Strategic Prediction of Employee Attrition Risk
Course Project: Classification Analysis

Team Members: Laman Guluzada & Javid Guliyev

February 24, 2026

1 Motivation

In modern HR management, identifying ”at-risk” employees before resignation is more cost-effective as losing employees leads to recruitment costs, training expenses, and/or productivity loss. For this reason, identifying employees who may be at risk of leaving has become increasingly important and benefit companies by keeping costs lower.
This project focuses on **multi-class attrition risk prediction**, where employees are categorized into Low (0), Medium (1), and High (2) risk levels using a dataset of 15,000 records. By analyzing demographic information, salary levels, job satisfaction, and workload-related variables, we aim to understand the key factors that influence attrition risk. The goal is to build and compare predictive models to classify employee into threes risk groups. We also explored binary classification set up using only high risk employees and others as targets.

2 Dataset and Preprocessing

The dataset consists of 15 variables with no missing values.

- **Class Distribution**: The target variable is moderately imbalanced: 50% Low, 35% Medium, and 15% High Risk.

- **Feature Engineering**: `Employee_ID` was removed. `Gender` was binary encoded, and `Job_Role` was one-hot encoded. Continuous variables were standardized using `StandardScaler` to prevent data leakage.

- **Data Integrity**: Outliers presented in less than 3% of observations and were retained, as their proportion was small and not likely to distort model estimates. Pairwise correlation analysis indicated moderate linear associations among tenure and age related variables, most visible between and `Years_at_Company` ($r \approx 0.63$) and between `Years_at_Company` and `Years_Since_Last_Promotion` ($r \approx 0.73$).

  Variance Inflation Factors (VIF) were computed for all predictors to observe multicollinearity formally. VIF values ranged from 1.00 to 6.72, with higher values observed for `Job_Satisfaction`, `Work_Life_Balance`, and `Num_Projects`. However, all VIF scores remained below the critical threshold of 10 which indicates no evidence of severe multicollinearity.

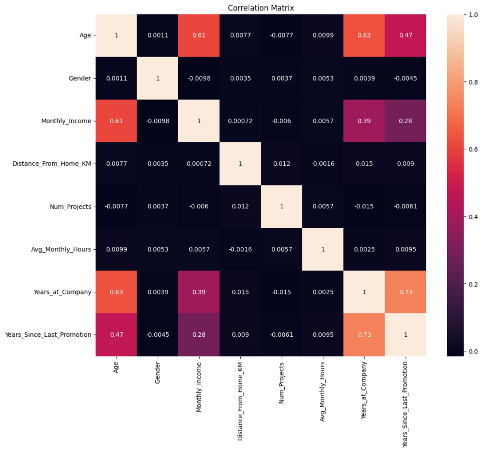


Figure 1: Correlation matrix of HR Dataset

3 Binary Logistic Regression

A binary logistic regression model was estimated to differentiate *high attrition risk* employees from all other risk categories. The model achieved a **Pseudo R^2 of 0.3958**, indicating strong explanatory power. The likelihood ratio test confirms that the model is statistically significant overall ($p < 0.001$).

3.1 Coefficient-Level Statistical Inference

Predictors with $p < 0.05$ are considered statistically significant.

3.1.1 High Risk vs. Non-High Risk

- **Years_at_Company**: $\beta = 1.8416, z = 35.444, p = 0.000$. A very strong positive predictor, indicating that employees with longer tenure have substantially higher odds of being classified as high attrition risk.

- **Job_Satisfaction**: $\beta = -0.8852, z = -24.475, p = 0.000$. A strong protective factor, where higher job satisfaction significantly reduces the likelihood of high attrition risk.

- **Work_Life_Balance**: $\beta = -0.6888, z = -17.933, p = 0.000$. Better work–life balance is associated with a lower probability of high attrition risk.

- **Avg_Monthly_Hours**: $\beta = 0.4790, z = 14.170, p = 0.000$. Increased workload significantly elevates the probability of high-risk attrition.

- **Distance_From_Home_KM**: $\beta = 0.1804, z = 5.727, p = 0.000$. Longer commuting distances are associated with higher attrition risk.

- **Monthly_Income**: $\beta = -0.1642, z = -3.927, p = 0.000$. Higher income significantly reduces the likelihood of high attrition risk.

- **Age**: $\beta = -0.1470, z = -2.549, p = 0.011$. Older employees are less likely to be classified as high risk.

- **Insignificant**: Gender ($p = 0.394$) and Num_Projects ($p = 0.462$) do not show a statistically significant association with high attrition risk.

Prediction Accuracy: The binary logistic regression model achieved a prediction accuracy of 0.884, indicating that it correctly classifies approximately 88.4% of observations in the test data.

4 Multinomial Logistic Regression

The model achieved a **Pseudo R^2 of 0.3695**, indicating significant explanatory power.

4.1 Coefficient-Level Statistical Inference

Predictors with $p < 0.05$ are considered statistically significant.

4.1.1 Medium Risk (Class 1) vs. Low Risk

- **Job_Satisfaction**: $\beta = -1.4356, z = -41.399, p = 0.000$. A strong protective effect.
- **Monthly_Income**: $\beta = -0.4024, SE = 0.033, z = -12.195, p = 0.000$.
- **Years_at_Company**: $\beta = 1.5714, p = 0.000$. Positive association with risk.
- **Work_Life_Balance**: $\beta = -1.1426, z = -34.113, p = 0.000$. Strong protective effect against medium attrition risk.
- **Avg_Monthly_Hours**: $\beta = 0.7110, z = 25.271, p = 0.000$. Higher workload substantially increases attrition risk.
- **Distance_From_Home_KM**: $\beta = 0.3259, z = 12.265, p = 0.000$. Longer commute distances are associated with higher risk.
- **Insignificant**: Age ($p = 0.221$), Gender ($p = 0.144$), and Num_Projects ($p = 0.773$).

4.1.2 High Risk (Class 2) vs. Low Risk

- **Years_at_Company**: $\beta = 3.0550$. Extremely strong predictor of high risk ($OR \approx e^{3.055}$).
- **Job_Satisfaction**: $\beta = -2.0149$. The strongest protective factor.
- **Age**: $\beta = -0.1738, p = 0.007$. Older employees are less likely to be in the high-risk group.
- **Work_Life_Balance**: $\beta = -1.5862, z = -32.645, p = 0.000$. Strongly reduces the likelihood of high-risk attrition.
- **Avg_Monthly_Hours**: $\beta = 1.0319, z = 25.013, p = 0.000$. One of the strongest positive predictors of high attrition risk.

- **Distance_From_Home_KM**: $\beta = 0.4418, z = 11.589, p = 0.000$. Increased commute distance significantly elevates high-risk attrition probability.
- **Monthly_Income**: $\beta = -0.4799, z = -9.748, p = 0.000$. Higher income substantially lowers high-risk attrition likelihood.

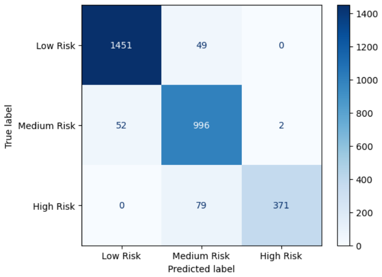
Prediction Accuracy: The model achieved a prediction accuracy of 0.716, meaning that it correctly classifies approximately 71.6% of observations in test data.

4.2 Confounding Variable Analysis

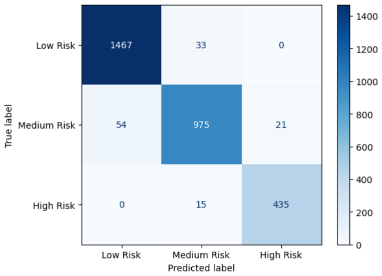
A confounding relationship was identified between **Age** and **Years_at_Company** ($r \approx 0.63$). When tenure was removed, the Age coefficient increased considerably, and Pseudo R^2 dropped from 0.3695 to 0.2176, indicating that Age's effect is partially explained by tenure.

5 Discriminative Analysis (LDA & QDA)

- **LDA**: Achieved **93.93% accuracy**. For the High-Risk class, the AUC was **0.997**. An optimized threshold of **0.25** yielded Precision = 0.964 and Recall = 0.904.
- **QDA**: Achieved the highest performance with **95.9% accuracy**. This suggests class-specific covariance matrices better capture the non-linear structure of the data.



Confusion matrix for LDA



Confusion matrix for QDA

6 Naive Bayes and Model Comparison

Naive Bayes achieved **78.8% accuracy**. Its performance was limited by the independence assumption, which was violated by correlations between Age, Income, and Tenure.

6.1 Model Comparison and Interpretation

Quadratic Discriminant Analysis (QDA) achieved the highest accuracy (95.9%), meaning that the underlying class structure is better captured by non-linear decision boundaries. Linear Discriminant Analysis (LDA) also performed strongly (93.9%), indicating that much of the class separation can be approximated linearly.
The Binary Logistic Regression model demonstrated high predictive performance (88.4%), particularly effective in separating high-risk employees from others. This indicates it is suitability for risk detection task.

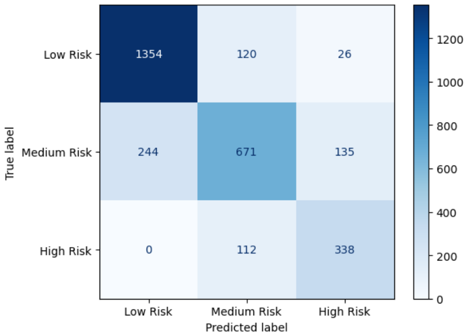


Figure 2: Confusion matrix for Naive Bayes

Naive Bayes showed moderate performance (78.8%), likely because of its strong independence assumption, which does not hold given the correlation among the data features such as income, workload, and tenure.
While Multinomial Logistic Regression resulted in the lowest accuracy (71.6%), it still remains highly valuable for statistical inference and interpretability. Unlike QDA and LDA, logistic regression provides us with direct coefficient estimates, odds ratios, and statistical significance measures. It makes LR preferable for explanatory analysis rather than purely predictive task of risk classification.
Overall, QDA is the strongest predictive model, whereas Logistic Regression (binary and multinomial) offers better interpretability and inferential opportunity.

Table 1: Model Performance Comparison

Model	**Accuracy**	**Key Insight**
QDA	**95.9%**	Best performer; non-linear class boundaries.
LDA	93.9%	High discriminative ability.
Binary Logistic Reg.	88.4%	Higher predictive accuracy for high-risk detection than multiclass.
Naïve Bayes	78.8%	Moderate; is sensitive to feature correlation.
Logistic Reg.	71.6%	Best for inference and/or coefficient analysis.

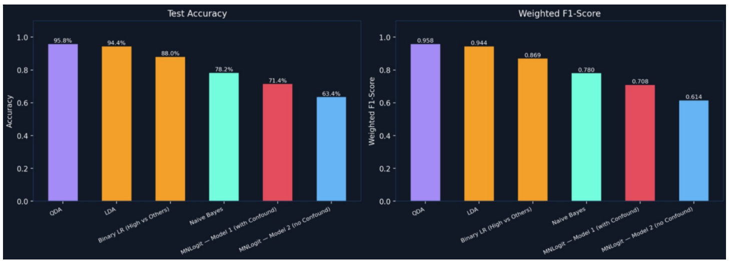


Figure 3: Logistic Regression, LDA, QDA and Naive Bayes models' performace comparison

7 Linear vs. Poisson Regression (Num_Projects)

Both OLS and Poisson models were used to model the count variable `Num_Projects`.

- **Results**: None of the predictors were significant ($p > 0.05$).
- **Fit**: OLS (AIC = 63228.63) vs. Poisson (AIC = 63455.90).
- **Conclusion**: $R^2 \approx 0$ for both, indicating negligible explanatory power for project count.

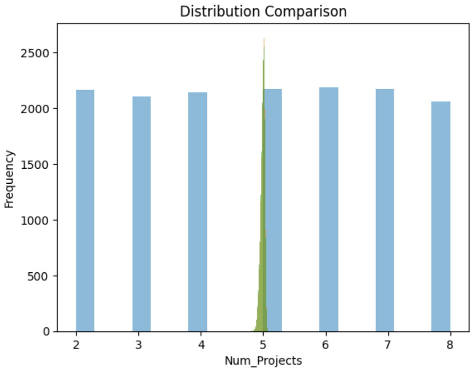


Figure 4: Distribution of Number of Projects

8 Conclusion and Team Contributions

QDA demonstrated the best predictive performance, followed by LDA, Naïve Bayes, and Logistic Regression. This means that the dataset exhibits class-specific variance patterns and non-linear separability which makes discriminant analysis methods more suitable than linear models.
Laman Guluzada worked on Preprocessing, Logistic Regression, LDA and QDA.
Cavid Guliyev focused on Poisson Analysis and OLS Analysis, Naïve Bayes, and the UI Design.

References

1. B. Marious Kono, “HR Analytics: Employee Attrition and Risk Levels,” Kaggle Dataset, 2025. Available: https://www.kaggle.com/datasets/bertnardomariouskono/hr-analytics-employee-attrition-and-risk-levels