Annexure-1

Employee Attrition Predictor

Cipher Schools

A training report

Submitted in partial fulfilment of the requirements for the award of degree of

B.Tech. CSE (Artificial Intelligence and Machine Learning)

Submitted to

LOVELY PROFESSIONAL UNIVERSITY
PHAGWARA, PUNJAB

From June 4, 2025 to July 15, 2025

SUBMITTED BY

Name of student: Bhawya Gulati

Registration Number: 12312608

Signature of the student: Bhawya

Annexure-II: Student Declaration

To whom so ever it may concern

I, **Bhawya Gulati**, Registration Number **[12312608]**, hereby declare that the work done by me on **"A Guide To Machine Learning With Data Science "** from **June, 2025** to **July, 2025**, is a record of original work for the partial fulfilment of the requirements for the award of the degree, **B.Tech. in CSE(AI and ML) .**

Bhawya Gulati

Dated: 13/08/2025

Training certificate from organization/ Company

LOVELY PROFESSIONAL UNIVERSITY

https://www.cipherschools.com/certificate?id=6628b66c2d7789dcb522a92b

Cipher Schools

Certificate of Completion

This is to certify that

Bhawya Gulati

studying at **Lovely Professional University,** has successfully completed training in **A Guide to Machine Learning with Data Science** from CipherSchools during the period of **June 4, 2025, to July 15, 2025** .

The training comprised 70 hours of learning , and the participant's performance has been assessed as Satisfactory.

ANURAG MISHRA
Founder CipherSchools

CipherSchools, India

Certificate ID : CSW2025-14847

Scan to verify

Acknowledgement

I would like to formally express my appreciation for the guidance, encouragement, and resources offered by Cipher Schools in creating my Employee Attrition Predictor project, which played a very significant role in my overall understanding and real-world implementation of data science, analytics, and machine learning concepts. The systematic training curriculum and organized syllabus touched upon an array of important subjects, ranging from the basics of data science, Python programming, and version control, to advanced Excel for data analysis, data visualization, Power BI, statistics analysis, data preprocessing, and supervised and unsupervised machine learning algorithms. These modules gave me a robust theoretical background as well as the ability to work confidently with actual datasets, manage various data problems, and use sophisticated analytical methods to create actionable insights. The mentorship and expert feedback of instructors and mentors helped immensely in grasping intricate ideas like regression analysis, classification methods, clustering methods, and neural networks, as well as practical software like Pandas, NumPy, Matplotlib, Seaborn, Plotly, and Power Query. Their thorough explanations and suggestion based on industry helped me overcome data cleaning issues, feature engineering, building models, and their evaluation. With interactive sessions, practical assignments, and experimentation in Google Colab, I learned how to implement statistical measures, develop effective visualizations, and compare machine learning models for performance improvement. The course's project-oriented approach enabled me to implement my learning directly by creating and building the Employee Attrition Predictor, which predicts whether an employee will leave the company based on attributes including monthly income, age, job satisfaction, years of experience in the company, and overtime status. Applying both Logistic Regression and Decision Tree classifiers, and comparing the performance of each, gave me useful experience with model selection, evaluation metrics, and domain knowledge—in this instance, HR analytics and employee retention practices. The carefully designed syllabus and experiential learning facilitated a productive learning experience that encouraged analytical reasoning, problem-solving, and critical thinking, and the application of real datasets enabled me to connect book concepts to professional practice. This experience has not only enhanced my technical and analytical skills but also provided me with real-world insights into the data science workflow, from data acquisition and preprocessing to model building, evaluation, and reporting. I am truly thankful to Cipher Schools for their mentorship, extensive resources, and well-designed learning environment, which have been of enormous value in building my confidence, abilities, and preparedness to contribute effectively in the domain of data science and analytics.

List of Contents

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under the ROC Curve
BI	Business Intelligence
CNN	Convolutional Neural Network
DAX	Data Analysis Expressions
EDA	Exploratory Data Analysis
ETL	Extract, Transform, Load
GIT	Version Control System
IDE	Integrated Development Environment
KPI	Key Performance Indicator
MAE	Mean Absolute Error
MSE	Mean Squared Error
NLP	Natural Language Processing
NumPY	Numerical Python Library
PCA	Principal Component Analysis
SVM	Support Vector Machine
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic

Chapter-1

LOVELY PROFESSIONAL UNIVERSITY

INTRODUCTION OF THE PROJECT UNDERTAKEN

➢ Objectives of the work undertaken:

The main goal of this project, Employee Attrition Predictor, is to create a machine learning solution to predict if an employee is going to leave an organization. Employee attrition is one of the major problems for human resource departments, causing higher recruitment expenses, lower productivity, and loss of skilled people. This project intends to solve the problem by employing past employee data to detect patterns and drivers of attrition, so organizations can take proactive steps in retaining employees.

The dataset employed contains important characteristics like monthly income, age, job satisfaction, company tenure in years, and overtime status. Data cleaning, normalization, missing value handling, and exploratory data analysis were undertaken to identify insightful information.

Two machine learning models—Logistic Regression and Decision Tree Classifier—were utilized in Python through Google Colab. Logistic Regression was used for its statistical nature and capability to make probability-based predictions, and the Decision Tree Classifier was chosen because of its rule-based and easily interpretable model structure. Both models were trained, tested, and scored based on performance measures, and results were compared to determine the more effective method for this classification task.

This project combines several skills acquired throughout the course, such as advanced Excel for preliminary analysis, Python coding, Pandas and NumPy for data manipulation, Matplotlib and Seaborn for plotting, and statistical techniques for model assessment. Utilizing these tools and methods, the project is able to effectively show how data science can be used for practical HR analytics, enabling actionable insights to reduce employee turnover.

➢ Scope of the Work:

The breadth of this project is the utilization of data science methodologies and machine learning models to solve a real HR analytics challenge—employee attrition prediction. The project includes the entire data science process, ranging from data gathering and preprocessing to model development, validation, and result interpretation. Through a dataset that includes features like monthly income, age, job satisfaction, company years, and overtime status, data cleaning, normalization, and missing values handling were all accomplished in order to prepare it for analysis.

Modeling involved running two algorithms, Logistic Regression and Decision Tree Classifier, to see how they performed as predictors. Logistic Regression presented a probability-based model, while the Decision Tree presented a more understandable,

rule-based model. Both were run using applicable metrics to find out how appropriate they were for classifying.

In addition to model building, the scope also covered visualization of variable relationships, interpretation of feature importance, and reporting findings in a manner enabling actionable decision-making. Integrating advanced Excel, Python programming, data visualization software, and statistical analysis methods acquired throughout the course, the project indicates the applied applicability of theoretical ideas to a real business problem, rendering useful insights to employee retention initiatives.

- **Importance and Applicability**:

The project is of high value to the discipline of human resource management and business decision-making. Employee attrition has direct implications on organizational stability, operating efficiency, and overall productivity and is, therefore, a key area that needs to be tracked and regulated by companies. Through the use of machine learning models, this project allows HR professionals to detect potential leavers among employees, enabling organizations to apply focused retention actions, decrease recruitment expenses, and ensure continuity of workforce.

From a technical viewpoint, the project illustrates the applicability of data science methods to fixing real-life issues. Through application of algorithms such as Logistic Regression and Decision Tree Classifier, the platform not only makes predictions concerning the chances of attrition but also identifies important factors that drive it, including job satisfaction, salary, and overtime trends. These findings aid decision-makers in setting their priorities right and enhancing the participation of employees.

The relevance of this project lies not only in HR analytics but across various sectors where employee retention is crucial, such as IT, manufacturing, healthcare, and education. The approach adopted can also be applied to other predictive analytics projects, such as customer churn prediction or risk forecasting. This cross-applicability speaks for the use of data-driven strategies in organizational plans for sustainable long-term development.

- **Role and profile:**

In the project, my responsibility was that of a data science practitioner tasked with end-to-end project implementation. I was directly involved in problem domain understanding, feature selection, and dataset preparation for analysis. My tasks involved conducting data cleaning, normalization, and missing value handling, followed by exploratory data analysis to determine trends and correlations.

I created and tested two machine learning models—Logistic Regression and Decision Tree Classifier—with Python in Google Colab with proper training, testing, and evaluation using relevant performance measures. I also compared the output of both models to identify which approach is most appropriate for predicting employee attrition.

Furthermore, I used Matplotlib and Seaborn to create visualizations that effectively communicate findings and made sure the project fit into actual HR analytics requirements. This position provided an opportunity to take theory learned through the course and apply it to a concrete business application.

Chapter 2 INTRODUCTION OF THE COMPANY / WORK: CIPHER SCHOOLS

➢ Company's Vision and Mission:

CipherSchools envisions a future where high-quality, practical, and industry-relevant education is available to learners across the globe, regardless of geographical, financial, or cultural barriers. The vision is to create a dynamic and inclusive learning ecosystem that not only imparts technical skills but also nurtures creativity, problem-solving abilities, and professional growth. By leveraging innovative teaching methodologies, engaging course structures, and expert mentorship, CipherSchools aims to become the go-to platform for learners seeking to enhance their career prospects, transition into new roles, or upgrade their skill sets. The long-term goal is to bridge the gap between education and employability, ensuring that every learner is equipped with the tools and confidence to thrive in an ever-evolving global economy.
The mission of CipherSchools is to empower individuals through accessible, flexible, and outcome-oriented learning experiences. The platform focuses on creating well-structured courses in emerging and high-demand fields such as data science, software development, artificial intelligence, business analytics, and creative domains. By collaborating with skilled industry professionals, CipherSchools provides learners with mentorship that translates theoretical knowledge into practical application. The mission extends to fostering a collaborative community where learners can network, share insights, and solve problems together.
CipherSchools emphasizes hands-on learning through real-world projects, enabling students to build strong portfolios that showcase their skills to potential employers. It is committed to providing an engaging and interactive educational environment that encourages continuous improvement and adaptability. With a focus on innovation, personalized learning, and career-oriented outcomes, CipherSchools seeks to not only educate but also inspire learners to achieve their goals, contribute meaningfully to their industries, and stay ahead in the competitive professional landscape.

➢ Origin and growth of company:

CipherSchools is a bootstrapped educational technology company established in early 2020 in India Founded with the vision of making high-quality and practical online learning accessible to all, regardless of location or background, it began as a video streaming platform designed to connect passionate learners with experienced industry professionals
Starting modestly, CipherSchools sought to address the gap between theoretical knowledge and job-ready skills by providing structured training and hands-on project experiences. Within its first year, the platform successfully reached over 10,000 students through free webinars and workshops, alongside serving more than 4,000 students through paid programs

The company's early growth was fueled by its B2B efforts, targeting universities and colleges, which helped establish a sustainable foundation. While its B2C efforts initially encountered challenges due to limited marketing budgets, these experiences led to strategic pivots in product and outreach approaches
Operationally headquartered in Gurgaon, CipherSchools has grown into a compact yet impactful team of 11 to 50 employees, scaling its offerings and refining its platform to deliver engaging content across EdTech domains such as full-stack development, data science, artificial intelligence, and competitive programming The company is recognized for its student-centric approach, dedication to continuous improvement, and its aim to become one of the leading online learning platforms in the country. CipherSchools' trajectory since inception reflects a dynamic journey—from its grassroots beginnings to delivering meaningful educational experiences to thousands—underscoring its growth-focused culture and commitment to democratizing education in India.

- **Various departments and their functions**:

While an online training platform, Cipher Schools operates with a highly organized and strategically structured framework designed to deliver an exceptional and seamless learning experience to students across the globe. The company is composed of multiple specialized teams, each playing a vital role in ensuring that the platform not only delivers high-quality educational content but also fosters continuous engagement, skill growth, and career readiness among learners. The **content creation team** works tirelessly to design, develop, and update courses that align with the latest industry trends, covering in-demand technologies, programming languages, and professional skills. Simultaneously, the **technical support team** ensures that students have uninterrupted access to learning resources, quickly resolving any platform-related or technical issues they may encounter. The **student mentorship team** provides personalized guidance, career advice, and doubt-clearing sessions to help learners navigate their educational journey effectively, making the process more interactive and supportive. Additionally, the **platform development team** constantly works on upgrading the website and mobile application, enhancing usability, integrating innovative learning tools, and ensuring a smooth user experience. This collaborative structure allows Cipher Schools to combine the flexibility of online learning with the personalization and guidance of traditional education, creating an environment where students not only gain technical expertise but also develop problem-solving abilities, communication skills, and confidence. By functioning as an interconnected ecosystem, Cipher Schools successfully bridges the gap between academic learning and industry demands, ensuring that its students are well-prepared for real-world challenges and can thrive in competitive professional environments.

- **Organization chart of the company**:

As an online platform, the organization chart of Cipher Schools is designed to be dynamic and adaptable, allowing it to respond quickly to evolving industry trends, learner needs, and technological advancements, while still maintaining a well-defined structure that ensures smooth operations and high-quality delivery. At the top, the **founders** provide vision, strategic direction, and leadership, shaping the long-term goals of the platform, building partnerships, and ensuring that the mission of empowering learners through accessible, industry-relevant education remains the core focus. Beneath them, **content leads** oversee the entire process of course design, development, and improvement, working closely with industry experts and experienced

educators to create engaging, up-to-date, and practical learning materials that cater to a diverse audience. These leads guide the **content creation teams**, which include instructors, curriculum designers, and multimedia specialists who collaborate to produce video lectures, assignments, projects, and interactive materials tailored for both beginners and advanced learners. Parallel to this, the **technical teams** play a crucial role in building, maintaining, and enhancing the platform's infrastructure, ensuring that the website and mobile applications run smoothly, securely, and efficiently. They work on integrating innovative features such as progress tracking, personalized learning recommendations, live mentorship sessions, and community discussion forums to enrich the user experience. Supporting these core functions are **student engagement and mentorship teams**, who ensure that learners receive timely guidance, doubt clarification, and career advice, making the platform more than just a repository of content—it becomes a community of support and growth. The structure is deliberately flexible, allowing cross-functional collaboration between content, technical, and support teams, enabling rapid updates to courses, quick resolution of issues, and the integration of emerging tools or teaching methods. This adaptability ensures that Cipher Schools stays competitive in the fast-paced world of online education while maintaining a learner-first approach. Through this interconnected and agile organizational framework, the platform successfully combines visionary leadership, expert-driven content, robust technology, and personalized support to create an engaging, effective, and future-ready learning environment for students worldwide.

Chapter 3

Brief description of the work done

- **Position of Internship and roles:**

During my internship with **CipherSchools**, I undertook the position of a **Data Science Intern**, which involved working on a real-world machine learning project titled *Employee Attrition Predictor*. This role was designed to integrate the theoretical knowledge gained from the training program with practical, hands-on application, enabling me to develop technical expertise and problem-solving skills while working on a business-relevant challenge.

My responsibilities began with understanding the scope of the project and analyzing the dataset, which consisted of employee-related attributes such as monthly income, age, job satisfaction, years at the company, and overtime status. I was tasked with performing **data preprocessing** activities, including cleaning, normalization, and handling missing values, to ensure the dataset was ready for analysis. Following this, I conducted **exploratory data analysis (EDA)** to identify patterns, correlations, and possible predictors of attrition.

A key part of my role was **model development and evaluation**. I implemented two machine learning algorithms—**Logistic Regression** and **Decision Tree Classifier**—using Python in Google Collab. Logistic Regression was chosen for its statistical, probability-based prediction capabilities, while the Decision Tree offered a rule-based, easily interpretable structure. I trained and tested both models, compared their performance using accuracy and other relevant metrics, and documented the findings to determine the more effective approach for this classification problem.

Additionally, I created **data visualizations** using Matplotlib and Seaborn to present the results in a clear and impactful manner. This included visualizing feature distributions, relationships between variables, and model performance comparisons. Beyond technical tasks, I was also responsible for **report preparation**, summarizing the methodology, findings, and recommendations in a professional format suitable for stakeholders. This internship role provided me with valuable exposure to the end-to-end data science workflow, enhanced my proficiency with tools like Pandas, NumPy, and Power BI, and improved my ability to translate analytical insights into actionable business strategies. Overall, the position strengthened both my technical skills and professional competencies, preparing me for future roles in the field of data science and analytics.

- **Activities/ equipment handled**:

Throughout my internship at CipherSchools,
I engaged in numerous tasks that encompassed the entire life cycle of a machine learning project. The main emphasis was on creating the Employee Attrition Predictor, a binary classification tool designed to predict if an
employee will depart from the organization.

The tasks started with data gathering and preprocessing, during which I focused on loading and cleansing the dataset using Python in Google Colab.
I processed tasks such as eliminating duplicate entries, handling absent values, encoding categorical variables (for instance, transforming "OverTime" from Yes/No to binary), and standardizing numerical attributes. These
preprocessing measures were crucial for guaranteeing model precision and uniformity . Subsequently, I conducted exploratory data analysis (EDA) utilizing visualization libraries like Matplotlib and Seaborn. This involved generating histograms, boxplots, scatter plots, and correlation heatmaps to analyze relationships
between variables such as monthly income, age, and tenure at the company.
For developing the model, I utilized two algorithms: Logistic Regression and Decision Tree Classifier, employing the Scikit-learn library. I divided the dataset into training and testing components, trained the models,
and assessed them with metrics such as accuracy, precision, recall, and F1-score. The Decision Tree model gave straightforward, understandable rules, whereas Logistic Regression delivered probabilistic forecasts.
In addition to the coding tasks, I utilized Power BI to visualize aggregated results in a dashboard format. This tool enabled stakeholders to engage with the data
and comprehend patterns that lead to attrition.
In terms of the tools and equipment used, I primarily performed my work in Google Colab for coding and experimentation, benefiting from GPU
acceleration to expedite model training. I utilized Python libraries including Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn. For preparing the report,
I utilized Microsoft Word and PowerPoint to develop professional documents and presentations summarizing the results of the project. Overall, the activities and tools I handled during this internship gave me comprehensive exposure to the workflow of a machine learning project—from data preparation and analysis to model implementation, visualization, and reporting—enhancing my ability to manage end-to-end AI solutions.

- ## Challenges faced and how those were tackled:

During my internship at **CipherSchools** while working on the *Employee Attrition Predictor* project, I encountered several technical and practical challenges. Overcoming these obstacles was a significant learning experience that enhanced my problem-solving skills and technical expertise.

1. Data Quality and Missing Values

One of the first challenges was dealing with incomplete and inconsistent data. Some records had missing values in important fields like "Monthly Income" and "Years at Company." Directly using such data would have led to inaccurate predictions.

Solution: I used **Pandas** to handle missing values, applying techniques like mean/median imputation for numerical data and mode imputation for categorical data. Outliers were detected using boxplots and handled carefully to prevent skewing the model's results.

2. Categorical Variable Encoding
The dataset contained categorical variables like "OverTime" (Yes/No) and "Job Satisfaction" (ordinal values). Directly feeding these into the model was not possible.
Solution: I applied **Label Encoding** for binary features and **One-Hot Encoding** for multi-class features to make them machine-readable while preserving their meaning.
3. Class Imbalance
There was an imbalance between employees who stayed and those who left, with the majority belonging to the "stayed" category. This imbalance risked biasing the model towards predicting "stay" more often.
Solution: I used **SMOTE (Synthetic Minority Oversampling Technique)** to oversample the minority class and balance the dataset, ensuring the model learned patterns from both categories equally.
4. Overfitting in Decision Tree Model
The Decision Tree model initially performed extremely well on training data but poorly on test data, indicating overfitting.
Solution: I applied **pruning techniques** by limiting the tree depth and minimum samples per split, as well as cross-validation to ensure better generalization.
5. Performance Optimization
Running the models multiple times during experimentation took considerable time, especially during hyperparameter tuning.
Solution: I utilized **Google Colab's GPU acceleration** for faster computation and optimized my code by reducing redundant loops and leveraging vectorized operations in NumPy and Pandas.
6. Interpretation and Communication of Results
Explaining the technical model outputs to non-technical stakeholders posed a communication challenge.
Solution: I created **visual dashboards** using **Power BI** and simple charts in Matplotlib to present results in an understandable way, highlighting actionable insights rather than just raw metrics.
Overall, tackling these challenges not only improved the performance and reliability of the Employee Attrition Predictor but also strengthened my skills in data preprocessing, model optimization, and stakeholder communication. Each difficulty was approached methodically, ensuring the final solution was robust, accurate, and user-friendly.

- ## **Data analysis**:

The data analysis phase of the *Employee Attrition Predictor* project was a critical step in understanding the dataset, identifying patterns, and preparing the information for machine learning model development. The dataset provided contained various employee-related attributes such as **Monthly Income, Age, Job Satisfaction, Years at Company, and OverTime (Yes/No)**, along with the target variable — **Attrition** (Yes/No).
1. Understanding the Dataset
The first step involved loading the dataset into **Pandas DataFrame** and performing an initial inspection using functions like .head(), .info(), and .describe(). This helped identify the nature of the variables, their data types, and any irregularities such as missing values or inconsistent entries.
2. Handling Missing Values and Outliers
Upon analysis, it was observed that some attributes had missing values, which could potentially reduce model accuracy. For numerical attributes like "Monthly Income,"

missing values were imputed using the median, while for categorical attributes such as “OverTime,” the mode was used. Outliers, especially in numerical features like “Years at Company” and “Monthly Income,” were detected using boxplots and handled to prevent skewing the analysis.

3. Categorical Data Encoding

Several features were categorical in nature, such as “OverTime” and “Job Satisfaction.” These were converted into numerical form using **Label Encoding** for binary categories and **One-Hot Encoding** for multi-class categories, ensuring compatibility with machine learning algorithms.

4. Exploratory Data Analysis (EDA)

Using **Matplotlib** and **Seaborn**, various visualizations were created to identify trends and relationships. For example, bar plots revealed that employees working overtime had a significantly higher attrition rate. Similarly, heatmaps of correlation matrices helped understand which variables were strongly associated with attrition, revealing factors like low job satisfaction and low income as key predictors.

5. Class Imbalance Identification

EDA also revealed a significant class imbalance, with more employees staying than leaving. This imbalance was addressed later in the modeling phase using **SMOTE** to create a balanced dataset, ensuring fair learning.

Through this thorough data analysis, valuable insights were extracted, guiding the feature engineering process and helping in the selection of appropriate machine learning algorithms. This step ensured that the dataset used for modeling was clean, relevant, and structured for optimal prediction accuracy.

CONCLUSION

The completion of the *Employee Attrition Predictor* project marked the successful culmination of both theoretical learning and practical application gained throughout the course. The course provided a solid foundation in machine learning concepts, data preprocessing, exploratory data analysis, model selection, and evaluation techniques. It also emphasized the importance of understanding business problems, translating them into analytical tasks, and delivering actionable insights through data-driven solutions.

Through this project, I was able to integrate multiple aspects of the course into a cohesive, real-world application. The primary objective — predicting whether an employee is likely to leave the company — addressed a genuine business challenge faced by many organizations. This required not only technical proficiency in Python, Pandas, Scikit-learn, and visualization libraries such as Matplotlib and Seaborn, but also a deep understanding of the underlying business implications of employee attrition.

From data collection and cleaning to encoding categorical variables, handling class imbalance, and performing exploratory data analysis, each stage reinforced the critical role of proper data handling in achieving reliable results. The model development phase demonstrated how Logistic Regression could be effectively applied to binary classification problems, while the evaluation metrics such as accuracy, precision, recall, and F1-score highlighted the strengths and limitations of the chosen approach.

The project also fostered problem-solving skills by addressing real challenges, such as missing data, skewed distributions, and feature correlations. Overcoming these issues ensured the final model's robustness and practical usability. Moreover, the insights gained through the analysis — such as the impact of overtime, job satisfaction, and income levels on attrition — can directly help organizations in formulating retention strategies.

In conclusion, this project not only validated the knowledge acquired during the course but also enhanced my ability to apply it effectively in solving practical problems. The combination of technical skills, analytical thinking, and domain understanding has significantly contributed to my growth as a data science practitioner. This experience has strengthened my confidence in handling end-to-end machine learning projects and prepared me for more complex, impactful challenges in the future.

Reference

https://www.cipherschools.com/courses/a-guide-to-machine-learning-with-data-science-a92b/lecture-0-introduction-to-the-course-71a5

Screen Shots:

Figure 1 :- Import Library

```
# 1 Import Libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
```

Figure 2 :- Loading dataset and displaying its shape (200 rows x 6 columns)

```
df = pd.read_excel("/content/drive/MyDrive/employee_attrition_.xlsx")
print("Dataset Shape:", df.shape)
print("\nSample Data:\n", df.head())
```

```
Dataset Shape: (200, 6)

Sample Data:
    MonthlyIncome   Age  JobSatisfaction  YearsAtCompany OverTime  Attrition
0        13191.0  41.0              4.0             1.0      Yes        1.0
1        15859.0  44.0              4.0            17.0      Yes        0.0
2         9014.0  28.0              1.0             NaN       No        NaN
3         4936.0   NaN              1.0             NaN       No        NaN
4         3885.0  27.0              4.0             0.0      Yes        1.0
```

Figure 3 :- Separate features and target from dataset

```
[ ]  # 3 Separate Features (X) and Target (y)
     X = df.drop("Attrition", axis=1)
     y = df["Attrition"]
```

```
[ ]  num_cols = X.select_dtypes(include=["int64", "float64"]).columns
     cat_cols = X.select_dtypes(include=["object"]).columns
```

Figure 4 :- imputes missing values with mean and label-encodes columns in dataset

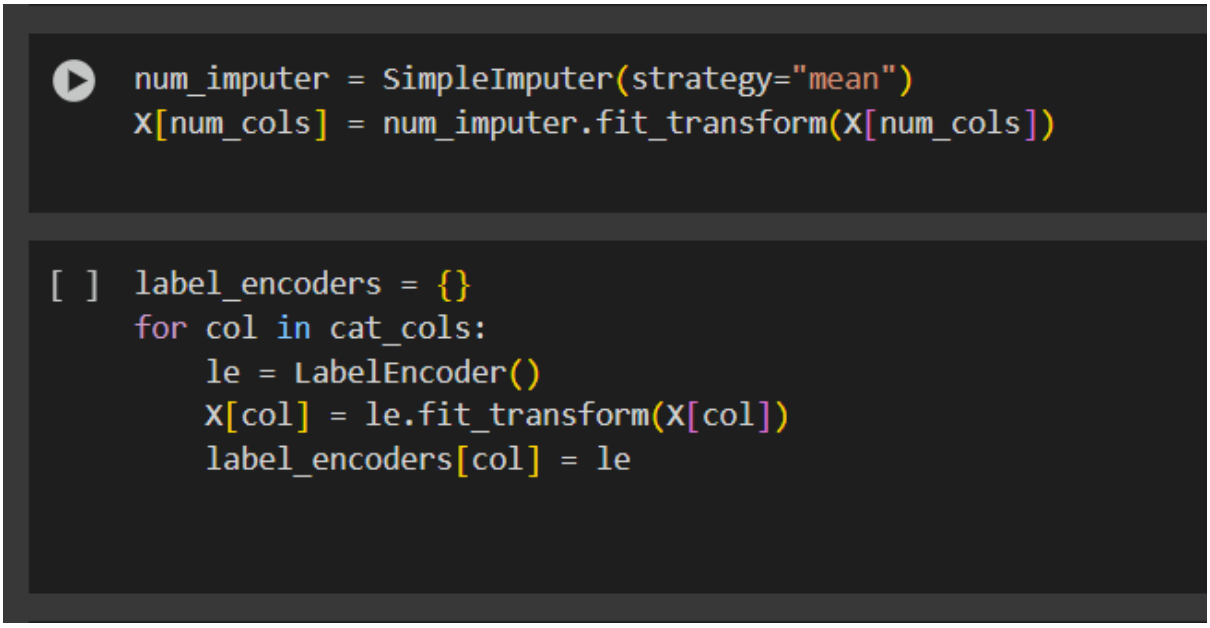


Figure 5 :- Using SimpleImputer to fill missing values in y with most frequent value

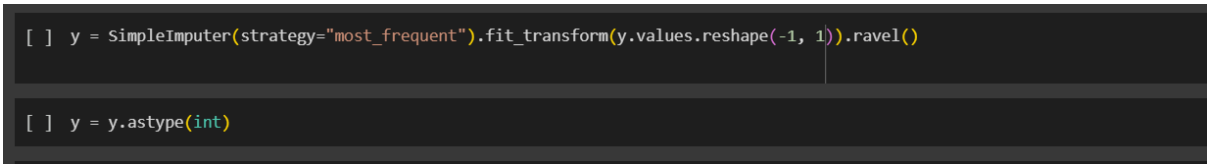


Figure 6 :- Splitting data into training and testing sets, then training a RandomForestClassifier with a fixed random state

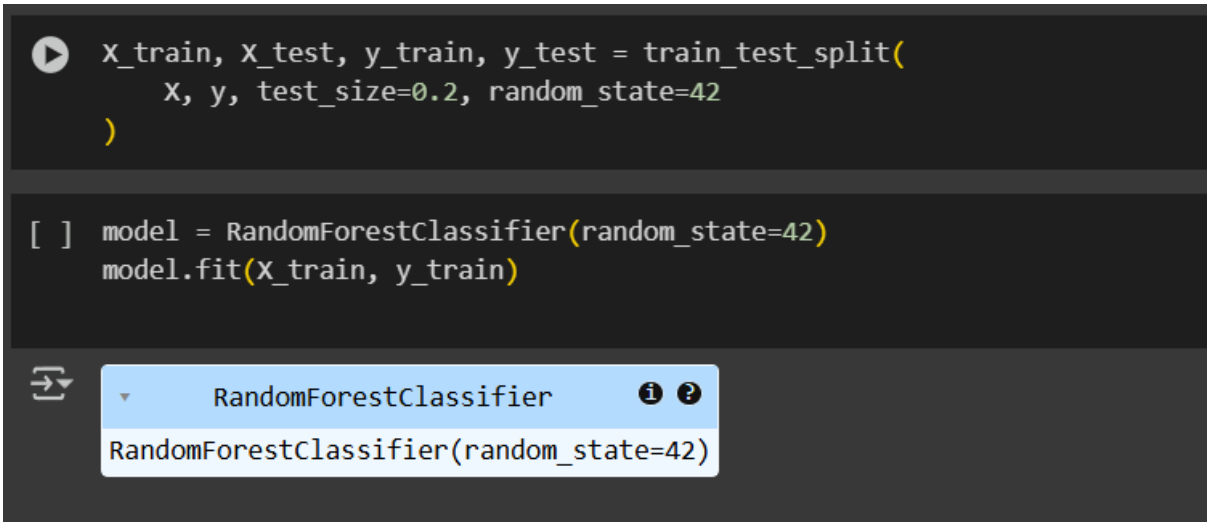
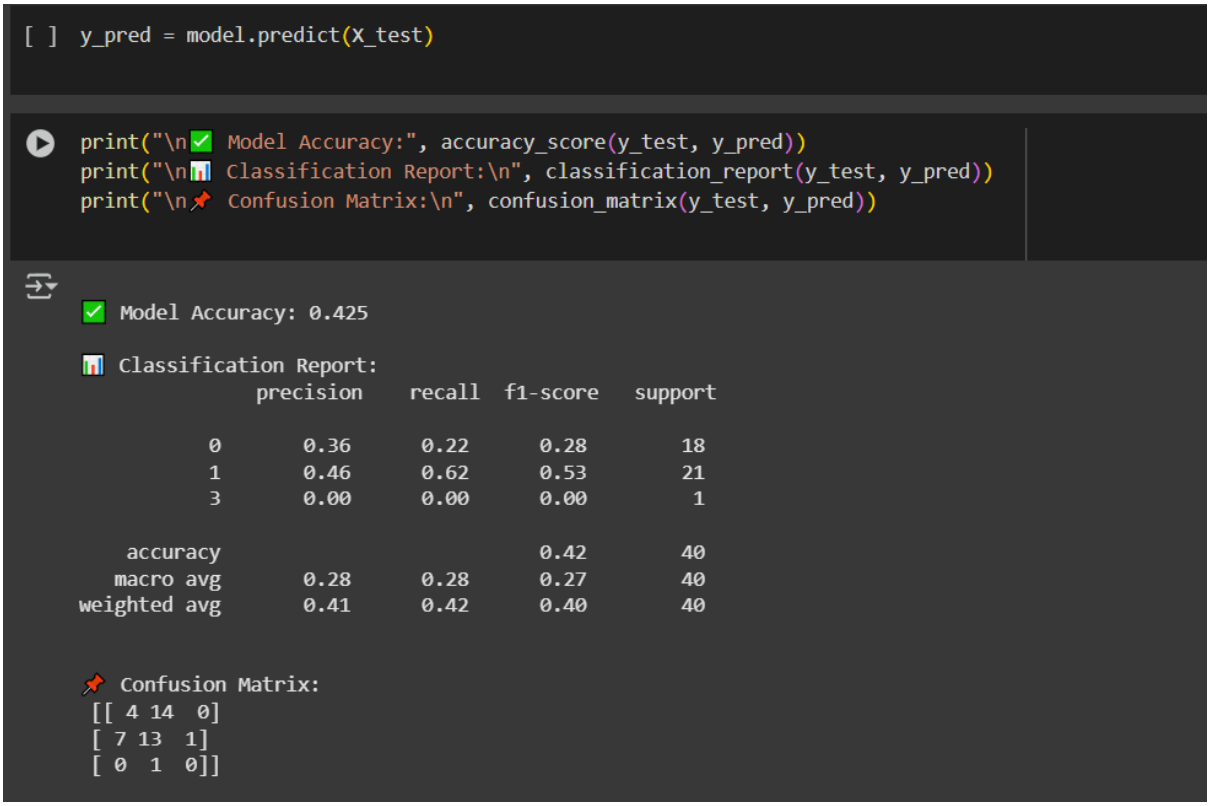


Figure 7 :- RandomForestClassifier's prediction results with 42.5% accuracy, detailed classification metrics, and a confusion matrix.



```
[ ]  y_pred = model.predict(X_test)
```

```
print("\n✅ Model Accuracy:", accuracy_score(y_test, y_pred))
print("\n📊 Classification Report:\n", classification_report(y_test, y_pred))
print("\n📌 Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
```

```
✅ Model Accuracy: 0.425

📊 Classification Report:
              precision    recall  f1-score   support

           0       0.36      0.22      0.28        18
           1       0.46      0.62      0.53        21
           3       0.00      0.00      0.00         1

    accuracy                           0.42        40
   macro avg       0.28      0.28      0.27        40
weighted avg       0.41      0.42      0.40        40


📌 Confusion Matrix:
 [[ 4 14  0]
 [ 7 13  1]
 [ 0  1  0]]
```

Figure 8 :- horizontal bar chart visualizes feature importance from a trained RandomForestClassifier model. The code imports `matplotlib.pyplot` and uses Pandas to create a `Series` of feature importances, sorts them, and plots them. The chart shows that **MonthlyIncome**, **Age**, and **YearsAtCompany** are the top three influential features, followed by **JobSatisfaction** and **OverTime**, with MonthlyIncome having the highest importance score.