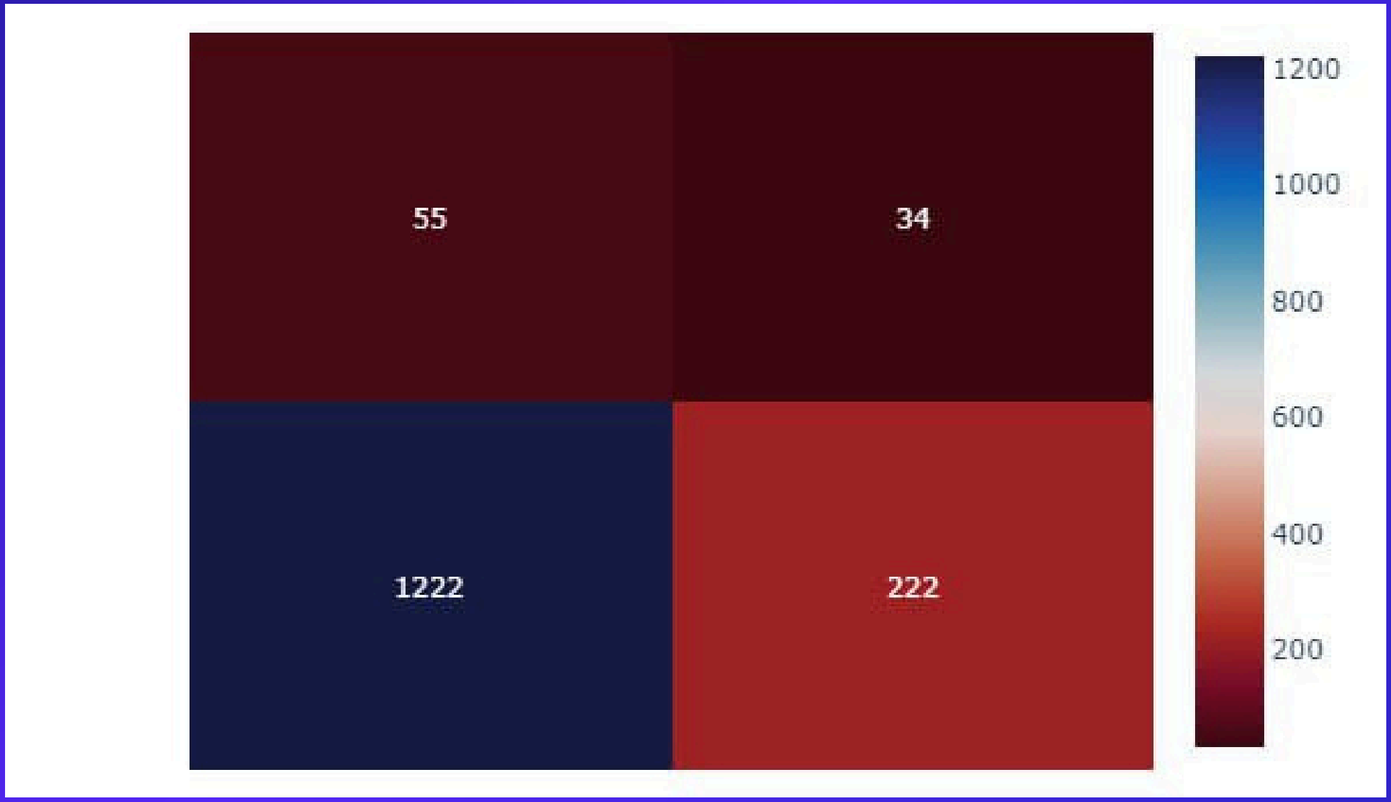
ANN model results

-Prediction accuracy: 82

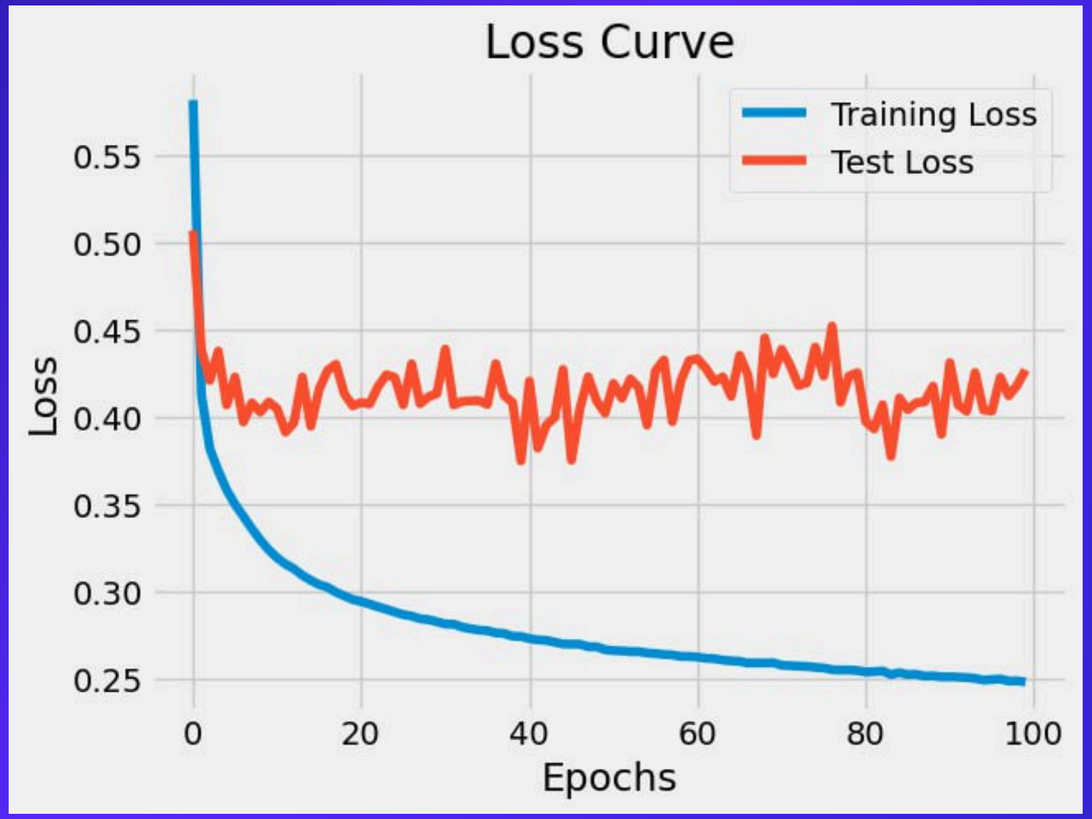
-Classification report diagram

	precision	recall	f1-score	support
0	0.96	0.84	0.89	1444
1	0.12	0.37	0.19	89
accuracy			0.81	1533
macro avg	0.54	0.61	0.54	1533
weighted avg	0.91	0.81	0.85	1533

Plot confusion matrix



Plot Loss curve



-Summary

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 16)	176
dense_1 (Dense)	(None, 8)	136
dense_2 (Dense)	(None, 1)	9

Total params: 321 (1.25 KB)

Trainable params: 321 (1.25 KB)

Non-trainable params: 0 (0.00 B)

Hyperparameters

- we add three layers :
- The first layer was added with 16 units and ReLU activation function
- The second layer was added with 8 units and ReLU activation function
- The last layer was added with 1 unit and sigmoid activation function for binary classification

APPLIED MODELS ON IT:

02

- Decision Tree Classifier: :

_ Before scaling number of features :6
_After scaling number of features : 6
Features Names :[age, hypertension, heart_disease, avg_glucose_level, bmi, stroke]

-Deicon tree report before graid search

```
Accuracy on Train set 1.0
Accuracy on Test set 0.8754076973255055
F1-score on Test set: 0.20746887966804978

              precision    recall  f1-score   support

           0       0.95      0.91      0.93      1444
           1       0.16      0.28      0.21        89

    accuracy                           0.88      1533
   macro avg       0.56      0.60      0.57      1533
weighted avg       0.91      0.88      0.89      1533
```

-Deicon tree report after graid search

```
Accuracy on Train set 0.8733899297423887
Accuracy on Test set 0.7958251793868232
F1-score on Test set: 0.2155388471177945

              precision    recall  f1-score   support

           0       0.96      0.82      0.88      1444
           1       0.14      0.48      0.22        89

    accuracy                           0.80      1533
   macro avg       0.55      0.65      0.55      1533
weighted avg       0.91      0.80      0.84      1533
```

Hyperparameters

- Max depth:10
- Max features:6
- Min samples split:90
- Min samples leaf:20

These hyperparameters are used in the DecisionTreeClassifier() initialization

SECOND DATASET:

Salary Dataset

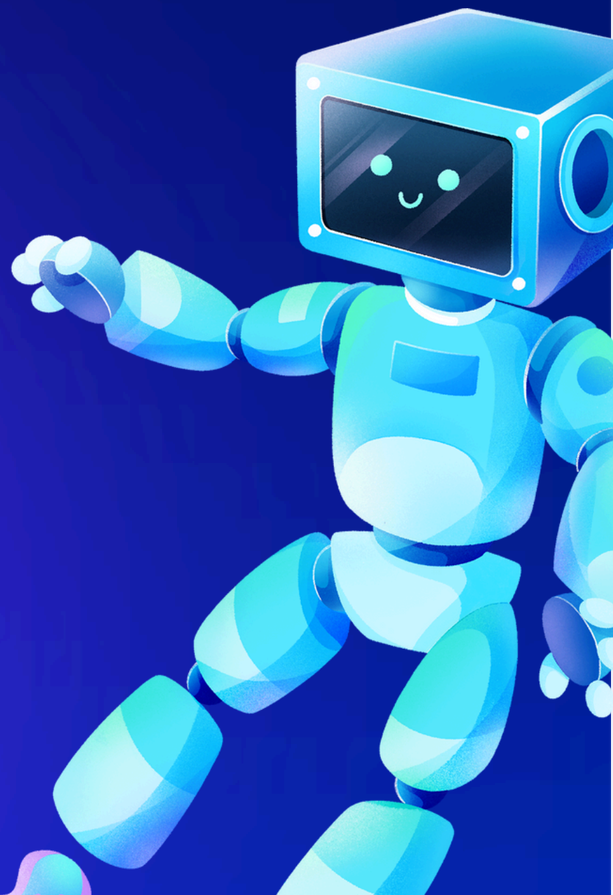
This dataset provides a comprehensive collection of salary information from various industries and regions across the globe. Sourced from reputable employment websites and surveys, it includes details on job titles, salaries, job sectors, geographic locations, and more. Analyze this data to gain insights into job market trends, compare compensation across different professions, and make informed decisions about your career or hiring strategies. The dataset is cleaned and preprocessed for ease of analysis and is available under an open license for research and data analysis purposes.

ABOUT SALARY DATASET:

we used it in regression

number of samples used in : 2758

- split of samples :
- training:80% of the dataset(2206 samples)
- testing:20% of the dataset (552 samples)



APPLIED MODELS ON IT:

01

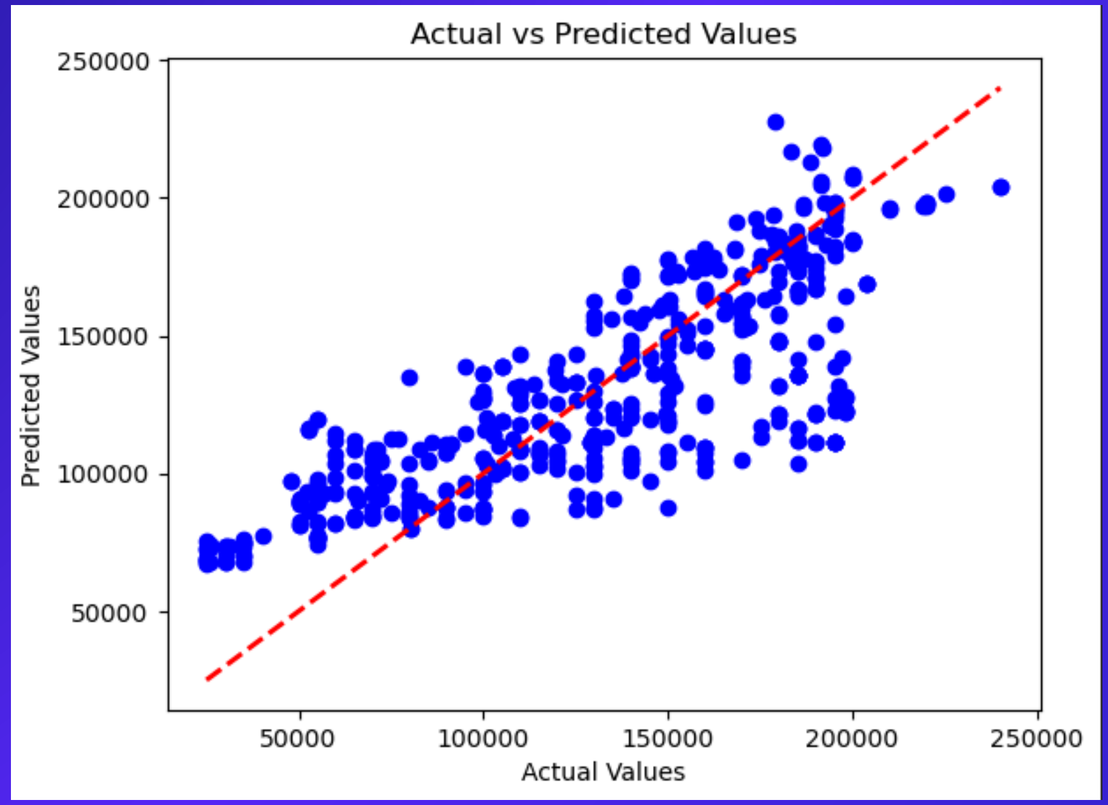
- Support Vector Machine (SVM) :

- Number of features : 7
- Features Names:[age, country, jop_title, education_level, gender, senior, years_of_experience]

Hyperparameters

- Kernal: linear
- Regulization parameter(c):100.0
- Kernal coefficient (gamma):0.01

Actual vs Predicted Values



svm results

- Accuracy with poly kernal:0.7533956904543017
- Accuracy with sigmoid kernal:0.6494171237383519

mean errors

- Mean Squared Error (MSE): 1722190897.9388
- Mean Absolute Error (MAE): 34370.4980
- R-squared (Coefficient of Determination): 0.3313

THANK YOU!

RESOURCE PAGE

Stroke Prediction Dataset

11 clinical features for predicting stroke events

k kaggle.com

Salary by Job Title and Country

A Cleaned Data For predicting income by job title and country.

k kaggle.com