ALY 6120: Leadership in Analytics

Module_4 – Assignment_4 (Group 3)

CRISP-DM Cycle: Data Preparation

Student Names: Parita Gala, Josephine Agbedoawu, Jiansheng Shentu, Godfred Akoto

Professor: Abeba N. Turi

Date: 26 November 2025

Abstract

This project explores employee attrition using the IBM HR Analytics Employee Attrition & Performance dataset, with the goal of identifying meaningful patterns that can strengthen workforce retention strategies. The dataset includes demographic, organizational, behavioral, and compensation-related variables representative of real HRIS systems. After detailing data acquisition through Kaggle, this report outlines internal versus external sourcing, data preparation processes, team collaboration needs, and data quality assessments. The modelling approach focuses on uncovering the key drivers of attrition and predicting employees at risk of leaving. Logistic regression is selected as an interpretable baseline model, supported by decision trees and Random Forest for deeper pattern detection and improved predictive performance. Together, these models provide actionable insights into employee turnover, enabling HR leadership to design targeted interventions grounded in data-driven evidence.

Identify Your Data Sources

For this project, the primary dataset is the IBM HR Analytics Employee Attrition & Performance dataset. This dataset includes: Employee Demographics: Age, gender, marital status, education, education field.

Job & Organizational Variables: Job role, department, job level, job satisfaction, environment satisfaction, work-life balance. Compensation & Benefits: Monthly income, hourly rate, stock option level, percent salary hike. Performance & Engagement: Performance rating, relationship satisfaction, training times last year. Workplace Behavior: Distance from home, overtime status, total working years, job involvement. Attrition Labels: Whether the employee left the company (Yes/No).

Determine Whether They Can Be Acquired Internally or Externally

Internal Acquisition: In a real corporate setting, HR departments would acquire this type of data internally from HRIS, payroll, performance management systems, and employee engagement systems. All the variables present in the IBM dataset map directly to internal company systems.

External Acquisition: For academic or training purposes, the IBM HR Analytics Attrition dataset is publicly available externally (e.g., via Kaggle). Therefore, while the dataset represents internal HR data, it is accessed externally for this project.

How The Data Will Be Acquired

We will obtain the IBM HR Analytics Employee Attrition & Performance dataset from Kaggle, a reliable source for publicly accessible datasets utilized in both academic and professional analytics.

Acquisition and Validity : We will go to Kaggle's official IBM HR Analytics repository, obtain the CSV file, and confirm that it includes the necessary variables (Age, Attrition, Job Satisfaction, Monthly Income, Years at Company, Work-Life Balance, etc.) without any corruption or missing information.

Protected Storage : The dataset would be stored in a well-organized project directory with distinct naming rules (e.g., IBM_HR_Analytics_Raw_2025-11-22.csv), keeping the original file distinct from its processed copies to guarantee reproducibility. Load into Tools: We would bring the data into Python using pandas (pd.read_csv()) for analysis and cleaning, or into Power BI through the "Get Data" connector for visualization. Depending on the modelling needs, we may also use R or Jupyter notebooks.

Document Source: We would maintain a README file documenting the source (Kaggle), download date, original creators (IBM Watson Analytics), and any preprocessing steps applied. Since this is a synthetic dataset created for educational purposes, it contains no real employee information, eliminating privacy concerns while providing a realistic HR analytics practice.

Team Members We Will Collaborate With

An HR Business partner or HR Analyst will provide domain expertise on workforce metrics and organizational practices. They would validate my interpretation of variables like Job Satisfaction and Work Life Balance, explain what constitutes concerning attrition rates, and help translate analytical findings into actionable retention strategies HR leadership can implement.

A Data Scientist or Analytics Manager oversees methodology and technical approach. They would review our feature selection for the attrition prediction model, recommend appropriate algorithms (logistic regression, decision trees, random forest), validate model performance interpretation, and ensure we follow best practices while avoiding pitfalls like overfitting.

IT/Database Professional (for real organizational data) manages data access, quality, and security. While unnecessary for this public Kaggle dataset, they become critical when working with actual employee information systems, providing data extraction, explaining table relationships, and ensuring proper security protocols.

HR Leadership (Stakeholders) represents the end users of our analysis. We shall work together with them to grasp their particular retention inquiries, specify what choices they must make, and deliver results in understandable formats that encourage action instead of merely generating technical documents.

Data Governance or Privacy Officer (in practical applications) guarantees adherence to regulations such as PIPEDA, FIPPA, or GDPR. Although this synthetic dataset doesn't need supervision, real HR data necessitates their authorization and direction regarding de-identification and privacy measures.

Within an academic environment, we would partner with our instructor for methodological advice and engage with classmates for peer evaluations and different analytical strategies. This cooperative method guarantees that the analysis is technically robust, relevant to the domain, and applicable for decision-making within the organization.

Data Preparation and Quality Assessment

In this assignment, before the data would be passed on for further modelling, we would employ several preparation steps to ensure that there is data quality, reliability, and suitability for the predictive modelling, considering that the focus for the analysis here is attrition.

First, missing data would be identified and quantified across all variables. Although there might be a few Nas in the dataset, any gaps in fields such as education, job role, or income would require attention. Missing values in numerical variables may be imputed using simple approaches like mean or median, or flagged for review, while missing categorical values may need mode imputation or a dedicated unknown category. Rows with excessive missingness might be removed after several deliberations from the group.

Next, the data would be checked for incorrect values, for example negative ages, unrealistic total working years, duplicated employee IDs, or inconsistent combinations such as “Years at Company” greater than “Total Working Years.” Categorical variables would also require validation to ensure labels are properly formatted and standardized. Outlier detection is essential for numerical fields such as monthly income, years of experience, distance from home, or daily rate. The outliers that would be identified may indicate data entry errors or genuinely extreme cases which would be evaluated and either corrected or removed depending on context.

The final step would be to ensure correct data types and documenting all cleaning decisions be to ensure transparency for the modelling team.

Models to Employ and Rationale

To discover meaningful patterns in employee attrition, I would employ a combination of descriptive, predictive, and explanatory models. Each type of model contributes a different layer of insight that supports both understanding and decision-making.

1. Logistic Regression (Primary Predictive Model)

Logistic regression is the most appropriate baseline model for attrition because the target variable—whether an employee left the company—is binary. This model offers:

- **Interpretability**, allowing HR partners and business leaders to understand the direction and strength of each predictor (e.g., how overtime or job satisfaction affects attrition odds).

- **Transparency**, which is critical for HR decision-making and ethical considerations.
- **Baseline benchmarking**, enabling comparisons with more complex models.

Its coefficients help translate statistical patterns into actionable insights, such as identifying high-risk groups or key drivers of turnover.

2. Decision Trees

Decision trees provide an intuitive, rule-based structure that is easy for non-technical stakeholders to understand. They reveal:

- **Clear segmentation patterns**, such as how combinations of variables (e.g., low job satisfaction + high commute distance) increase attrition likelihood.
- **Non-linear relationships** that logistic regression might not capture.

Decision trees are particularly useful for creating **profile-based retention strategies**, as they visually show which employee groups are at highest risk.

3. Random Forest (Ensemble Model for Enhanced Accuracy)

Random Forest is an ensemble method that aggregates many decision trees, improving predictive performance and reducing overfitting. This model is ideal for:

- **Handling high-dimensional HR data** with multiple interacting variables.
- **Identifying important predictors** through built-in feature importance metrics.
- **Capturing complex, non-linear patterns** in attrition behavior.

Random Forests are typically one of the strongest models for this dataset, offering both robustness and high accuracy.

4. Gradient Boosting Models (e.g., XGBoost) – Optional for Performance Optimization

If the project requires maximizing predictive accuracy, gradient boosting techniques like XGBoost can be used. They often outperform simpler models by:

- Sequentially correcting errors of previous trees
- Handling imbalanced datasets more effectively
- Capturing subtle interactions among predictors

However, these models are less interpretable, so they would be used only after ensuring fairness and transparency.

5. Exploratory Models: Cluster Analysis

Before predictive modelling, cluster analysis (e.g., K-Means) can uncover:

- **Natural patterns in employee groups**
- **Hidden segments** such as early-career employees with long commutes or high-income employees with low engagement
- **Profiles where attrition risk is highest**

These insights strengthen feature engineering and help HR design targeted interventions.

Summary

A **tiered modelling approach**—starting with logistic regression for interpretability, followed by tree-based models for deeper pattern discovery, and optionally advanced ensemble models for accuracy—ensures both analytical rigor and practical relevance. Together, these models allow us to uncover the drivers of attrition, predict high-risk employees, and design evidence-based retention strategies aligned with organizational priorities.

Reference

IBM. (n.d.). *IBM HR Analytics Employee Attrition & Performance*. IBM Watson Analytics. https://www.ibm.com/

Kaggle. (n.d.). *IBM HR Analytics Employee Attrition & Performance*. Retrieved from https://www.kaggle.com

Delen, D., & Ram, S. (2018). *Predictive analytics in HR: A framework and research agenda*. Journal of Business Analytics, 1(1), 1–22.

Shmueli, G., Bruce, P., Gedeck, P., & Patel, N. (2020). *Data mining for business analytics: Concepts, techniques, and applications in R*. Wiley.

Kuhn, M., & Johnson, K. (2019). *Applied predictive modeling*. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.