GROUP 8

IBM-HR-ANALYTICS
EMPLOYEE ATTRITION & PERFORMANCE PREDICTION

Artificial Intelligence

TABLE OF CONTENTS

- Introduction & Recap of Briefing 1
- Data & Survival Problem Formulation 2
- Modeling Pipeline & Algorithms 3
- Quantitative Results 4
- Risk Stratification & HR Use Cases 5
- Limitations & Future Work 6
- Conclusion 7

MEET THE TEAM

Jenny - 1137131

Tee - 1137168

Harley - 1137158

INTRODUCTION & RECAP OF BRIEFING

INTRODUCTION

- **HR today:** hiring, development, pay, retention → data scattered, late signals, costly turnover.
- **Pain points:** manual reporting, missed early signs, inconsistent decisions, tight budgets.
- **We propose:** an AI co-pilot + time-to-event view to act before resignations.
- **Objectives:**
  a. Understand drivers of attrition.
  b. Estimate risk over time (3/6/12 months).
- **Research questions:** who is at risk soon? which factors shorten/extend tenure? how to guide targeted, cost-savingactions?

RECAP OF PREVIOUS BRIEFING

1. DATA & VARIABLES

IBM HR Analytics.

2. PLANNED APPROACH:

Treat attrition as a time-to-event problem using survival analysis; also add a classification model and Explainable AI.

3. EXPECTED OUTPUTS

- Survival curves for key groups,
- Comparison of survival models,
- Risk stratification table for 3, 6 and 12 months,
- Overfitting analysis and SHAP explanations.

WHAT THE FINAL PRESENTATION WILL COVER?

- Data & survival formulation.
- Modeling pipeline (Cox, Random Survival Forest, XGBoost + SHAP).
- Evaluation results and overfitting analysis.
- HR insights, limitations, and future directions.

DATA & SURVIVAL PROBLEM FORMULATION

DATASET & PREPARATION

BM HR Analytics Employee Attrition and Performance.

1,470 employees, 35 variables:

- **Demographics**: Age, Gender, MaritalStatus, etc.
- **Job-related:** Department, JobRole, BusinessTravel, OverTime, JobLevel, etc.
- **Tenure-related:** YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, etc.

SURVIVAL PROBLEM FORMULATION

Define event:

- event = 1 if Attrition = “Yes”
- event = 0 if Attrition = “No” (censored observation).

Define duration:

- duration = YearsAtCompany × 12 → tenure in months.

Explain concept of censoring:

- i. Employees who have not left yet at the end of observation are censored, but still provide information about survival time.

RESEARCH QUESTIONS

- **Q1:** Which employees are at highest risk of leaving in the next 3, 6, and 12 months?
- **Q2:** Which factors are most strongly associated with shorter or longer tenure?
- **Q3:** How can these models support HR decision-making?

MODELING PIPELINE & ALGORITHMS

MODELING PIPELINE & ALGORITHMS

Technical Setup & Preprocessing

- Python 3.11 + Conda env ai_hr_survival, VS Code
- **Main libraries:** pandas, matplotlib, lifelines, sksurv, xgboost, shap
- **Data after preprocessing:** 1,470 employees × 46 features
- **Drop non-informative columns:** EmployeeNumber, EmployeeCount, Over18, StandardHours,...
- **One-hot encoding for categorical vars** (BusinessTravel, Department, JobRole, MaritalStatus, OverTime,...)
- **Train-test split:** 70% / 30%, random_state = 42
- **Survival formulation:**

  - **Event** = **1 if Attrition** = **"Yes", else 0** (censored).

  - **Duration** = **YearsAtCompany** (in months)

  - **Cox sample:** 1,029 obs – 176 events, 853 censored

```
from lifelines import KaplanMeierFitter, CoxPHFitter
from lifelines.utils import concordance_index

from sklearn.model_selection import train_test_split
from sksurv.ensemble import RandomSurvivalForest
from sksurv.metrics import (
```

```
drop_cols = ["EmployeeNumber", "Attrition", "EmployeeCount", "Over18", "StandardHours"]
drop_cols = [c for c in drop_cols if c in df.columns]
data = df.drop(columns=drop_cols)
```

```
<lifelines.CoxPHFitter: fitted with 1029 total observations, 853 right-censored observations>
             duration col = 'duration'
                event col = 'event'
      baseline estimation = breslow
   number of observations = 1029
number of events observed = 176
   partial log-likelihood = -659.54
         time fit was run = 2025-12-12 07:43:33 UTC
```

SURVIVAL MODELS

Cox Proportional Hazards (lifelines)

- Linear log-hazard model, interpretable baseline
- Uses all encoded features, ***duration*** and ***event***

```
cph = CoxPHFitter()
cph.fit(train_df, duration_col="duration", event_col="event")
cph.print_summary()

pred_partial_haz_train = cph.predict_partial_hazard(train_df)
pred_partial_haz_test = cph.predict_partial_hazard(test_df)
```

Random Survival Forest (sksurv)

- Ensemble of survival trees, captures non-linear effects & interactions
- ~300 trees, **max_features = "sqrt"**
- Conservative **min_samples_split** and **min_samples_leaf** to control overfitting

```
rsf = RandomSurvivalForest(
    n_estimators=300,
    min_samples_split=10,
    min_samples_leaf=5,
    max_features="sqrt",
    n_jobs=-1,
    random_state=42,
```

```
ci_rsf_train = concordance_index_censored(
    (1 - yevent_train).astype(bool),
    ytime_train,
    rsf.predict(X_train),
)[0]

ci_rsf_test = concordance_index_censored(
    (1 - yevent_test).astype(bool),
    ytime_test,
    rsf.predict(X_test),
)[0]
```

XGBOOST & SHAP

1. **Binary target: Attrition (Yes / No).**

```
y_clf = y_event

Xtr, Xte, ytr, yte = train_test_split(X, y_clf, test_size=0.3, random_state=42)

xgb_model = xgb.XGBClassifier(
    objective="binary:logistic",
    eval_metric="logloss",
    n_estimators=200,
    max_depth=4,
    learning_rate=0.05,
    subsample=0.8,
    colsample_bytree=0.8,
    random_state=42,
)
```

2. **Monitor train vs test logloss across boosting rounds**

```
results = xgb_model.evals_result()
train_logloss = results["validation_0"]["logloss"]
test_logloss = results["validation_1"]["logloss"]
epochs = range(1, len(train_logloss) + 1)
```

3. **Provides a second view on overfitting behaviour**

4. **Explainable AI:**

- Use TreeExplainer to compute SHAP values
- SHAP summary plot ranks features by average impact on attrition probability

```
explainer = shap.TreeExplainer(xgb_model)
shap_values = explainer.shap_values(Xte)

plt.figure()
shap.summary_plot(shap_values, Xte, show=False)
plt.tight_layout()
plt.savefig("shap_summary.png", bbox_inches="tight")
plt.close()
```

QUANTITATIVE RESULTS & OVERFITTING ANALYSIS

Cox Model: Performance & Fit

```
Concordance = 0.98
Partial AIC = 1407.08
log-likelihood ratio test = 905.08 on 44 df
-log2(p) of ll-ratio test = 533.02

C-index Cox train: 0.983
C-index Cox test : 0.974
```

Cox Model: Key Effects

covariate	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%
Age	-0.02	0.98	0.01	-0.05	0.00	0.95	1.00
DailyRate	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00
DistanceFromHome	0.02	1.02	0.01	0.00	0.04	1.00	1.04
Education	0.06	1.06	0.08	-0.10	0.22	0.91	1.24
EnvironmentSatisfaction	-0.26	0.77	0.08	-0.41	-0.11	0.66	0.90
HourlyRate	0.00	1.00	0.00	-0.01	0.01	0.99	1.01
JobInvolvement	-0.26	0.77	0.11	-0.48	-0.05	0.62	0.95
JobLevel	0.20	1.22	0.29	-0.36	0.76	0.70	2.15
JobSatisfaction	-0.29	0.75	0.07	-0.43	-0.15	0.65	0.86
MonthlyIncome	0.00	1.00	0.00	-0.00	0.00	1.00	1.00
MonthlyRate	0.00	1.00	0.00	-0.00	0.00	1.00	1.00
NumCompaniesWorked	0.19	1.21	0.03	0.12	0.25	1.13	1.29
PercentSalaryHike	-0.00	1.00	0.03	-0.07	0.06	0.93	1.07
PerformanceRating	-0.13	0.88	0.36	-0.82	0.57	0.44	1.77
RelationshipSatisfaction	-0.17	0.85	0.08	-0.32	-0.02	0.73	0.98
StockOptionLevel	-0.11	0.89	0.14	-0.40	0.17	0.67	1.19
TotalWorkingYears	-0.07	0.93	0.03	-0.13	-0.02	0.88	0.98
TrainingTimesLastYear	-0.12	0.89	0.07	-0.25	0.02	0.78	1.02
WorkLifeBalance	-0.22	0.80	0.11	-0.45	-0.00	0.64	1.00
YearsAtCompany	-1.14	0.32	0.13	-1.39	-0.88	0.25	0.41
YearsInCurrentRole	-0.16	0.85	0.05	-0.26	-0.06	0.77	0.94
YearsSinceLastPromotion	0.11	1.12	0.04	0.03	0.20	1.03	1.22
YearsWithCurrManager	-0.10	0.91	0.05	-0.20	0.00	0.82	1.00
BusinessTravel_Travel_Frequently	1.42	4.13	0.41	0.60	2.23	1.83	9.30
BusinessTravel_Travel_Rarely	0.90	2.46	0.39	0.13	1.67	1.14	5.32
Department_Research & Development	2.81	16.58	3.17	-3.41	9.02	0.03	8301.42
Department_Sales	3.17	23.72	3.19	-3.09	9.42	0.05	12367.73
EducationField_Life Sciences	-0.75	0.47	0.79	-2.31	0.80	0.10	2.22
EducationField_Marketing	-0.32	0.72	0.82	-1.94	1.29	0.14	3.63
EducationField_Medical	-0.52	0.60	0.79	-2.06	1.02	0.13	2.78
EducationField_Other	-0.51	0.60	0.83	-2.15	1.12	0.12	3.07
EducationField_Technical Degree	-0.14	0.87	0.80	-1.72	1.43	0.18	4.19
Gender_Male	0.20	1.23	0.17	-0.13	0.53	0.88	1.71
JobRole_Human Resources	4.69	109.03	3.27	-1.71	11.09	0.18	65710.94
JobRole_Laboratory Technician	2.09	8.05	0.62	0.88	3.30	2.40	27.03
JobRole_Manager	1.13	3.09	0.94	-0.72	2.98	0.49	19.60
JobRole_Manufacturing Director	1.06	2.88	0.66	-0.24	2.36	0.79	10.55
JobRole_Research Director	-0.23	0.79	1.06	-2.31	1.85	0.10	6.39
JobRole_Research Scientist	1.48	4.39	0.63	0.25	2.71	1.29	15.00
JobRole_Sales Executive	1.16	3.20	1.13	-1.06	3.38	0.35	29.48
JobRole_Sales Representative	2.07	7.93	1.17	-0.23	4.37	0.80	78.92
MaritalStatus_Married	0.45	1.57	0.29	-0.11	1.02	0.89	2.76
MaritalStatus_Single	1.08	2.95	0.35	0.39	1.78	1.47	5.91
OverTime_Yes	1.31	3.72	0.16	0.99	1.64	2.69	5.13

Random Survival Forest Results

Evaluation on same split as Cox

- **C-index:**
  - Train: 0.714
  - Test: 0.704

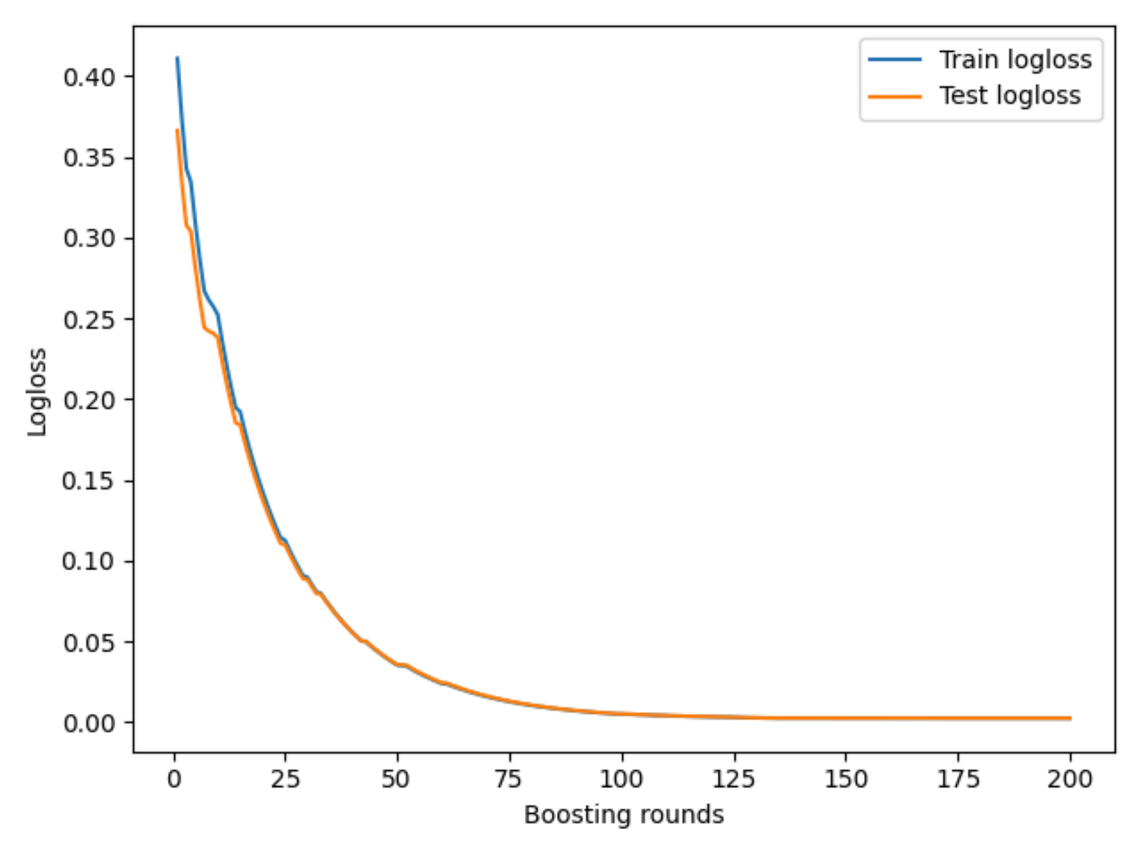
```
C-index RSF train: 0.714
C-index RSF test : 0.704
```

- **Interpretation:**
  - Lower C-index than Cox, but train ≈ test
  - RSF generalizes well, captures non-linear patterns with modest overfitting

- **Time-dependent AUC(t) & Brier at 12/24/36 months**
  - Could not be computed stably (censoring survival function = 0)
  - Use C-index as main performance metric

QUANTITATIVE RESULTS & OVERFITTING ANALYSIS

XGBoost Logloss & Overfitting



- **X-axis:** boosting rounds; **Y-axis:** logloss
- Both train and test logloss decrease rapidly at early rounds
- **Around ~60–80 rounds:**
  – Test logloss already very low and curves almost overlap
- **Curves stay close and flatten near zero:**
  → No strong overfitting at chosen complexity
- **Illustrates concept:**
  – If we kept adding trees, train loss ↓ but test loss ↑
  → overfitting
  – In practice: use early stopping / stronger regularization

QUANTITATIVE RESULTS & OVERFITTING ANALYSIS

SHAP Summary: Main Drivers of Attrition

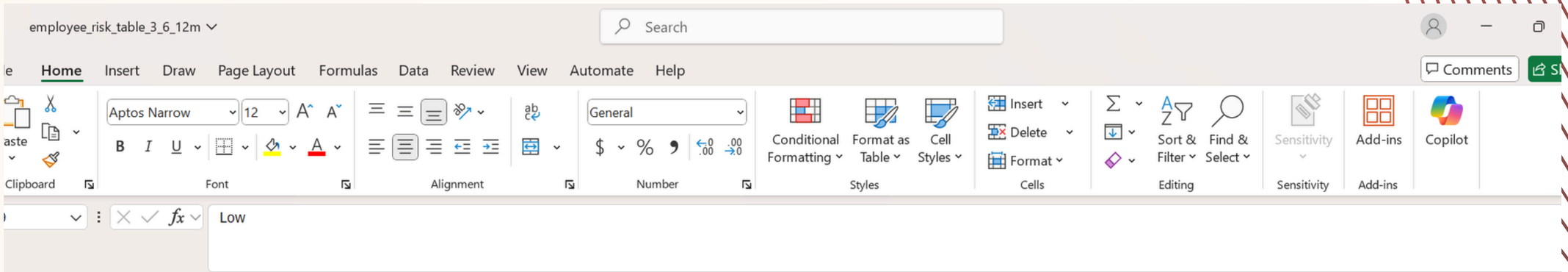
SHAP value = contribution of each feature to predicted attrition risk

- **Top HR-related drivers (ignoring technical event variable):**
  - **OverTime_Yes** – high overtime pushes risk up.
  - **StockOptionLevel** – lower options associated with higher risk.
  - **Age & MonthlyIncome** – younger and lower income → higher risk.
  - **MaritalStatus_Single** – single employees more likely to leave.
  - **Tenure variables** (YearsAtCompany, YearsInCurrentRole, TotalWorkingYears) – longer tenure → lower risk.
  - **Satisfaction & involvement** – higher satisfaction/involvement → lower predicted attrition.
  - **BusinessTravel_Travel_Frequently**, **DistanceFromHome**, **some JobRoles** also contribute.

SHAP explanations are consistent with Cox results and with HR intuition.

RISK STRATIFICATION & HR USE CASES

RISK TABLE OUTPUT (3 / 6 / 12 MONTHS)

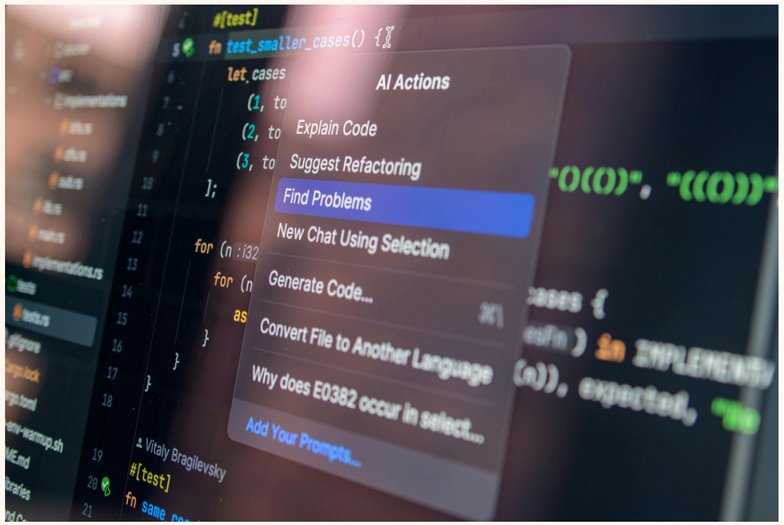


A	B	C	D	E	F	G
EmployeeID	Prob_leave_3m	Prob_leave_6m	Prob_leave_12m	RiskTier_6m	Department	JobRole
1	0.18%	0.18%	1.47%	Low	Sales	Sales Executive
2	0.00%	0.00%	0.00%	Low	Research & Development	Research Scientist
4	46.30%	46.30%	64.39%	Medium	Research & Development	Laboratory Technician
5	0.21%	0.21%	1.91%	Low	Research & Development	Research Scientist
7	0.09%	0.09%	1.18%	Low	Research & Development	Laboratory Technician
8	0.00%	0.00%	0.04%	Low	Research & Development	Laboratory Technician
10	0.24%	0.24%	6.59%	Low	Research & Development	Laboratory Technician
11	0.59%	0.59%	12.69%	Low	Research & Development	Laboratory Technician
12	0.00%	0.00%	0.00%	Low	Research & Development	Manufacturing Director
13	0.00%	0.00%	0.00%	Low	Research & Development	Healthcare Representative
14	0.06%	0.06%	0.20%	Low	Research & Development	Laboratory Technician
15	0.03%	0.03%	0.19%	Low	Research & Development	Laboratory Technician
16	0.00%	0.00%	0.19%	Low	Research & Development	Research Scientist
18	0.00%	0.00%	0.59%	Low	Research & Development	Laboratory Technician
19	1.13%	1.13%	4.29%	Low	Research & Development	Laboratory Technician
20	0.00%	0.00%	0.00%	Low	Research & Development	Manufacturing Director

Suggested actions:

- 1-on-1 check-ins & short pulse surveys; adjust OT/workload; rotation/promotion plans; targeted compensation offers.
- Operating cycle: Run quarterly; track movement across Risk Tiers after interventions.

LIMITATIONS & FUTURE WORK



LIMITATIONS:

- Data: IBM HR dataset is synthetic/demo → may differ from live environments.
- Evaluation: Some AUC(t)/Brier at 12/24/36 months can be unstable due to censoring; use C-index as primary.
- Modeling: Limited tuning for RSF/XGBoost; missing engagement/behavioral features.

FUTURE WORK:

- Systematic Grid/Random Search for RSF & XGBoost; try DeepSurv / GBM survival.
- Enrich features with performance reviews, eNPS, mobility/training history.
- Ship an HR dashboard (filters, export, alerting).

CONCLUSION

End-to-end pipeline: EDA → Kaplan–Meier → Cox/RSF → XGBoost + SHAP → 3/6/12-month risk table.
Model results:

- Cox: High, stable C-index on train/test (good generalization).
- RSF: C-index > 0.5 with small train–test gap.
- XGBoost: Clearly demonstrates overfitting and the need for early stopping/regularization.

Impact for HR: Risk tiers + explanations (SHAP) enable proactive retention instead of reactive responses.

Built a complete pipeline with reliable Cox/RSF performance, highlighted XGBoost overfitting and fixes, and deliver a final 3/6/12-month risk table with explanation
=> Enough for HR to prioritize actions and track outcomes over time.

Acknowledge course learning outcomes: hands-on experience with survival analysis, model evaluation, overfitting, and explainable AI in an HR context.

REFERENCES

- 01] Mishra S N, Lama D R and Pal Y 2016 Human Resource Predictive Analytics (HRPA) for HR Management in Organizations International Journal Of Scientific & Technology Research 5(5) 33-35
- 02] Hoffman M and Tadelis S 2018 People Management Skills, Employee Attrition, and Manager Rewards: An Empirical Analysis National Bureau of Economic Research
- 03] Frye A, Boomhower C, Smith M, Vitovsky L and Fabricant S 2018 Employee Attrition: What Makes an Employee Quit? MU Data Science Review 1(1)
- 04] S. Rabiyathul Basariya, Ramyar Rzgar Ahmed, A STUDY ON ATTRITION - TURNOVER INTENTIONS OF EMPLOYEES, International Journal of Civil Engineering and Technology (IJCIET), 2019, 10(1), PP2594-2601
- 05] Halkos, George & Bousinakis, Dimitrios, 2017. "The effect of stress and dissatisfaction on employees during crisis," Economic Analysis and Policy, Elsevier, vol. 55(C), pages 25-34.
- 06] Glavas, A., & Willness, C. (2020). Employee (dis)engagement in corporate social responsibility. In D. Haski-Leventhal, L. Roza, & S. Brammer (Eds.), Employee engagement in corporate social responsibility (pp. 10–27). Sage Publications Ltd. https://doi.org/10.4135/9781529739176.n2
- 07] S. Yadav, A. Jain and D. Singh, "Early Prediction of Employee Attrition using Data Mining Techniques," 2018 IEEE 8th International Advance Computing Conference (IACC), Greater Noida, India, 2018, pp. 349-354, doi: 10.1109/IADCC.2018.8692137.
- 08] R. Jain and A. Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach," 2018 International Conference on System Modeling & Advancement in Research Trends (SMART), Moradabad, India, 2018, pp. 113-120, doi: 10.1109/SYSMART.2018.8746940.
- 09] Alduayj, Sarah & Rajpoot, Kashif. (2018). Predicting Employee Attrition using Machine Learning. 93-98. 10.1109/INNOVATIONS.2018.8605976.
- 10] Setiawan, Irwan & Suprihanto, Suprihanto & Nugraha, Ade & Hutahaean, Jonner. (2020). HR analytics: Employee attrition analysis using logistic regression. IOP Conference Series: Materials Science and Engineering. 830. 032001. 10.1088/1757-899X/830/3/032001.
- 11] Yadav, Sandeep & Jain, Aman & Singh, Deepti. (2018). Early Prediction of Employee Attrition using Data Mining Techniques. 349-354. 10.1109/IADCC.2018.8692137.
- 12] I. Ballal, S. Kavathekar, S. Janwe, P. Shete, and N. Bhirud, “People Leaving the Job-An Approach for Prediction Using Machine Learning,” Int. J. Res. Anal. Rev., vol. 7, no. 1, pp. 8–10, 2020, [Online]. Available: www.ijrar.org
- 13] A. Qutub, A. Al-Mehmadi, M. Al-Hssan, R. Aljohani, and H.S. Alghamdi, “Prediction of Employee Attrition Using Machine Learning and Ensemble Methods,” Int. J. Mach. Learn. Comput., vol. 11, no. 2, pp. 110–114, 2021, doi:10.18178/ijmlc.2021.11.2.1022.
- 14] N. Mansor, N. S. Sani, and M. Aliff, “Machine Learning for Predicting Employee Attrition,” Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 11, pp. 435–445, 2021, doi: 10.14569/IJACSA.2021.0121149.
- 15] D. Saisanthiya, V. M. Gayathri, and P. Supraja, “Employee Attrition Prediction Using Machine Learning and Sentiment Analysis,” Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 5, pp. 7550–7557, 2020, doi: 10.30534/ijatcse/2020/91952020.

THANK YOU