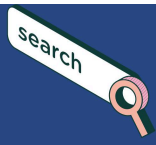


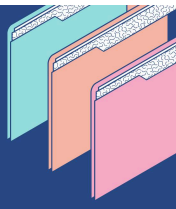
SPRING BOARD

By Infosys

Machine learning

M.SREEHARSHA
D.AJITH KUMAR
K.RAJESH REDDY

Employee attrition prediction

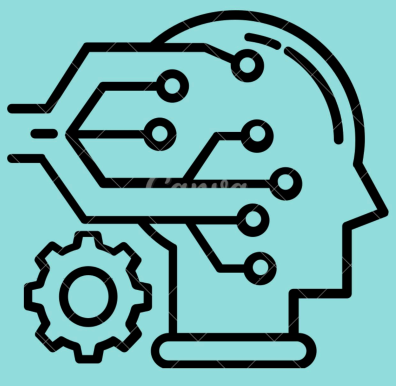


PROBLEM STATEMENT :Develop a predictive model that can identify employees at risk of leaving the company.

Objective : The primary objective of this project is to leverage historical HR data to build a robust predictive model that can effectively forecast employee attrition within the organization. By examining key features such as employee satisfaction, workload, departmental dynamics, and salary structures, the model will be able to identify patterns and trends that indicate potential turnover. This predictive tool will enable management to proactively address retention issues and implement targeted strategies to mitigate employee attrition.

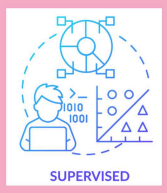
Machine learning

Machine learning (ML) is defined as a discipline of artificial intelligence (AI) that provides machines the ability to automatically learn from data and past experiences to identify patterns and make predictions with minimal human intervention.



Types of ML:
1.Supervised ML
2.Un supervised ML

Supervised Machine learning



- Input data is labeled
- used for predection
- data is classifed based on training dataset
- alogorithms: decesion tres,logistic regression ,support vector macchine

further classified into

1. Regression
2. Classification

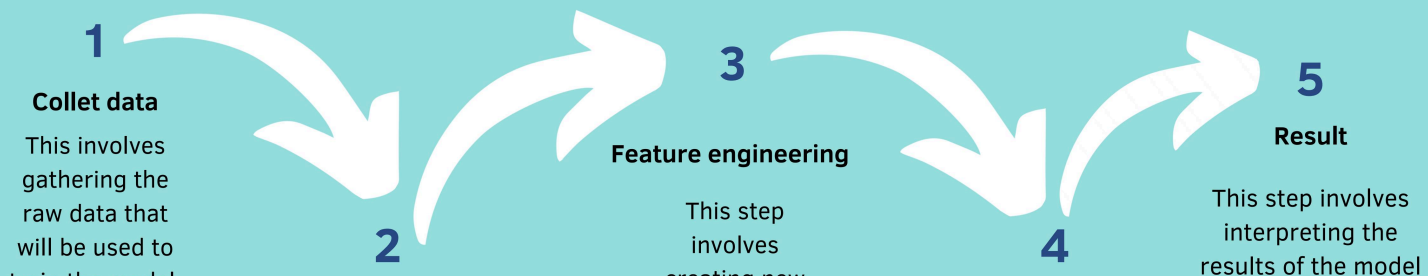
UnSupervised Macine learning

- input data in unlabeled
- used for analysis
- assign properties of given data to classify it
- algorithms: k-means clustering, hierarchical clustering, apriori

further classified into

1. Association
2. Clustering

WORKING



1

Collet data

This involves gathering the raw data that will be used to train the model.

2

Train data

This may involve cleaning the data, removing errors, and formatting it in a way that the model can understand.

3

Feature engineering

This step involves creating new features from the existing data.

4

Model validation

This step involves evaluating the performance of the model on a separate dataset that was not used to train the model.

5

Result

This step involves interpreting the results of the model and using them to make decisions or predictions.

About our project

I.Explore the Dataset:

exploring the given dataset. Understand its structure, features, and data types.

Check for missing values, duplicates and decide on appropriate data cleaning strategies.

2.Data Preprocessing:

Handle missing data, outliers, and perform any necessary data transformations.
Split the dataset into features (independent variables) and the target variable .

3.Exploratory Data Analysis (EDA):

Conduct exploratory data analysis to understand the distribution of variables.

Visualize relationships between variables, especially focusing on the satisfaction level, working hours, department, promotion, and salary level.

4.Correlation Analysis:

Use statistical methods to determine the relationship between satisfaction level and working

5.Feature Importance Analysis:

Analyze the effect of satisfaction level, department, promotion in the last 5 years, and salary level on employee exits.

About our project

6.Machine Learning Model Building:

Choose an appropriate machine learning algorithm for predicting employee exits. Common algorithms for binary classification problems include Logistic Regression, Decision Trees, Random Forest, or Gradient Boosting.

7.Model Evaluation and Tuning:

Assess the model's performance using metrics such as accuracy, precision, recall.

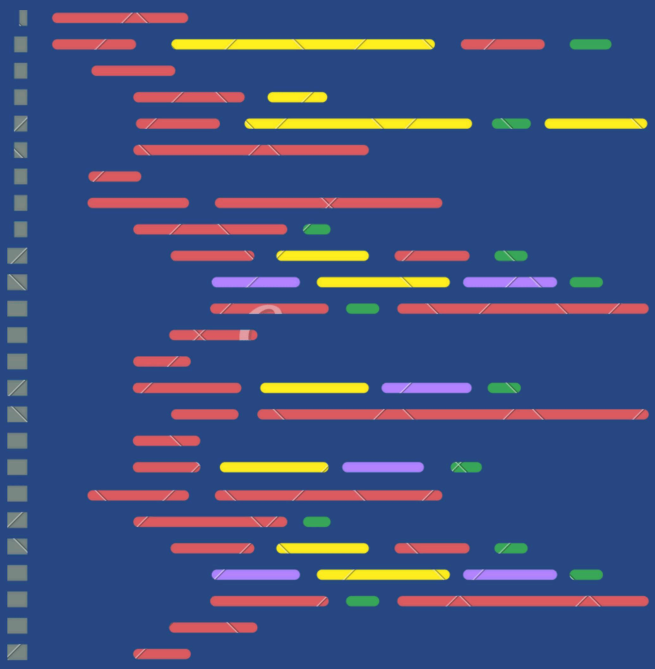
8.Interpretation:

Interpret the results and provide insights into the key factors contributing to employee exits.

9.Implementation and Monitoring:

Once satisfied with the model's performance, implement it in the HR system.
Set up monitoring mechanisms to track employee satisfaction and other relevant factors over time.

IMPLEMENTATION



*The scatter plot is helpful for visualizing the relationship between satisfaction level and average monthly hours, especially for employees who left.

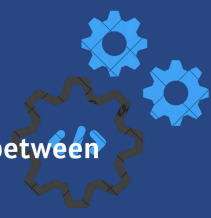
aPandas: Used for data manipulation and analysis.

Matplotlib and Seaborn: Used for data visualization.

Scikit-learn: Used for machine learning tasks, including Random Forest Classifier, train_test_split, accuracy_score, and classification_report.

NumPy: Used for numerical operations.

1.Loading Data: The dataset is loaded using Pandas from a CSV file.



2.Exploratory Data Analysis (EDA): A scatter plot is created using Seaborn to visualize the relationship between satisfaction level and average monthly hours for employees who left the organization (left == 1) .

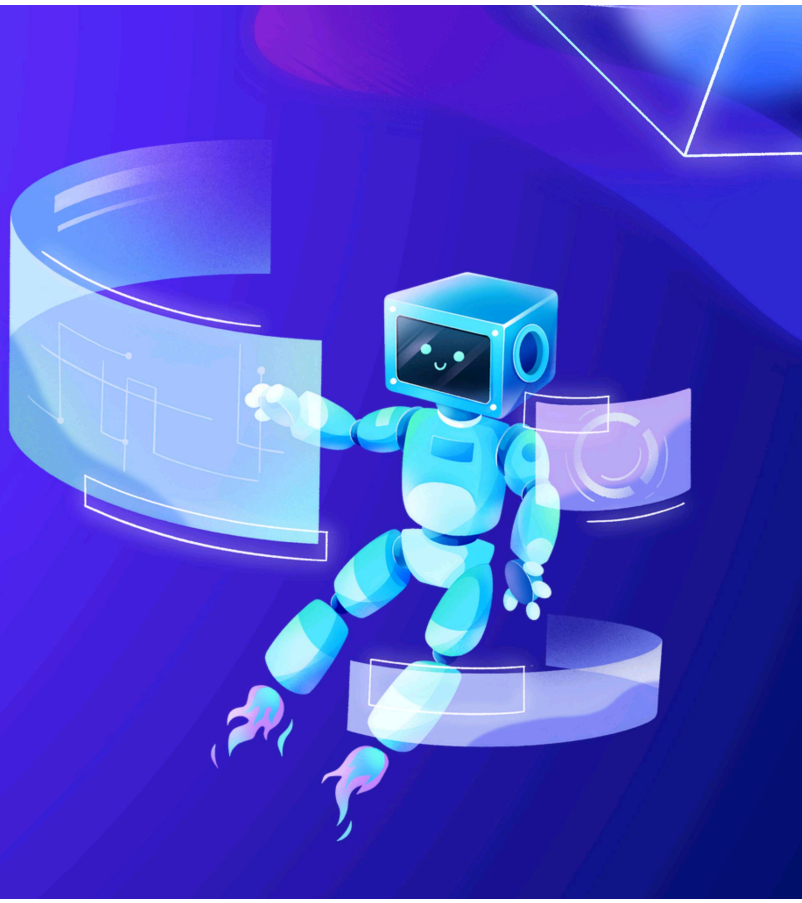
3.Data Preprocessing: Label encoding is applied to convert categorical variables ('Department' and 'salary') into numeric format.

4.Data Splitting: The dataset is split into features (X) and the target variable (Y). Further, the dataset is split into training and testing sets using train_test_split.

5.Random Forest Model Building: A Random Forest classifier is instantiated with 100 trees (n_estimators=100) and a random state for reproducibility. The model is trained using the training set.

6.Model Prediction and Evaluation: The trained model is used to make predictions on the testing set (Y_pred). The accuracy score and classification report (precision, recall, F1-score) are printed to evaluate the model's performance.

THANK YOU!



Presentation link:

https://www.canva.com/design/DAF-AocMCZE/hQAyerJgG8oUQETAJWer7g/edit