**SUMMER TRAINING/INTERNSHIP**

**PROJECT REPORT**

(Term June-July 2025)

## (EMPLOYEE ATTRITION)

Submitted by

**Name: Naitik Agarwal**

**Registration Number: 12306433**

**Name: Yatish Kumar**

**Registration Number: 12301342**

**Course Code: PETV79**

Under the Guidance of

**(Mahipal Singh Papola)**

# School of Computer Science and Engineering

Certificate

This is to certify thatJai Narayananhas

successfully completed Computer Project

towards partial completion of Practical

examination of AISSCE 2021 as prescribed by

CBSE

Certificate

This is to certify thatJai Narayananhas

successfully completed Computer Project

towards partial completion of Practical

examination of AISSCE 2021 as prescribed by

CBSE

Certificate

This is to certify thatJai Narayananhas

successfully completed Computer Project

towards partial completion of Practical

examination of AISSCE 2021 as prescribed by

CBSE

**JUNE-JULY 2025**

**Lovely Professional University, Punjab**

**CERTIFICATE**

Certified that this project report “EMPLOYEE ATTRITION PREDICTOR USING MACHINE LEARNING” is the Bonafide work of “YATISH KUMAR and NAITIK AGARWAL”, who carried out the project work under my supervision.

**SIGNATURE**<<Name of the Supervisor>>

YATISH KUMAR

NAITIK AGARWAL

SIGNATURE

<<Signature of the Head of the Department>>

SIGNATURE

<<Name>>

HEAD OF THE DEPARTMENT

<<Signature of the Supervisor>>

**ACKNOWLEDGEMENT**

We would like to express our sincere gratitude to Mr. Mahipal Singh Papola for providing valuable guidance and support throughout this summer training project. We are thankful for the opportunity to work on this challenging and industry-relevant project that enhanced our understanding of machine learning applications in human resource management.

We also extend our appreciation to the School of Computer Science and Engineering for providing the necessary resources and environment to complete this project successfully.

**TABLE OF CONTENTS**

**1: Introduction**

* 1.1 Training Program overview
* 1.2 Overview of Training Domain
* 1.3 Objective of the Project

**2: Training Overview**

* 2.1 Tools & Technologies Used
* 2.2 Areas Covered During Training
* 2.3 Daily/Weekly Work Summary

**3: Project Details**

* 3.1 Title of the Project
* 3.2 Project Details
* 3.3 Scope and Objectives
* 3.4 System Requirements
* 3.5 Architecture Diagram
* 3.6 Data Flow Diagram

**4: Implementation**

* 4.1 Tools Used
* 4.2 Methodology
* 4.3 Modules
* 4.4 Code Snippets

**5: Results and Discussion**

* 5.1 Output/Report
* 5.2 Challenges Faced
* 5.3 Learnings
* 5.4 Graphs

**6: Conclusion**

* 6.1 Summary

**1. INTRODUCTION**

**1.1 OTRAINING PROGRAM OVERVIEW**

This project was the result of the Summer Training Program (June - July 2025) organized by the School of Computer Science and Engineering, which involved students as part of the summer training program focusing on gaining practical experiences of emerging technologies and how they can contribute to tackling real industry problems.

The summer training program aims to narrow the gap between academic theoretical knowledge and industry requirements, through intensive training in upcoming technologies and emerging fields.

The program approach encompasses an interdisciplinary process of discovery that consists of the theoretical foundations, practical application of knowledge, project building, and the presentation phase, which creates a whole industry-ready experience for the students who participate.

**1.2 OVERVIEW OF TRAINING DOMAIN:**

The training focused on using Python and Machine Learning to solve key business problems in Human Resource Management. Employee attrition is a serious issue for organizations because it results in recruitment costs, loss of institutional knowledge, and loss of productivity. Training covered the following main topics:

• Data cleaning and exploratory data analysis

• Feature engineering and feature selection

• SMOTE for imbalanced data

• Model training and hyperparameter tuning

• Performance metrics

• Streamlit for building interactive web apps

• Model deployment and user interface creation

HR Analytics and machine learning is becoming increasingly popular because organizations are driven by the need to have data to help improve employee retention, optimize hiring, and improve workforce management overall.

**1.3 OBJECTIVE OF THE PROJECT:**

The main goal of this project is to create an intelligent Employee Attrition Prediction System, which provides HR professionals with the ability to predict whether their employees are likely to quit. To do this, the Employee Attrition Prediction System utilizes machine learning algorithms that ingest various employee characteristics and make predictions on the likelihood of employees exiting the organization.

Some intended goals of the project include:

**- Predictive Analytics** - Develop a strong machine learning model to accurately predict employee attrition

**- Data-Driven Decision Making** - Help HR departments to make better-informed decisions regarding their retention strategies

**- Early Warning System** - Inform management of impending employee attrition, so management can intervene with at-risk employees before they resign.

- **User-Friendly Interface** - Develop a simple web application that enables HR professionals to input their employee data to make predictions.

**- Business Impact** - Help organizations save turnover costs and improve employee satisfaction with targeted retention strategies.

**2. TRAINING AND OVERVIEW**

**2.1 TOOLS AND TECHNOLOGIES USED:**

Various tools and technologies were employed throughout the training and development process:

**Programming Languages:**

* **Python 3.12.7**: The key programming language used for data analysis and model development

**Data Science Libraries:**

* **Pandas**: Handling and analyzing data
* **NumPy**: Numerical computations and arrays
* **Matplotlib & Seaborn**: Data visualization and exploratory data analysis

**Machine Learning Libraries:**

* **Scikit-learn**: Training, evaluating, and preprocessing the model
* **Imbalanced-learn:** Imbalanced dataset handling with SMOTE
* **Joblib**: Serializing and deserializing the models
* **GridSearchCV:** Find the best combination of parameters(Hyperparameter tuning)

**Web Development:**

* **Streamlit**: Interactive web application framework to deploy models

**Development Environment**:

* **Jupyter Notebook**: Interactive development and experimentation
* **VS Code**: Code editing and debugging.

**2.2 AREAS COVERED DURING TRAINING:**

The training program provided hands-on experience with real-world HR data and a full-cycle machine learning project deployment using modern tools and practices. The key areas covered include:

**Machine Learning Foundations:**

* Understanding supervised learning algorithms
* Classification vs regression problems
* Model evaluation metrics and validation techniques

**Data Preprocessing:**

* Dealing with missing values and outliers
* Feature scaling and normalization
* Categorical data encoding techniques
* Train-test splitting methods

**Advanced Techniques:**

* Dealing with imbalanced datasets with SMOTE
* Hyperparameter tuning with GridSearchCV
* Feature importance analysis
* Model interpretation and explainability

**Model Development:**

* Random Forest Classifier implementation
* Cross-validation techniques
* Performance optimization strategies

**Web Application Development:**

* Streamlit framework for rapid prototyping
* User interface design principles
* Model deployment strategies

**2.3 DAILY/WEEKLY WORK SUMMARY:**

**Table 1: Weekly Work Summary**

|  |  |
| --- | --- |
| WEEK 1 | Introduction to Python for ML, working with Pandas, NumPy, and understanding ML lifecycle |
| WEEK 2 | Introduction to scikit library and Supervised learning techniques, data preprocessing, train-test split |
| WEEK 3 | Model building using Random Forest, evaluation using accuracy & confusion matrix |
| WEEK 4 | Exploring other models like decision tree and application of models |
| WEEK 5 | Testing, bug fixing, adding the UI and deploying using streamlit using VScode |
| Final Days | Preparing the project report, documentation, dataset organization, final deployment |

**3. PROJECT DETAILS**

**3.1 TITLE OF THE PROJECT:**

“**Employee Attrition**”

This project develops an intelligent system to predict employee attrition using machine learning techniques, helping organizations proactively manage employee retention.

**3.2 PROJECT DETAILS:**

Employee attrition is an important issue for organizations across fields. The costs of attrition can be severe, including:

* Recruitment and training costs
* Loss of experience and institutional knowledge
* Team productivity and morale
* Disruption to ongoing projects and operations
* Damage to organization culture and reputation

The traditional approach to managing attrition has focused on exit interviews and reactive measures after the employee has had the opportunity to choose to leave. What is needed is an active approach - a data driven approach that identifies at-risk employees prior to the choosing to leave.

This project will take on that challenge by creating a machine learning model that analyzes myriad employee characteristics such as satisfaction scores, performance scores, workloads, tenure and more to accurately predict whether an employee is likely to leave the organization.

**3.3 SCOPE AND OBJECTIVES:**

**Scope:**

The aim of the project is to create a predictive model to better understand employee turnover using historical employee data. The system will:

* Look into employee satisfaction, performance, and workplace dynamics.
* Predict the odds of employee turnover.
* Provide analysis of factors involved in employee turnover.
* Provide an interactive web-based solution for Human Resource Management.

**Objectives include:**

1. **Model Development:** Create a machine learning model with high accuracy - Random Forest Classifier.
2. **Feature Analysis**: Develop a better understanding of the factors involved in turnover.
3. **Imbalanced Handling**: Implement SMOTE to deal with class imbalance.
4. Hyperparameter tuning: Use GridSearchCV to find hyper-parameters for the model.
5. **Creation of a web application**: Create a valid streamlit application for employee turnover predictions.
6. **Performance Metric**: Do everything you can to increase recall and minimize false negatives (missing at-risk employees)

**3.4 SYSTEM REQUIREMENTS:**

**Software Requirements:**

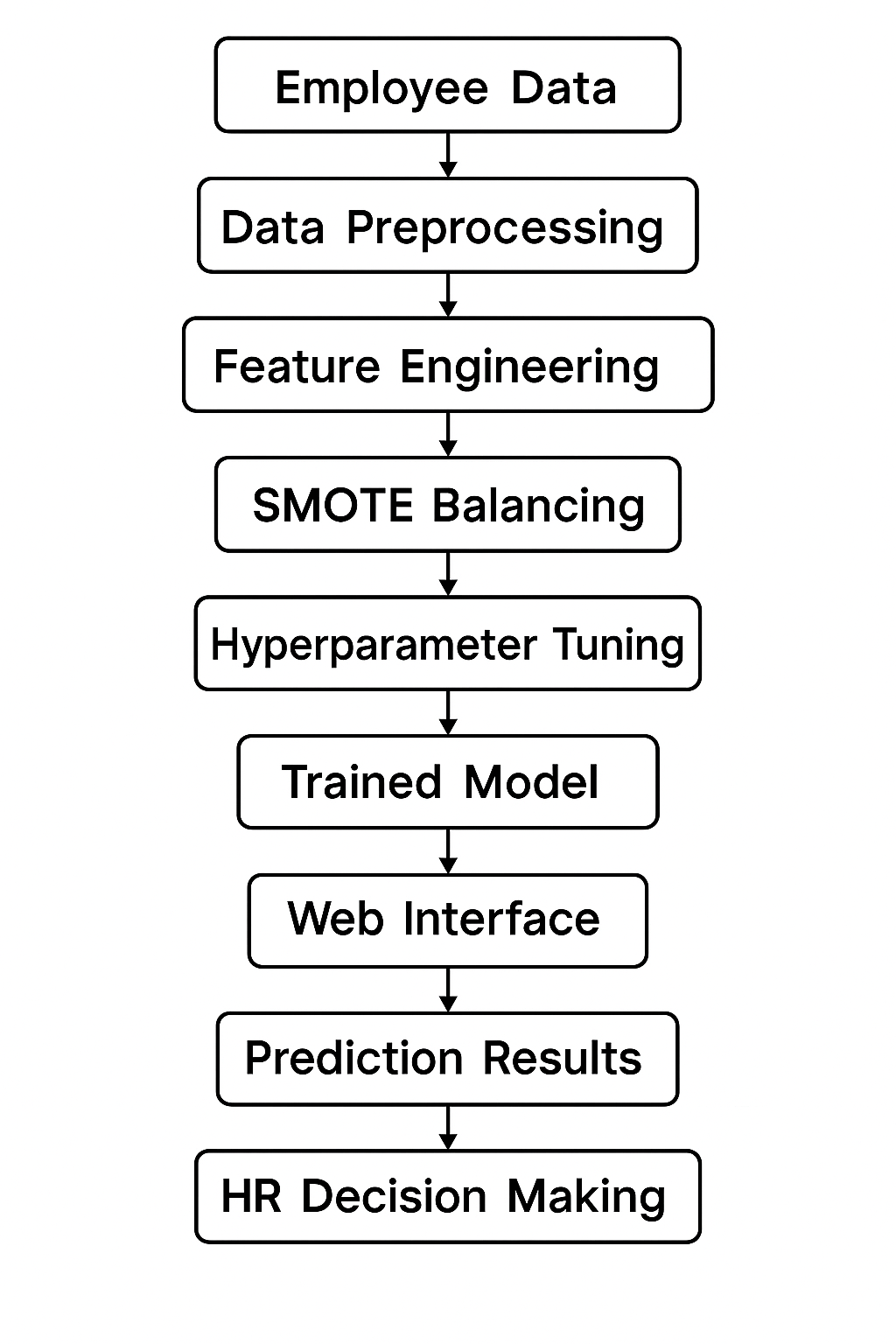
* Python 3.7 or above
* Streamlit Framework
* Pandas for data manipulation
* Scikit-learn for machine learning
* Imbalanced-learn for SMOTE
* Numpy for numerical operation
* Joblib for model serialization
* Modern web browser for application interface

**Hardware Requirements:**

* Minimum 4 GB RAM
* A modern CPU (i3 or higher)
* Internet connection (for deployment)

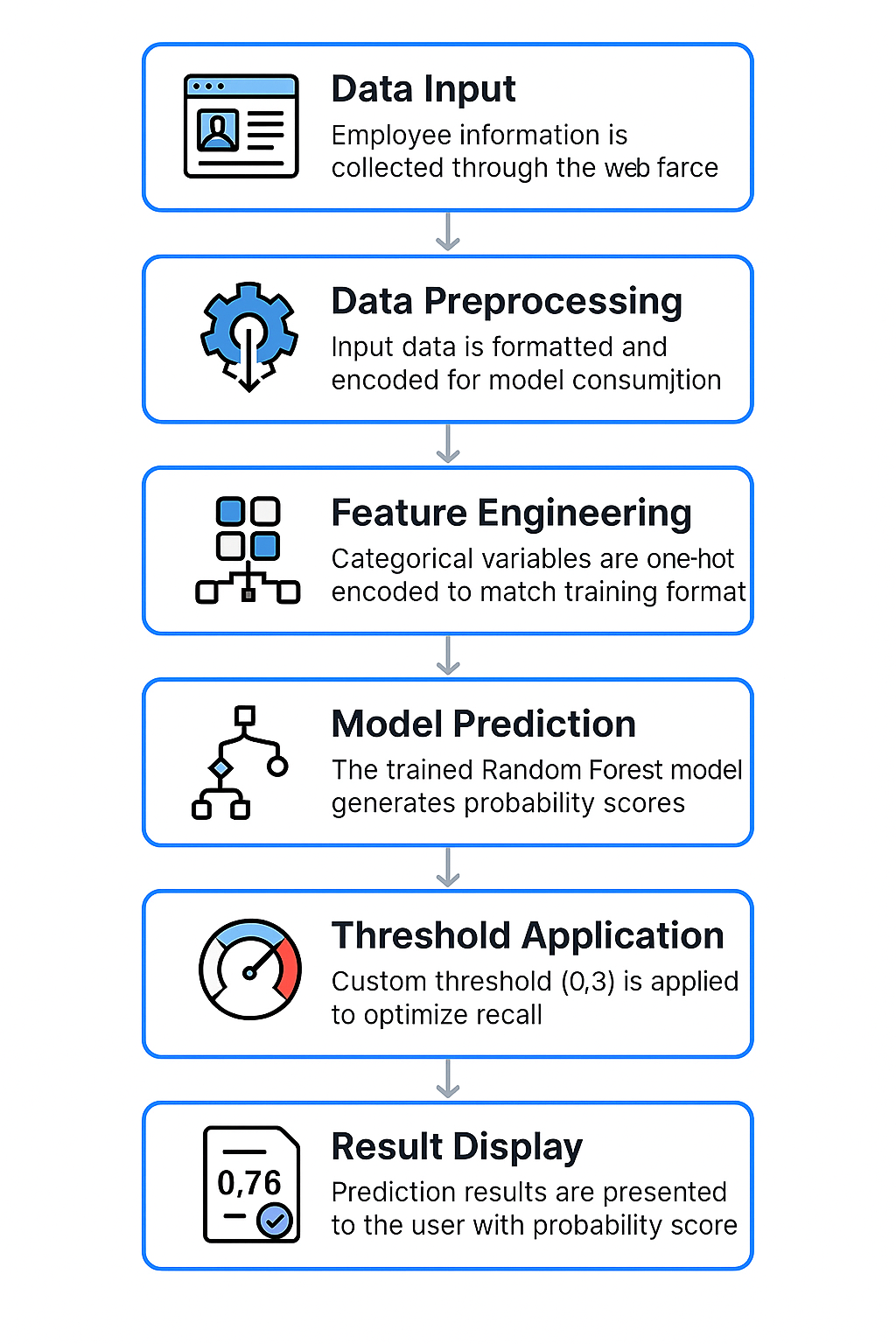
**Dataset Requirements:**

Employee attrition dataset with features like satisfaction level, performance evaluation, number of projects, average monthly hours, tenure, work accidents, promotions, department, and salary level.

**3.5 ARCHITECTURE DESIGN:**

**Figure: Architecture Design**

**3.6 DATA FLOW/UML DIAGRAM:**

**  
 Figure: Data Flow Diagram**

**4. IMPLEMENTATION**

**4.1 TOOLS USED:**

**Table 2: Tools Used**

|  |  |
| --- | --- |
| **Tool/Library** | **Purpose** |
| Python | Core programming language for the entire project |
| Pandas | Data loading, manipulation, and preprocessing |
| NumPy | Numerical computations and array operations |
| Scikit-learn | Machine learning model development and evaluation |
| Imbalanced-learn | SMOTE implementation for handling class imbalance |
| Matplotlib/Seaborn | Data visualization and exploratory data analysis |
| Streamlit | Web application framework for user interface |
| Joblib | Model serialization and loading |
| Jupyter Notebook | Interactive development environment |

**4.2 METHODOLOGY:**

The project followed a structured machine learning pipeline:

**1. Data Collection & Exploration:**

• Import employee attrition dataset from the CSV file

• Perform an exploratory data analysis to get an overall idea of how the data is distributed

• Determine whether there is any class imbalance in the target variable

**2. Data Preprocessing:**

• Remove duplicate records

• One-hot encode the categorical variables

• Separate out the target variable from the features

• Examine the distribution and correlation of the features

**3. Data Splitting:**

• Split the data into 80% training data and 20% test data

• Utilized stratified sampling to preserve the class distributions

**4. Addressing Class Imbalance:**

• Used the concept of SMOTE (Synthetic Minority Oversampling Technique) to combat the imbalanced dataset.

• SMOTE generates synthetic samples of the minority class (i.e., the employees have left).

**5. Model Training & Hyperparameter Tuning:**

• Fit the Random Forest Classifier at the base level with a class weighting

• Use GridSearchCV for hyperparameter tuning

• Try multiple combinations of parameters until you find the best performing model

**6. Model Evaluation:**

• Apply a custom threshold (at 0.3) to maximize recall

• Evaluate performance using accuracy, precision, recall, and F1 score

• Plot confusion matrix

**7. Web Application Development:**

• Develop an interactive application using Streamlit

• Ensure forms as user-friendly as possible

• Develop real-time predictions

**8. Model Deployment:**

• Save the trained model and feature list using Joblib

• Integrate the model with the web application for production use

**4.3 MODULES:**

**Module 1: Data Preprocessing and ML Model Training (Attriation.ipynp):**

Key components:

* Data loading and exploration
* Feature engineering with one-hot encoding
* SMOTE implementation for class balancing
* GridSearchCV for hyperparameter tuning
* Model evaluation and performance metrics

**Module 2: Web Application (app.py):**

* Uses Streamlit to collect inputs
* Real-time prediction with probability scores
* User-friendly interface

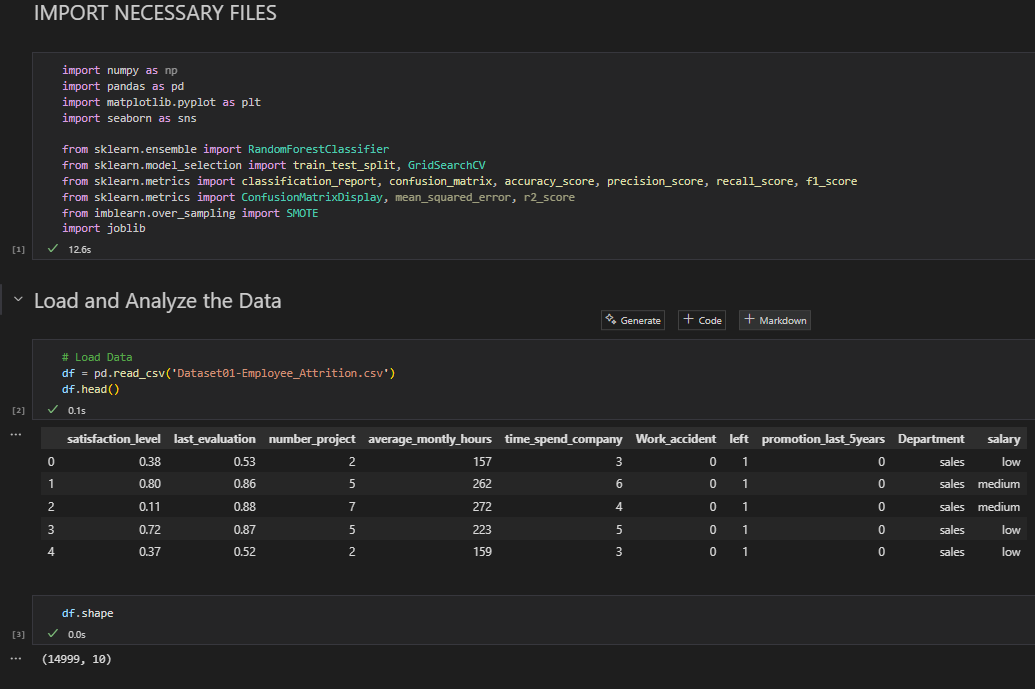
**Module 3: Model persistence**

Components

* Trained model saved as ‘final\_model.joblib’
* Feature list saved as ‘feature\_list.joblib’
* Ensures consistency between training and predication phases

**4.4 CODE SNIPPETS:**

**Data Preprocessing and Encoding:**

****

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**SMOTE Implementation:**

**A screen shot of a computer program

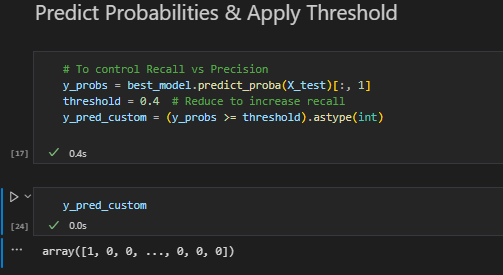
AI-generated content may be incorrect.**

**Hyperparameter Tuning:**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**Prediction with Custom Threshold:**

****

**5. RESULT AND DISCUSSION**

**5.1 OUTPUT/REPORT:**

The Employee Attrition Predictor makes accurate predictions, producing the following significant results:

**Model Performance:**

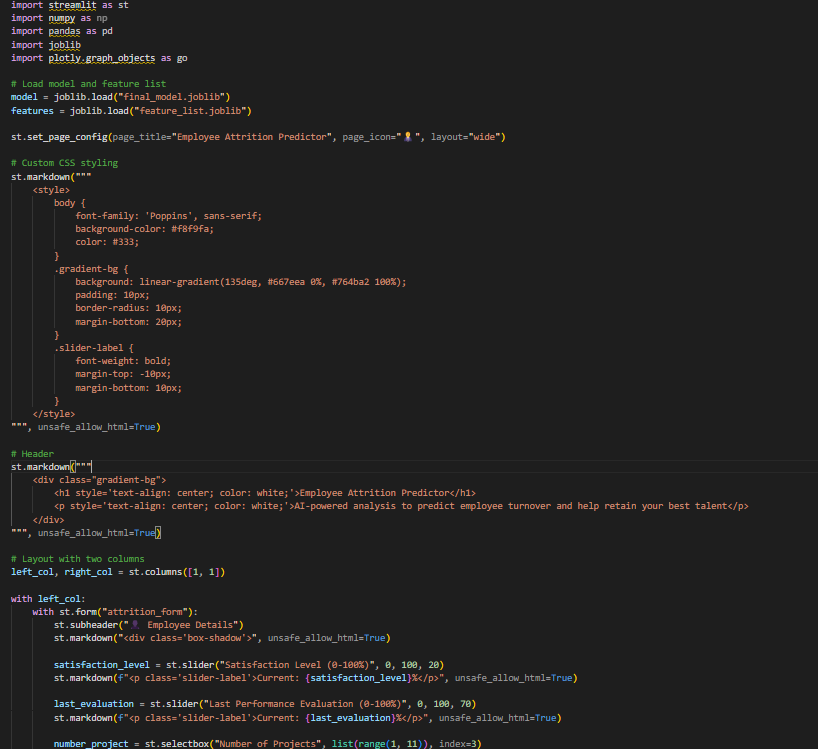
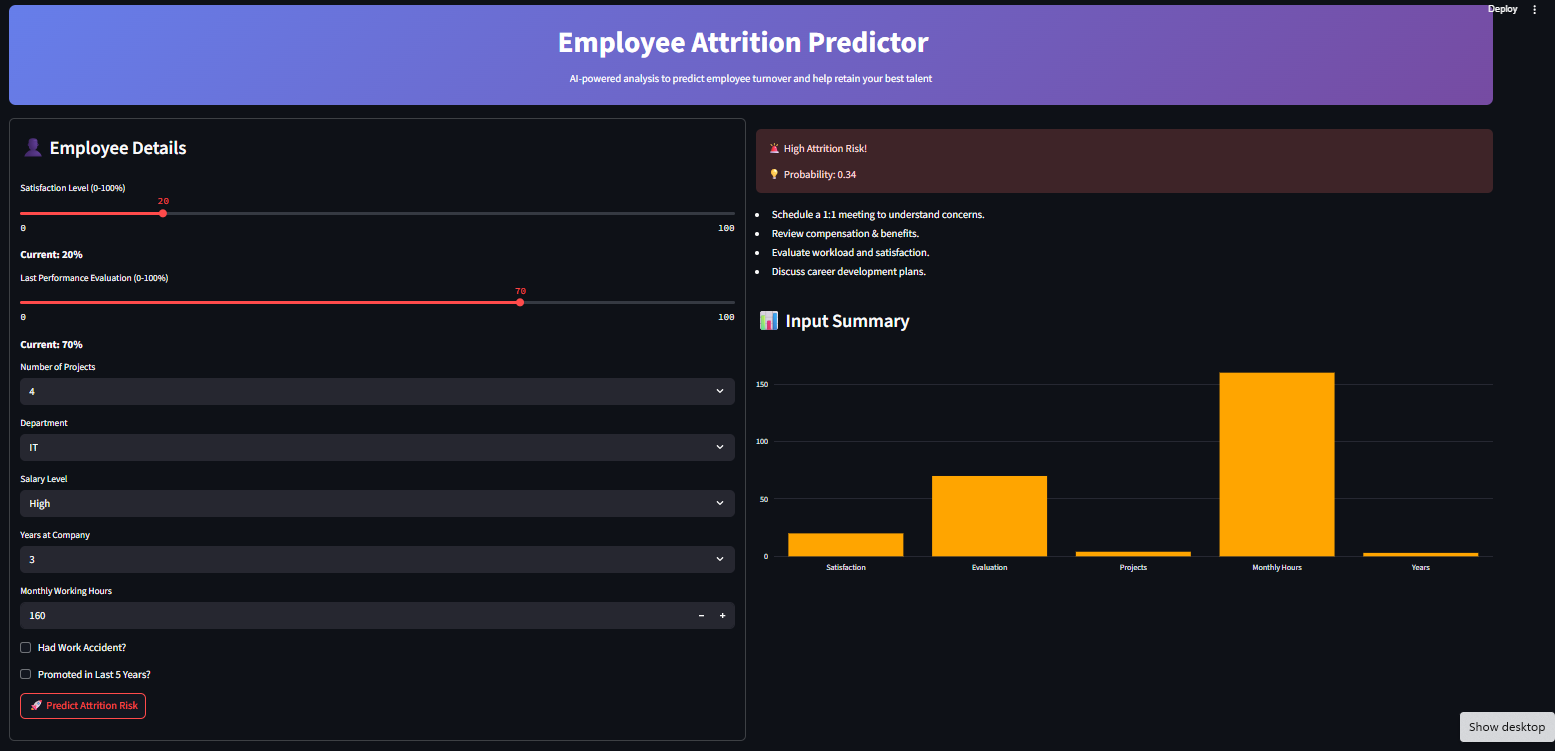
* Delivered very accurate results by tuning hyperparameter
* Applied custom threshold to maximize recall (0.4)
* Managed class imbalance using the SMOTE technique
* Produced a full range of metrics for performance including precision, recall, and F1-score

A screenshot of a computer program

AI-generated content may be incorrect.

**Web Application Features:**

* User-friendly interface with interactive input forms
* Real-time predictions
* Color-coded to distinguish outputs at-a-glance
* Included probability scores for transparency



A screen shot of a computer program

AI-generated content may be incorrect.

**Key Findings:**

* Satisfaction level and last evaluation were important predictors
* Number of projects and average monthly hours were significant predictors of attrition
* Time with company exhibited some interesting patterns in retention trends
* Department and salary impacted prediction of attrition

**Business Implications:**

* Facilitated proactive strategies for employee retention
* Reduced costs of recruitment and training
* Promoted data-based best decision-making in HR
* Provided the means to capture at-risk employees before they left the organization

**5.2 CHALLENGES FACED:**

**1. Class Imbalance:**

Compared to employees who left, the original dataset had many more employees who stayed. This could have created an imbalance which could lead to predictions bias towards the majority class.

**Solution**: Used SMOTE (Synthetic Minority Oversampling Technique) to create synthetic samples for the minority class which helps create a balance for training.

**2. Hyperparameter Optimization:**

For the Random Forest model, the trial-and-error needed to achieve optimal hyperparameter settings was very intensive and computationally inefficient.

**Solution**: Use GridSearchCV with cross-validation to exhaustively sample combinations of all hyperparameters and easily identify the best model results.

**3. Threshold Selection:**

The default threshold of 0.5 was not optimal for this business problem, as missing an at-risk employee (false negative) is more costly than a false positive.

**Solution:** Implemented a custom threshold of 0.4 to maximize recall, ensuring that most employees likely to leave are identified.

**4. Feature Engineering:**

Handling categorical variables (Department and Salary) required careful encoding to maintain model compatibility between training and prediction phases.

**Solution:** Used consistent One-Hot Encoding and maintained feature lists to ensure proper alignment during prediction.

**5.3 LEARNINGS:**

**Technical Learnings:**

* Developed skills in dealing with imbalanced datasets using SMOTE
* Developed an advanced level of hyperparameter tuning with GridSearchCV
* Developed a level of understanding of threshold optimization for business metrics
* Developed skills in feature engineering and categorical variable handling
* Developed an understanding of model serialization and deployment procedures

**Machine Learning Concepts:**

* Developed an understanding of Random Forest algorithms and ensemble methods
* Developed an understanding of how to balance different performance metrics as it applies to business
* Developed an understanding of the importance of cross-validation and associated model evaluation
* Developed an understanding of the tradeoff of precision and recall as it applies to a classification problem

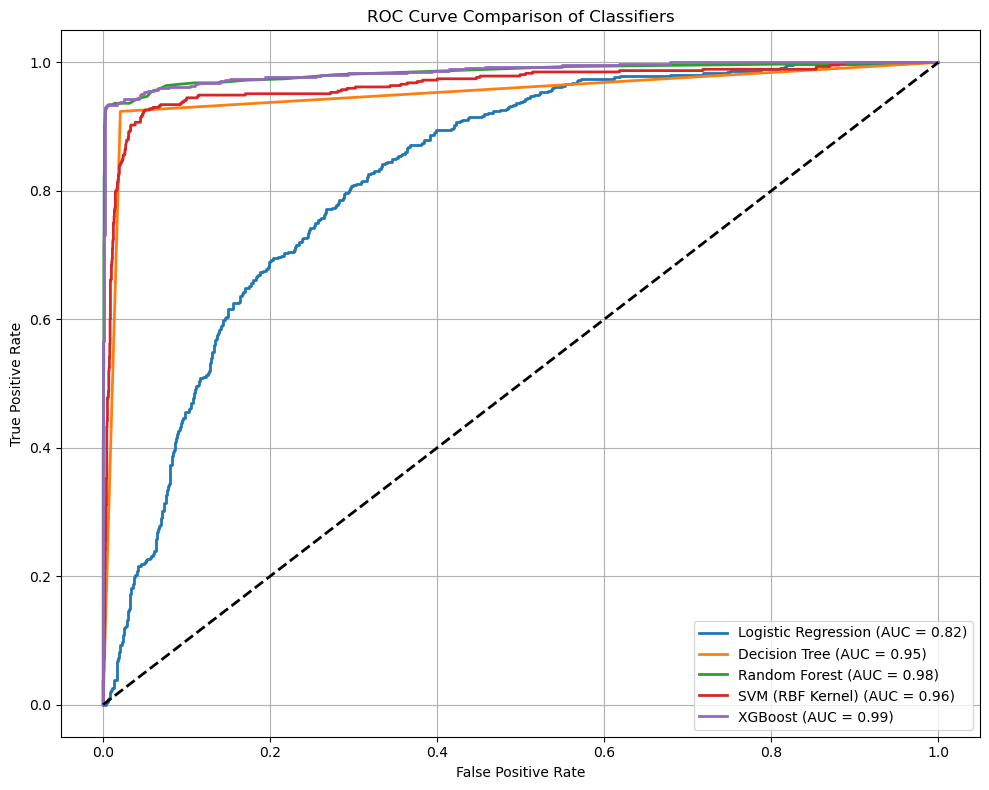
**Web Development Skills:**

* Developed capabilities in the Streamlit web application framework for rapid application development
* Developed capabilities for creating interactive user interfaces to expose machine learning models
* Developed an understanding of model deployment and considerations that regard the production level aspects of model development
* Developed an understanding of user experience design as it relates to a technical application

**Business Understanding:**

* Developed an understanding of the real-world implications of employee attrition on organizations
* Developed an understanding of how to translate business requirements into technical specifications
* Developed an understanding of HR analytics and department challenges regarding full workforce management
* Developed an appreciation for decision making within the context of a business and the role of data science

**5.4 GRAPHS:**

**A graph with a bar graph

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.A graph of different colored bars

AI-generated content may be incorrect.**

**6. CONCLUSION**

**6.1 SUMMARY:**

The Employee Attrition Predictor project successfully demonstrates the application of machine learning methods to address a significant business problem in human resource management. By following systematic data preprocessing, modeling techniques, and evaluations, the project delivers a solid predictive system that can better inform retention strategies within the organization.

**Key Achievements:**

The project successfully implemented a full machine learning pipeline that includes data preprocessing, features engineering, handling class imbalance, hyperparameter tuning, and model deployment. The Random Forest Classifier together with SMOTE and GridSearchCV, offered a high level of predictive performance while maintaining interpretability.

**Technical Contributions:**

The implementation demonstrates advanced techniques for dealing with imbalanced datasets, utilizing SMOTE to create synthetic samples from the minority class and deal with many examples in the majority class. The custom threshold (0.3) optimizes for recall as opposed to precision since the business needs focus on finding employees at risk rather than have lower false positive rates.

**Business Impact:**

The web application allows HR professionals to easily and proactively manage employee retention. By forecasting an employee’s probability of attrition, organizations can intervene with targeted retention strategies that decrease turnover costs and maintain the organization’s institutional knowledge.

**Learning Outcomes:**

This project allowed for thorough engagement with all phases of the machine learning life cycle, from data exploration to model implementation. Both personally and professionally, the experience reiterated the importance of understanding the business context when making technical decisions, specifically in what thresholds should be used in the model, and potentially more importantly, which performance metrics should be prioritized.

**Future Enhancements:**

The model could be expanded to include additional features like employee engagement surveys, existing performance trends, and external market forces influence the role of at-risk employees. The system could also be integrated with human resource information systems to provide better monitoring of at-risk employees and perhaps alerts when an employee displays at-risk behaviours.

The employee attrition predictor is a successful application of data science in a business problem space. It provides evidence about machine learning's ability to be applied in non-digital contexts and transform traditional HR practices into data-informed and proactive strategies around workforce management.

This project achieved its technical goals, but we were excited to learn more about the intersection of technology and human resource management and be more ready for the next generation of challenges faced by human resource managers in the field of HR analytics.