Day 1 :Getting started with Machine Learning 101

Google Developer Groups

On Campus • Techno Main Salt Lake

Speakers

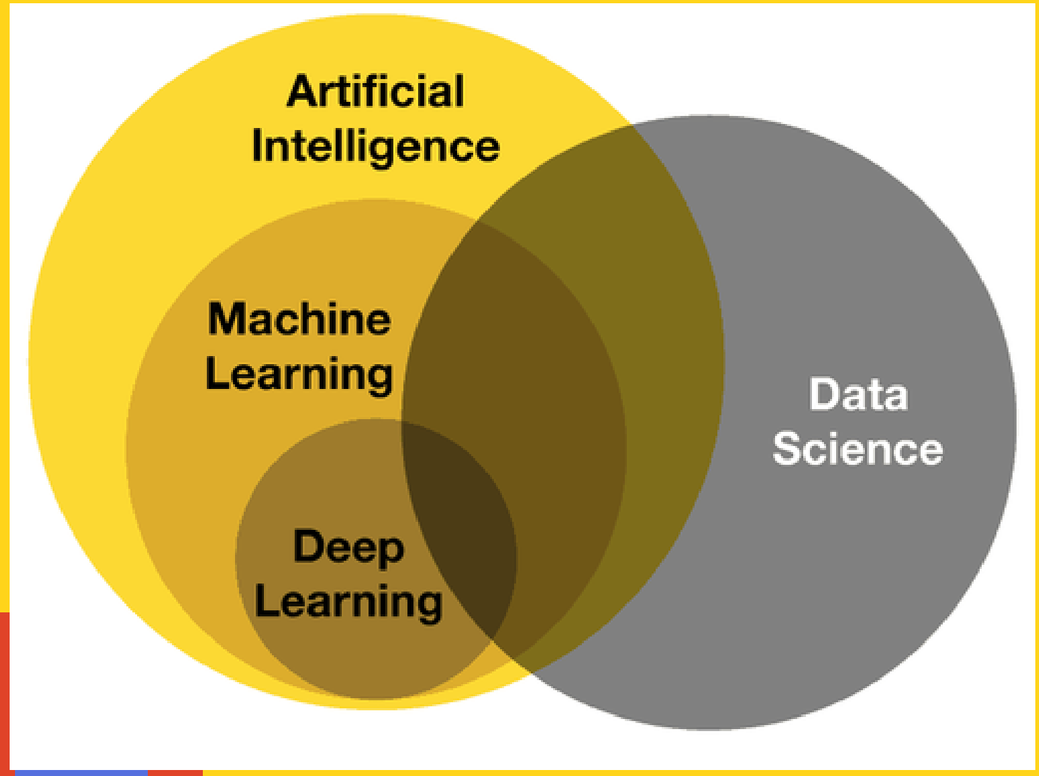
Rishi Bhattasali

Linkedin

Ayush Agarwal

Linkedin

What is Machine Learning?



Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" with data, without being **explicitly programmed**

What can you build with Machine Learning?

- Spam Ham Classifier
- Image Classification
- Sentiment Analysis
- Recommendation Systems
- Speech Recognition
- Natural Language Processing (NLP)

Agenda

01

Types of ML

02

Key Concepts and Terminology

03

Machine Learning Workflow

04

Overview of Algorithms

01 Types of ML

- **Supervised Learning -** A model is trained using data with known outcomes to learn how features relate to labels for making predictions.
  Eg - Regression, Classification
- **Unsupervised Learning -** A model is trained on data without labels to find patterns or groupings within the data**.** Eg -Clustering , Anamoly Detection.
- **Semisupervised Learning-**Semi-supervised learning is a machine learning approach that uses a small amount of labeled data along with a large amount of unlabeled data to improve model performance.
- **Reinforcement Learning -**Reinforcement learning trains a model to make decisions by rewarding good actions and penalizing bad ones.

02

Key Concepts and Terminologies

- **Features**: These are the input variables (independent variables)
- **Labels:** These are the output variables (dependent variable)

Training Data: The data used to train the model and learn patterns.
Testing Data: The data used to evaluate the performance of the model after training.

Overfitting: This results in poor performance on new, unseen data.
Underfitting: This results in poor performance on even testing data.

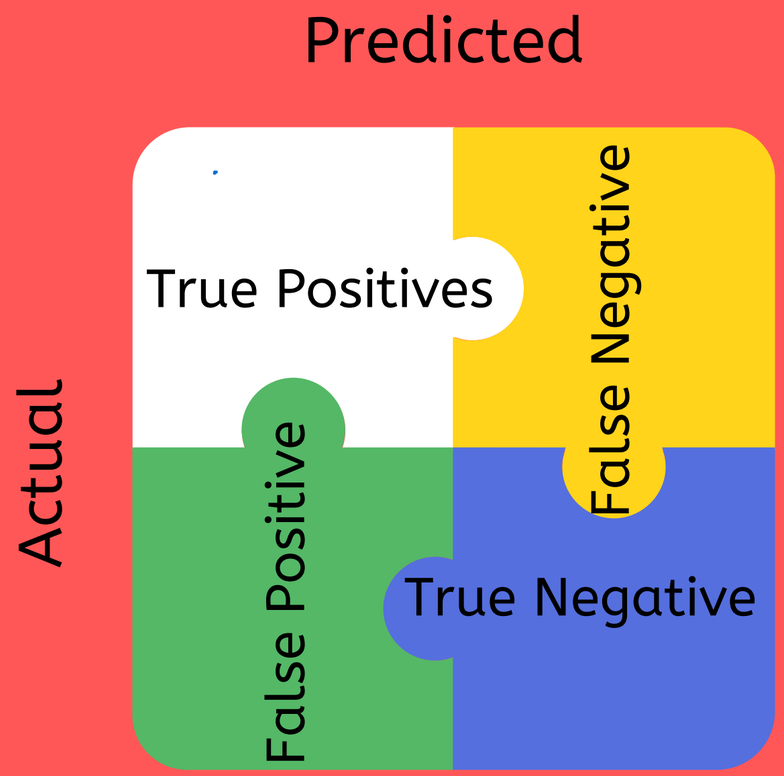
Model Evaluation Metrics

Accuracy: *The percentage of correct predictions made by the model.*

$$\frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Recall: *The ratio of true positive predictions to the total actual positives.*

$$\frac{TP}{(TP + FN)}$$



Precision: *The ratio of true positive predictions to the total predicted positives.*

$$\frac{TP}{(TP + FP)}$$

F1-Score: *The harmonic mean of precision and recall.*

$$\frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

Usecase 1 ⇒ Spam classification

Mail → Predict → Spam / Not a Spam

	1	0
1	TP	FP
0	FN	TN

FP↓↓↓

Mail → Spam
Model → Spam
} Good scenario

0 ⇐ Mail → Not a Spam
1 ⇐ Model → Spam
} ⇐ Blunder (FP is Important)

1 ⇐ Mail → Spam
0 ⇐ Model → Not a Spam
} ⇒ FN

⇓

PRECISION PERFORMANCE

Usecase 2 ⇒ FN is Important

To predict whether a person has diabetes or not

	Diabetes	No Diabetes → Actual
Diabetes	TP	FP
No Diabetes	FN↓↓↓	TN

1. Actual → Diabetes
   Model → Diabetes
   } Good

2. Actual → Diabetes
   Model → No. Diabetes
   } Blunder ⇒ FN↓↓↓ ⇒ Important

3. Actual → No Diabetes
   Model → Diabetes
   } FP ⇒ Wrong Prediction

4. Actual → No Diabetes
   Model → No Diabetes
   } Correct

Machine Learning Workflow

Data Collection and Preprocessing

Feature Selection and Engineering

Model Selection

Training the Model

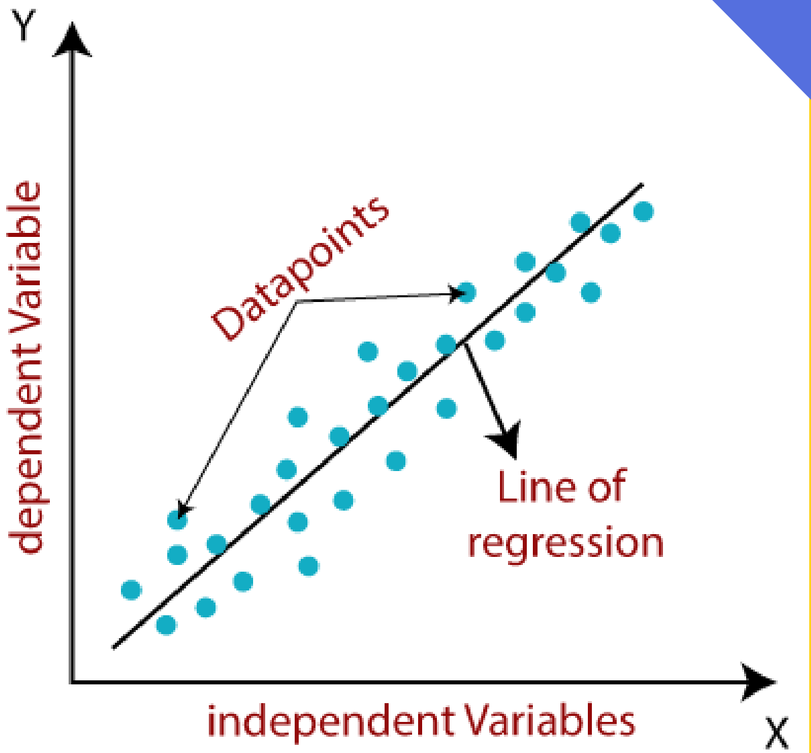
Evaluating and Improving the Model

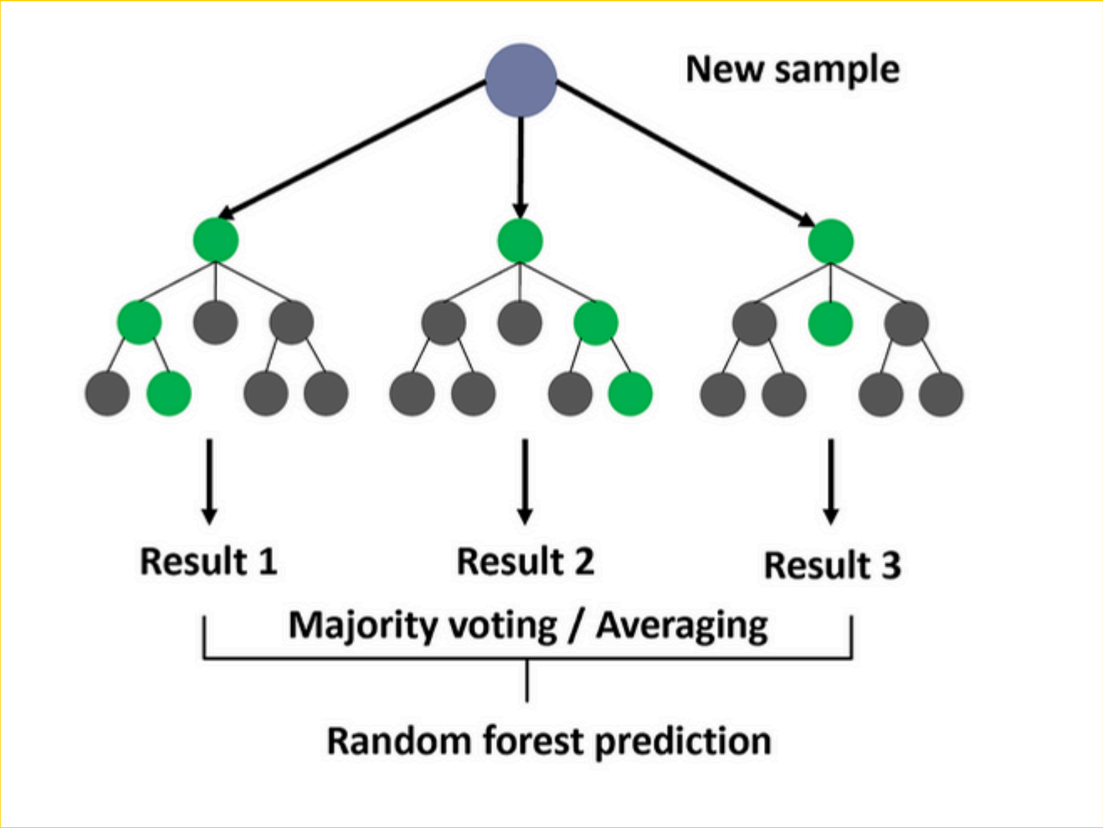
Deploying the Model

Supervised Learning

Linear Regression

This algorithm finds the best-fit straight line to show how one or more features relate to a continuous target variable.





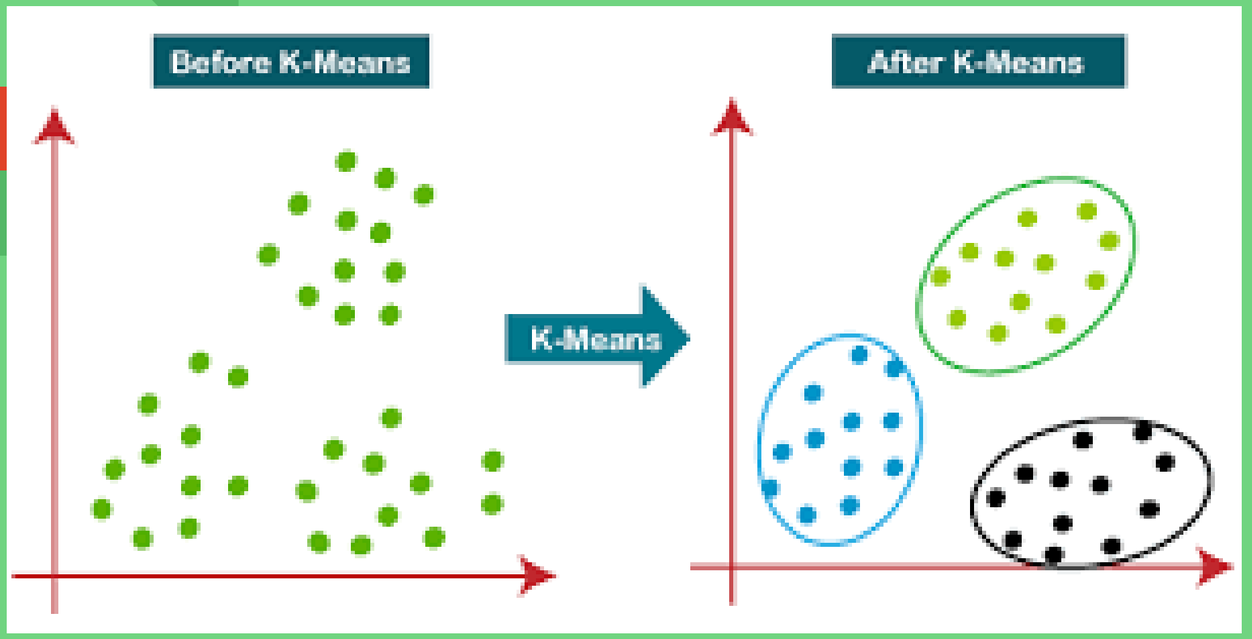
Decision Trees

split data into branches based on feature values to make decisions.

Random Forest

Combine multiple decision trees parallely to give the best ouput overcoming overfitting.

Unsupervised Learning



Clustering (K Means)

This algorithm sorts data into groups based on how similar the data points are, without needing any pre-labeled categories

Use Case

Customer segmentation in marketing, grouping similar images in photo libraries

Future Trends

Deep Learning

Deep learning is a type of artificial intelligence (AI) that teaches computers to process data in a similar way to the human brain.

Transformers

A transformer model is a neural network that learns context and meaning by tracking relationships in sequential data, like the words in this sentence.

AutoML

Automated machine learning, is the process of automating the time-consuming, iterative tasks of machine learning model development.

DAY - 2
Hands-On with Scikit-Learn and Regression Models

Google Developer Groups

On Campus • Techno Main Salt Lake

Pretisha Sahoo

Linkedin

Ritarshi Bandyopadhyay

Linkedin

Today’s Task

1 Understanding the Problem Statement

2 About the dataset

3 Data Preprocessing & Visualization

4 Model Building and Evaluation

Problem Satement

The problem is to predict the CO_2 emissions (in grams per kilometer) of vehicles based on various features such as engine size, fuel consumption, and vehicle specifications.

About the Dataset

Dataset Link

Columns : Make, Model, Vehicle Class, Engine Size(L), Cylinders, Fuel Consumption City (L/100 km), Fuel Consumption Hwy (L/100 km), Fuel Consumption Comb (L/100 km), Fuel Consumption Comb (mpg), CO_2 Emissions(g/km)

Preprocessing

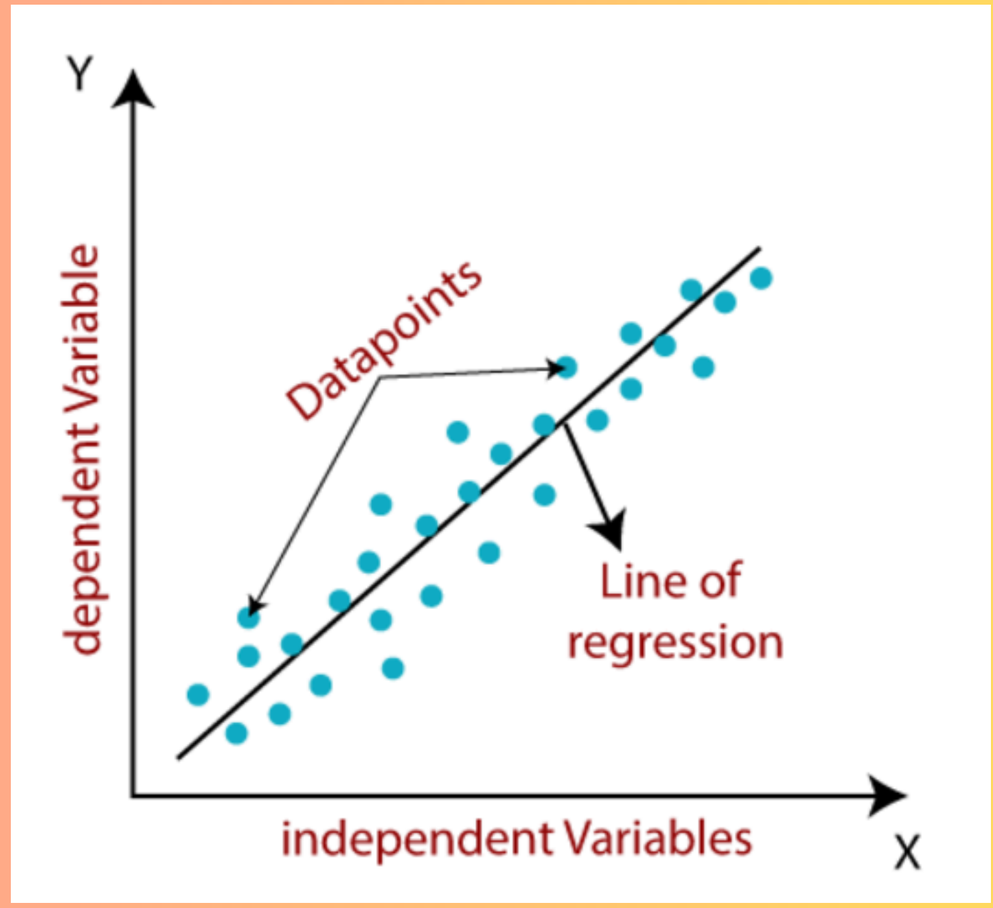
- Check Missing values , duplicates , data type
- Convert all Objects and Strings to numbers
- Feature selection based on correlation
- replace missing and null values
- Outlier detection

Linear Regression

It is Supervised Machine Learning Algorithm .

It predicts the continuous output variables based on the independent input variable.

Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a straight line (or hyperplane in higher dimensions) to minimize the difference between predicted and actual values.

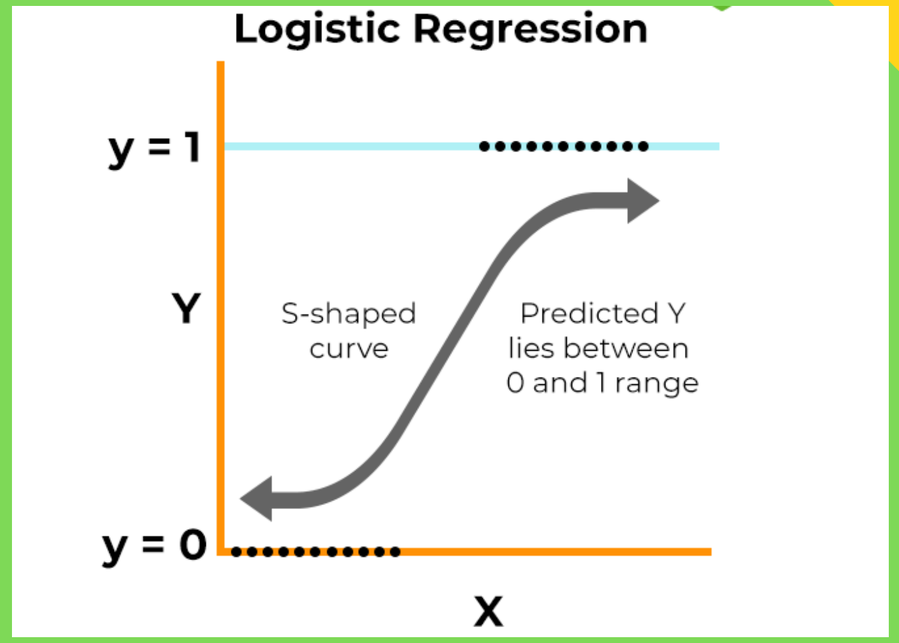


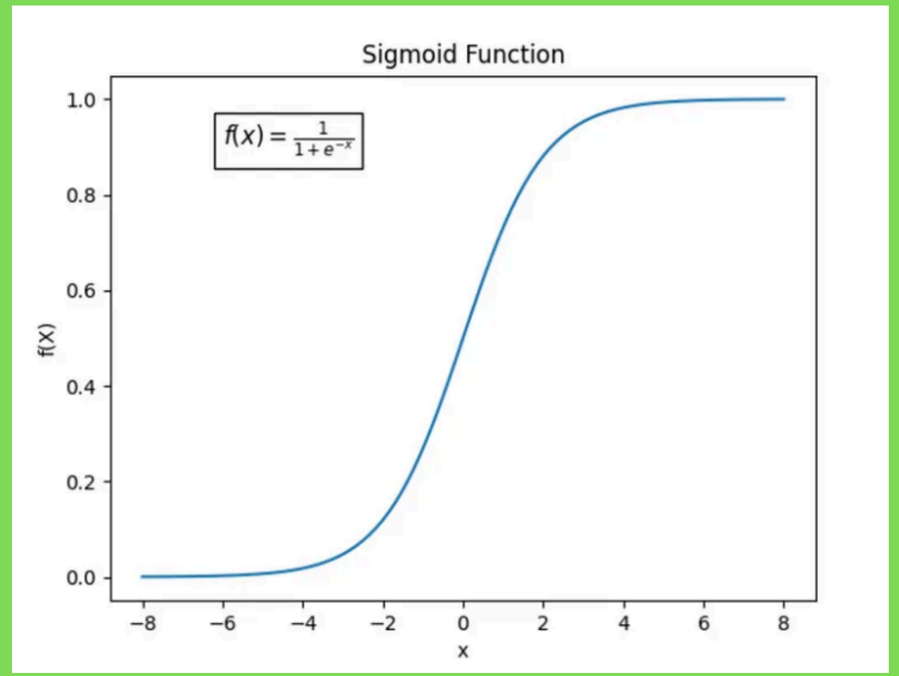
Logistic Regression

Logistic Regression predicts the probability of a binary outcome (e.g., yes/no, true/false) using a sigmoid function to map values between 0 and 1.

It assumes a linear relationship between the independent variables and the log-odds of the target variable.

Widely used in fields like healthcare (disease prediction), marketing, and finance (loan approval).





Evaluation Metrics

R2 Score

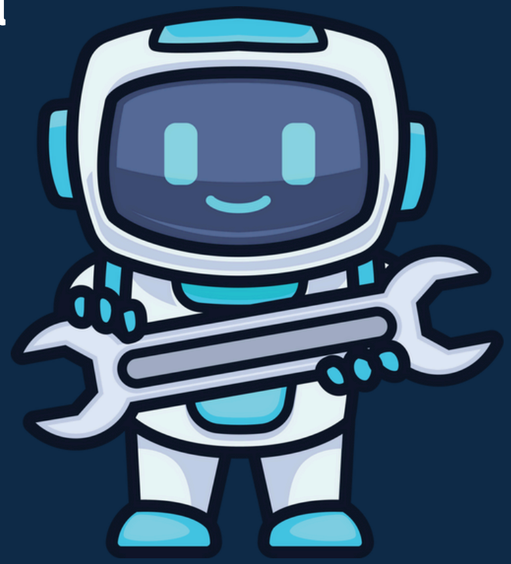
Models with higher R2 Score values are better.

Mean Squared Error

Models with lower Mean Squared Error (MSE) values are better

Not robust to Outliers.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$



Mean Absolute Error

Models with lower Mean Absolute Error (MSE) values are better.

Robust to Outliers

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} | x_i - x |$$

Day-3
Exploring Decision Trees with Scikit-Learn

Google Developer Groups

On Campus • Techno Main Salt Lake

Speakers

Rishi Bhattasali

Linkedin

Aishwarya Chowdhury

Linkedin

Today’s Agenda

1 Understanding the Problem Statement

2 About the dataset

3 Data Preprocessing & Visualization (EDA)

4 Model Building and Evaluation

Problem Satement

Predict the likelihood of stroke occurrence using patient demographic, health, and lifestyle data to aid in early detection.

About the Dataset:

1. Demographics: ID, gender, age, marital status, residence type.
2. Health Metrics: Hypertension, heart disease, BMI, glucose level.
3. Lifestyle: Smoking status, work type.
4. Target: Stroke occurrence (1 = Yes, 0 = No).

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Overview

Data Cleaning:

1. Removed irrelevant column (id).
2. Imputed missing values in the bmi column with its mean.

Feature Encoding:

1. Transformed categorical variables (e.g., gender, work_type) into numerical values using Label Encoding.

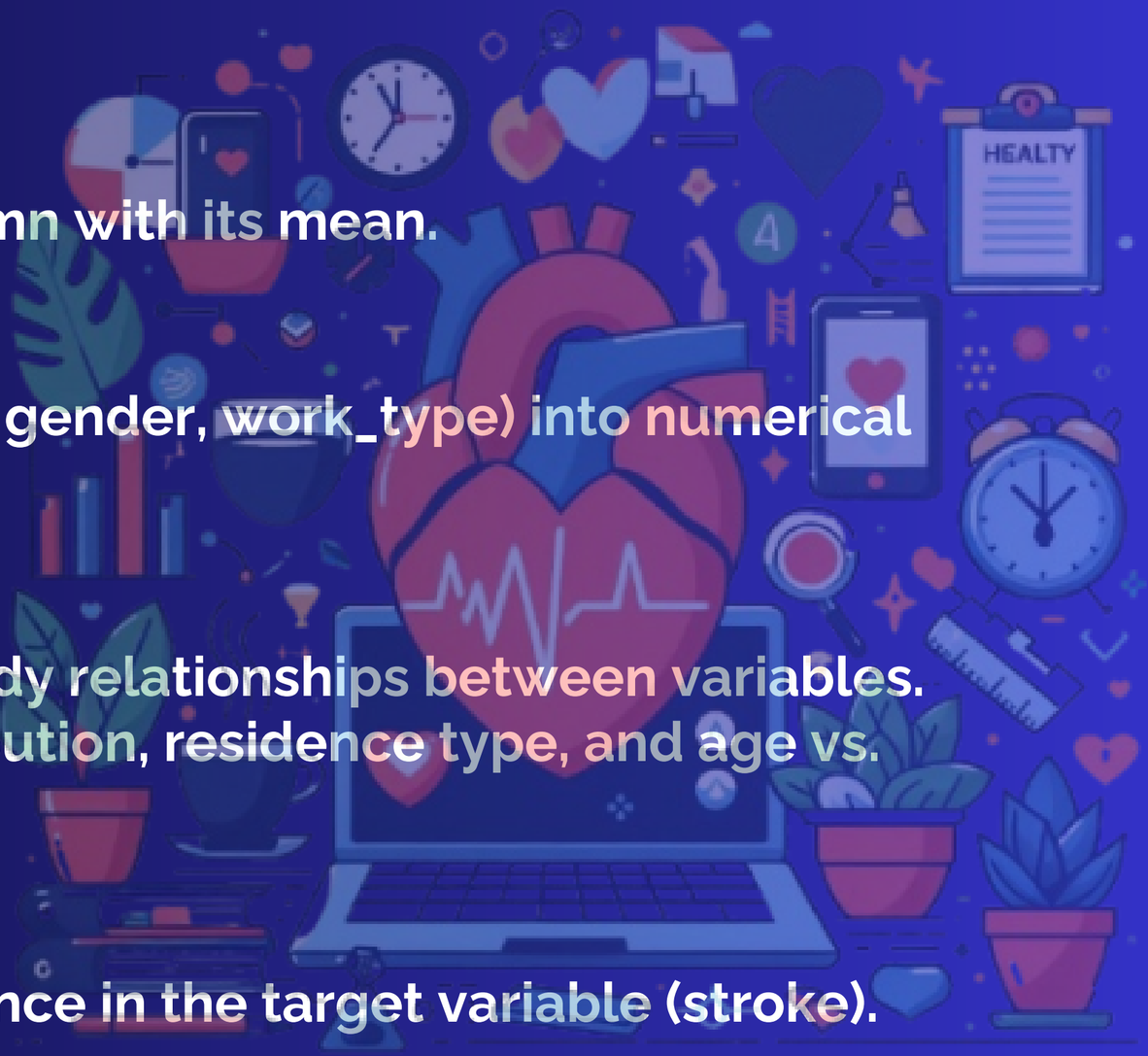
Exploratory Data Analysis (EDA):

1. Generated a correlation heatmap to study relationships between variables.
2. Created visualizations for gender distribution, residence type, and age vs. stroke trends.

Class Balancing:

1. Applied SMOTE to address class imbalance in the target variable (stroke).

Model Building and Evaluation



Exploratory Data Analysis

Process of analyzing and visualizing data to uncover patterns, detect anomalies, and prepare it for modeling.

Key Insights from the Correlation Heatmap:

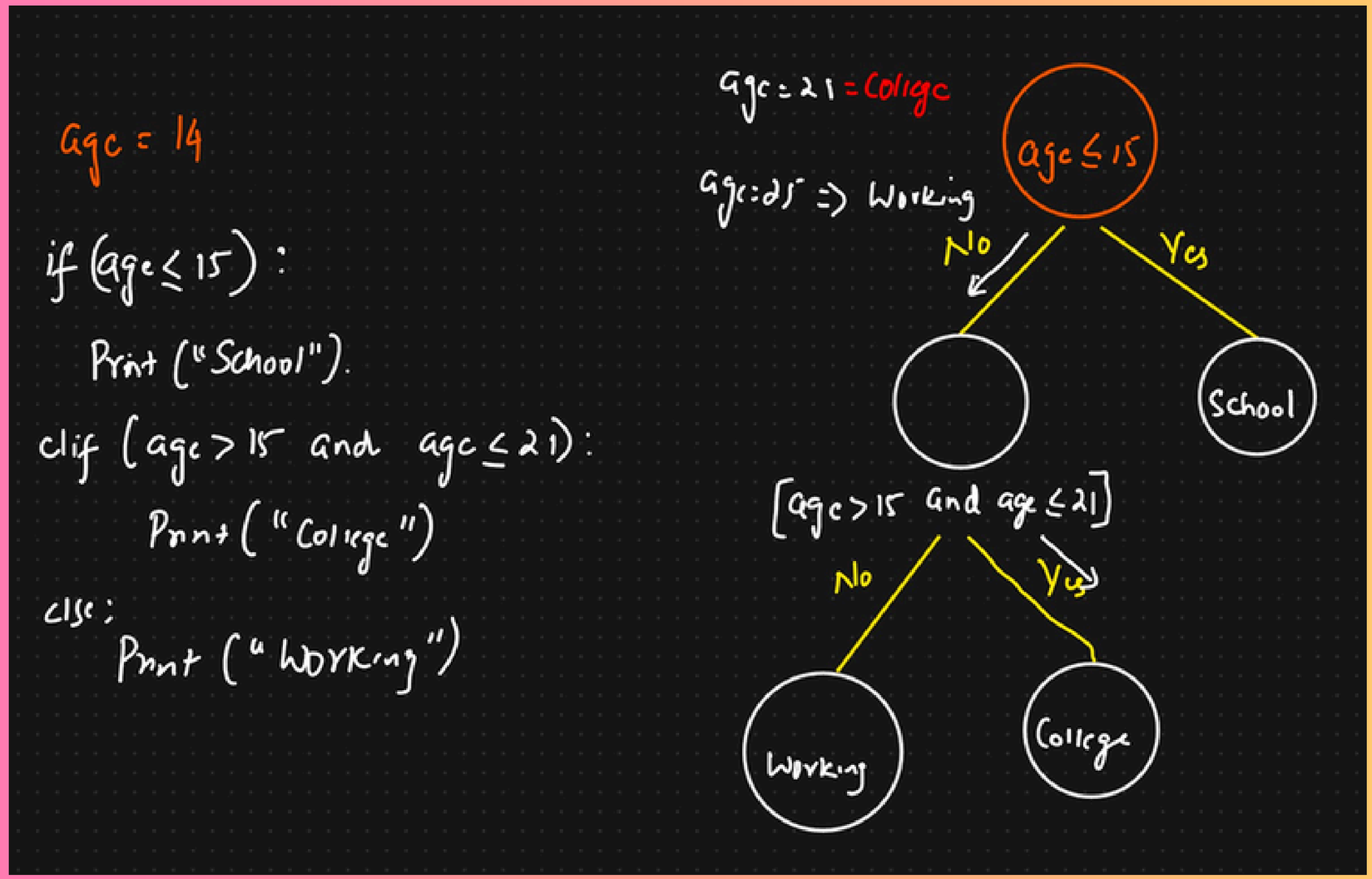
1. Strong Correlations:
   - `age` and `ever_married` are highly correlated (0.68), as older individuals are more likely to be married.
   - `bmi` has moderate positive correlations with `ever_married` (0.34) and `age` (0.33).

2. Stroke Risk Factors:
   - `stroke` shows moderate correlations with `age` (0.25), `hypertension` (0.13), and `heart_disease` (0.13).
   - Negligible impact from features like `gender` and `Residence_type`.

3. Work Type Observations:
   - Lower `work_type` values (e.g., Govt_job, Never_worked) have weak negative correlations with `age`, `bmi`, and `ever_married`.

4. Actionable Focus:
   - Older individuals with high `bmi`, hypertension, or heart disease prone to stroke.
   - Age and health factors are stronger predictors of stroke than lifestyle or work-related features.

Decision Trees



Thank You !!!