Vivekanand Education Society's
Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai, Approved by AICTE & Recognised by Govt. of Maharashtra)
NAAC accredited with 'A' grade
V.E.S. Since 1962

AI & DS - I Review 1

Title of the Project: Attrition & Salary Insights

Domain: AI/ML

Group Members:
Member 1: Aditya Kirtane (26)
Member 2: Sairam Konar (27)
Member 3: Bhushan Kor (28)

Mentor Name: Dr. Ravita Mishra

Content

- Problem Statement
- Objectives
- Literature Survey
- Proposed System
- Algorithms
- Implementation
- Result Analysis
- Conclusion
- References

Problem Statement

Employee attrition and fraud detection are critical challenges in workforce management. High employee turnover leads to increased hiring costs, loss of knowledge, and disruptions in workflow, while fraudulent activities such as salary anomalies and invalid documentation can cause financial losses. Organizations need an AI-driven solution to predict employee attrition based on key performance indicators (KPIs) like salary hikes, tenure, and work experience, as well as detect fraudulent patterns in salary and documentation. This project aims to build a predictive model using Random Forest and Isolation Forest algorithms to analyze employee data and identify high-risk attrition cases and fraud anomalies.

Objectives of the project

- **Attrition Prediction:** Develop a machine learning model to predict whether an employee is likely to leave based on salary trends, work experience, and promotion history.

- **Fraud Detection:** Identify employees with suspicious salary hikes and document inconsistencies using anomaly detection techniques.

- **Feature Analysis:** Evaluate the key factors influencing employee attrition and fraud using correlation analysis and feature importance rankings.

Literature Survey

Sr. No.	Title of Technical Paper	Name of Author	Year of Publication	Name of Journal	Methodology	Results/ Conclusions	Drawbacks/ Limitations
1	Predicting Employee Attrition Using Machine Learning Techniques[1]	Francesca Fallucchi, Marco Coladangelo, Romeo Giuliano, Ernesto William De Luca	2020	Computers, MDPI	Real-world HR dataset (IBM, ~1500 entries, 35 features); applied multiple ML classification models: Gaussian Naive Bayes, Logistic Regression, KNN, Decision Tree, Random Forest, SVM. Dataset was split (70:30), feature engineering & cross-validation used.	Gaussian Naive Bayes gave the best recall (0.541) and lowest false negatives (4.5%). The model helps HR make proactive, data-driven decisions to reduce attrition.	Limited dataset size; imbalance in attrition classes; precision was sacrificed to maximize recall. Results may improve with larger and more recent datasets.

Literature Survey

Sr. No.	Title of Technical Paper	Name of Author	Year of Publication	Name of Journal	Methodology	Results/ Conclusions	Drawbacks/ Limitations
2	Improving Salary Offer Processes With Classification Based Machine Learning Models[2]	Rukiye Kaya, Mehtap Saatçi, Mehmet Gökhan Bakal	2024	IEEE – 8th Int'l AI & Data Processing Symposium (IDAP'24)	Dataset from a private firm (1820 samples, 11 selected features); 7 classification models + 5 ensemble models tested for predicting salary categories; used correlation, K-Best, and mutual info for feature selection.	ANN achieved highest accuracy (58.2%), followed by KNN (56.8%). Hard Voting Ensemble reached 59.3% accuracy. The system provides fair, data-driven salary suggestions.	Accuracy limited due to class imbalance and unknown salary category mapping; some classes had very low representation. Results could improve with more balanced and enriched datasets.

Proposed System

- **Data Collection & Processing** – Gather, clean, and preprocess HR data with feature engineering.
- **Attrition Prediction** – Use machine learning (Random Forest) to predict employee attrition.
- **Fraud Detection** – Identify salary anomalies and fraudulent activities using Isolation Forest.
- **Model Deployment & Dashboard** – Deploy models in a web-based HR dashboard for real-time insights.

Algorithms

1) Random Forest algorithm vs XGBoost Algorithm

- Random Forest is an ensemble learning algorithm that builds multiple decision trees and averages their predictions (for regression) or uses majority voting (for classification) to improve accuracy and reduce overfitting.

- XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting algorithm that sequentially improves weak learners (decision trees) by minimizing errors using boosting techniques.

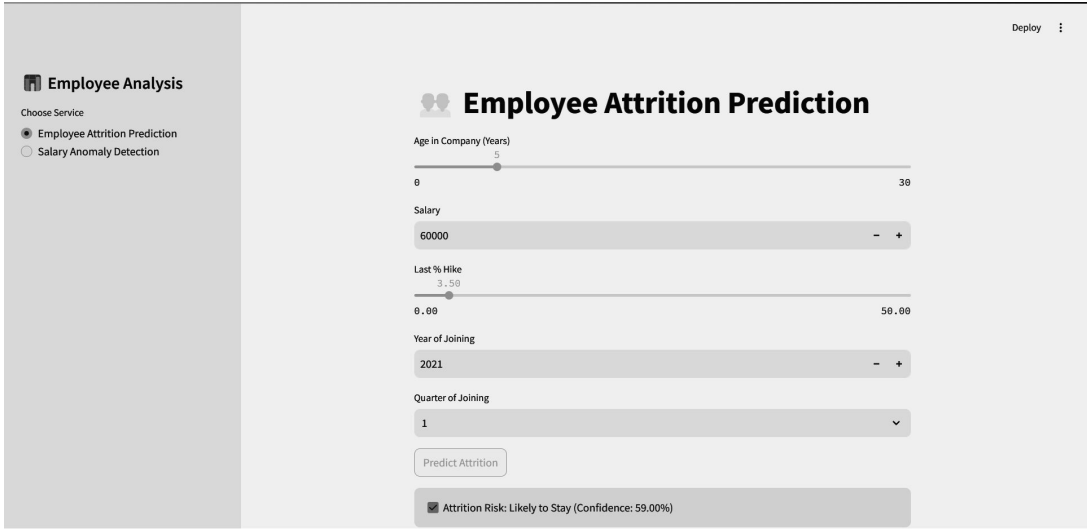
Algorithms

2) Isolation Forest Algorithm

- Isolation Forest is an anomaly detection algorithm that works by randomly selecting a feature and splitting the data recursively to isolate outliers.

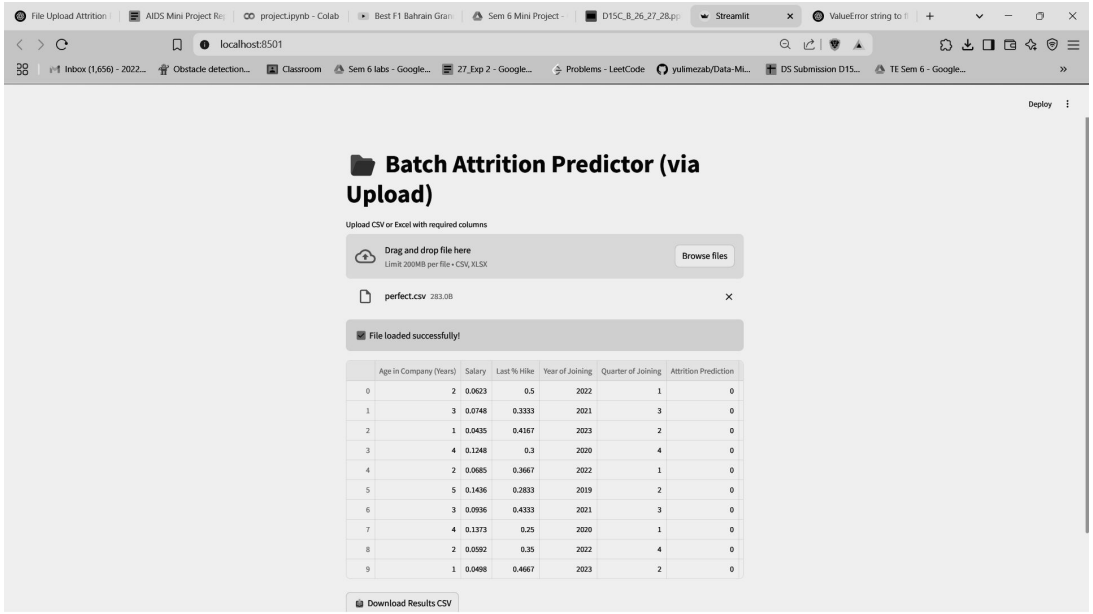
- Anomalies (fraudulent or rare data points) are isolated faster with fewer splits than normal instances.

Implementation



Employee Attrition Prediction

Implementation



	Age in Company (Years)	Salary	Last % Hike	Year of Joining	Quarter of Joining	Attrition Prediction
0	2	0.0623	0.5	2022	1	0
1	3	0.0748	0.3333	2021	3	0
2	1	0.0435	0.4167	2023	2	0
3	4	0.1248	0.3	2020	4	0
4	2	0.0685	0.3667	2022	1	0
5	5	0.1436	0.2833	2019	2	0
6	3	0.0936	0.4333	2021	3	0
7	4	0.1373	0.25	2020	1	0
8	2	0.0592	0.35	2022	4	0
9	1	0.0498	0.4667	2023	2	0

Employee Attrition (Upload file)

Result Analysis

Decision Tree Accuracy: 0.8331

	precision	recall	f1-score	support
0	0.99	0.75	0.85	129462
1	0.68	0.99	0.81	70538
accuracy			0.83	200000
macro avg	0.84	0.87	0.83	200000
weighted avg	0.88	0.83	0.84	200000

Logistic Regression Accuracy: 0.6995

	precision	recall	f1-score	support
0	0.72	0.88	0.79	129462
1	0.63	0.36	0.46	70538
accuracy			0.70	200000
macro avg	0.67	0.62	0.63	200000
weighted avg	0.69	0.70	0.67	200000

Random Forest Accuracy: 0.8010

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.88	0.80	0.84	129462
1	0.69	0.80	0.74	70538
accuracy			0.80	200000
macro avg	0.78	0.80	0.79	200000
weighted avg	0.81	0.80	0.80	200000

Random Forest ROC-AUC Score: 0.8015

XGBoost Accuracy: 0.833415

	precision	recall	f1-score	support
0	1.00	0.74	0.85	129462
1	0.68	1.00	0.81	70538
accuracy			0.83	200000
macro avg	0.84	0.87	0.83	200000
weighted avg	0.89	0.83	0.84	200000

Conclusion

- The Isolation Forest Algorithm is essential for detecting fraud in financial transactions by identifying anomalies efficiently.
- Predicting employee attrition using machine learning helps organizations retain talent and reduce turnover costs.
- Both applications enhance decision-making by providing data-driven insights for financial security and workforce management.
- Implementing these models ensures business stability, reduces risks, and improves overall organizational efficiency.

References

[1] Predicting Employee Attrition Using Machine Learning Techniques

[2]Improving Salary Offer Processes With Classification Based Machine Learning Models

[3] Dataset - https://excelbianalytics.com/wp/downloads-21-sample-csv-files-data-sets-for-testing-till-5-million-records-hr-analytics-for-attrition/

Publications (if any)