UNIVERSITI MALAYSIA SARAWAK
Faculty of Computer Science and Information Technology

Assignment/Report Cover Sheet

Group: JojaCorp

Student Name	Student Id Number	Lecture Group	Signature
NORATHIRAH IZZATI BINTI MUSA	82783	01	
AINUR SOFIYA BINTI JAMHARI	83119	01	
CALISTA SUPANG ANAK DANDY VICTOR	83411	01	
NURUL FAZLINA BINTI MEDIHI	85360	01	
VIONNA KHO KAI XIN	86001	01	

Subject Code: TMI4033		**Subject Name:** Collective Intelligence	
Assignment Title:	Assignment 2	**Lecturer:**	Dr Suhaila Binti Saee
Due Date: 9 December 2025, 12.05P.M.		**Date Submitted:** 8 December 2025	

This cover sheet must be completed, signed and firmly attached to the front of the submission. All work must be submitted by the due date. If an extension of work is granted, an assignment extension acknowledgement slip must be signed by the lecturer/tutor and attached to assignment. Please note that is your responsibility to retain copies of your assignment.

Plagiarism and Collusion are methods of cheating that falls under Peraturan Akademik Universiti Malaysia Sarawak para 11: Etika Akademik

Plagiarism
Plagiarism is the presentation of work which has been copied in whole or in part from another person's work, or from any other source such as the internet, published books or periodicals without due acknowledgement given in the text.

Collusion
Collusion is the presentation of work that is the result in whole or in part of unauthorized collaboration with another person or persons. Where there are reasonable grounds for believing that cheating has occurred, the only action that may be taken when plagiarism or collusion is detected is for the staff member not to mark the item of work and to report or refer the matter to the Dean. This may result in work being disallowed and given a fail grade or if the circumstances warrant, the matter may be referred to a committee of inquiry for investigation. Such investigation may result in the matter being referred to the University Discipline Committee, which has the power to exclude a student.

Upon placing signature above, I certify that I have not plagiarized the work of others or participated in unauthorized collusion when preparing this assignment.

I also certify that I have taken proper care in safeguarding my work and have made all reasonable efforts to ensure that my work is not able to be copied.

MARK:

Table of Contents

1. Brief Explanation on the Proposed Application

1.1 Clarity and Conciseness

As discussed in assignment 1, the proposed application is the Predictive Employee Retention System (PERS). The purpose of PERS is to serve as an early warning system that helps address the costly and disruptive problem of unexpected employee attrition. Its main purpose is to help Human Resources (HR) keep valuable employees by predicting what factors might make them leave and identifying those who are at high risk of resigning before they resign, maintaining or improving employee retention.

The system works by combining Machine Learning (ML) with a Collective Intelligence (CI) approach. It uses supervised learning trained on the organization's combined workforce data such as Human Resources Information System (HRIS) records, performance scores, and employee engagement results. It starts with a data pre-processing engine that cleans and organizes this raw data. The clean data is then fed into the ML prediction model, which analyses it and produces an attrition risk score for each employee. To ensure high accuracy and stable performance, the system relies on powerful ensemble models like Random Forest (RF) and Extra Trees Classifier (ETC). The PERS application serves as the user interface, showing the results and highlighting employees who are at high risk of leaving. This helps the HR team understand the reasons behind the risks and make better decisions to retain staff.

1.2 Relevance and Scope

The main goal of PERS is to turn raw organisational data into useful predictions that assist in smarter decision-making for Human Resource Management (HRM). The system continuously collects and analyses internal workforce data such as HRIS information, employee engagement levels, and performance records. It aims to shift HR departments away from reactive and subjective decision-making to more proactive strategies guided by patterns observed across the entire workforce.

To achieve this, the system uses ensemble machine learning methods that will capture the organisation's collective behaviour data and convert it into reliable predictions. The Extra Trees Classifier (ETC) is selected specifically because research shows it can achieve high accuracy making it highly dependable in identifying whether an employee is "High-Risk", "Medium-Risk", and "Low-Risk."

PERS is designed to directly tackle employee attrition, a major issue that leads to long-term financial loss when experienced or essential employees leave. By compiling historical data from the entire workforce to train its predictive models, the system forms a shared organizational knowledge base that is far more reliable and actionable than relying on individual managerial intuition. This approach reflects the growing shift toward data-driven

HR practices. PERS essentially digitises the organisation's collective intelligence on employee satisfaction and behaviour patterns, giving HR teams the tools they need to identify potential risks early and make informed decisions based on evidence rather than assumptions.

1.3 Context and Justification

The application is highly relevant because employees are one of an organization's most valuable assets, crucial for its reputation and long-term success. Failing to maintain a stable workforce can result in significant long-term losses. A study reveals that Malaysia's voluntary turnover rate has reached 9.5%, one of the highest in Southeast Asia, highlighting the country's growing challenge in retaining talent (Sim, 2024). Traditional methods of managing attrition, which rely on subjective interpretations, are often inefficient and fail to prevent employee departures. PERS addresses this gap by enabling HR decisions to be based on objective, predictive data analysis.

One key user benefit of this application is its ability to enable targeted early interventions. PERS gives HR to take timely actions such as offering flexible schedules, adjusting workloads, or providing career development opportunities before an employee decides to resign. The application also enhances transparency and communication. By identifying the specific factors that place an employee at high risk of attrition, PERS promotes fairness, clearer communication, and a better understanding between HR and managerial staff when designing retention strategies.

Additionally, the system provides objective decision support. HR teams can base critical retention decisions on data-driven insights derived from employees' historical records, consistently outperforming decisions made solely through manual judgment or intuition.

Finally, PERS continuously improves over time. It is designed with a learning loop that updates the predictive models as HR inputs feedback on interventions or actual reasons for resignations. This allows the system to adapt to changing organizational conditions, ensuring that its predictions remain accurate and relevant.

2. Pipeline of the Proposed Application

2.1 Comprehensive Stages

2.1.1 Data Collection

Employee data is collected from HRIS, including attendance records, performance evaluations, engagement metrics, and demographic records. This may include details such as the number of previous companies worked at, monthly income, job satisfaction, job role, department, job level, and total working hours. All these inputs are integrated and merged into a single, structured dataset for further processing.

2.1.2 Data Preprocessing

Data preprocessing must be carried out to prepare the dataset for modelling and includes the following tasks:

a. **Data Cleaning**

   Values that are missing, duplicated, or outliers are removed and repaired to ensure data reliability.

b. **Normalisation and Standardisation**

   Numerical features are scaled to consistent ranges so the model can process them effectively.

c. **Encoding Categorical Variables**

   Fields such as gender, job role, or department are converted into numeric form using techniques such as one-hot encoding or label encoding.

d. **Feature Selection**

   Using correlation analysis and variance thresholds, low-impact or irrelevant predictors are removed to ensure the accuracy of prediction.

e. **Train-Test Split:**

   Dataset is split for example, 70% for training and 30% for testing to ensure proper model evaluation and validation.

2.1.3 Model Training

The cleaned and transformed dataset is used to train the predictive model. Ensemble ML methods such as RF and ETC are applied during this stage, where the models learn patterns associated with employee attrition.

2.1.4 Model Evaluation

The best-performing ML model is selected for deployment based on their performance assessed using the previously separated test set. The key evaluation includes accuracy, precision and recall, F1-score, and Receiver Operating Characteristic – Area Under the Curve (ROC-AUC).

2.1.5 Prediction and Risk Scoring

The final model generates an attrition risk score for every employee based on the processed input data. The system classifies the employees into categories such as “High-risk”, “Medium-risk” and “Low-risk” to indicate their likelihood of resigning. These results are produced in real time whenever new or updated employee data is fed into the system, giving HR continuous, up-to-date insights into workforce retention levels.

2.1.6 CI-Based Knowledge Integration

This stage integrates the ML predictions with inputs from the organisation's CI. It combines HR feedback, such as reasons of resignation and the outcomes of retention interventions, as well as insights from managers. Organisational knowledge, such as policies, workload trends, and promotion cycles is also considered. Combing these sources of information helps to refine model's predictions.

2.1.7 Decision Support and Intervention Planning

The PERS interface presents insights to HR such as list of high-risk employees, contributing factors to attrition, and the recommended retention strategies.

2.1.8 Continuous Learning Feedback Loop

The PERS collects feedback as HR implements the early intervention to prevent attrition and constantly updates the system. These updates are used to retrain the model and to refresh feature importance, ensuring PERS stays accurate and relevant over time.

2.2 Logical Flow

The pipeline follows a structured and sequential flow to ensure accurate employee attrition prediction. First, the system collects raw employee information from HRIS sources and consolidates it into a unified dataset. This dataset then flows into the preprocessing stage, where it is systematically transformed into a machine-readable form. Once prepared, the processed data moves into model training, where the machine learning algorithms learn the underlying attrition patterns.

The trained models are then evaluated to determine which performs best, and the selected model is used for generating attrition risk predictions. These predictions are enriched through organisational knowledge integration, where HR insights, managerial feedback, and contextual company information add greater depth to the results. The refined insights then progress into the decision-support stage, allowing HR to plan targeted retention strategies.

Finally, the outcomes of HR interventions loop back into the system through a continuous learning mechanism, enabling the model to be retrained and updated over time. This creates a closed, adaptive pipeline where every stage depends on the previous one, and feedback continually enhances future predictions.

2.3 Task Descriptions

Table 1 – Task Descriptions of each Pipeline Stages.

Pipeline Stage	**Task Description**
Data Collection	Employee data is gathered from HRIS sources such as attendance logs, salary records, demographic details, performance ratings, and engagement survey results. This stage ensures the machine learning model has sufficient organizational knowledge to identify attrition behavior.
Data Preprocessing	The collected data is cleaned, structured, and transformed so it is ready for Machine Learning. Techniques include missing value treatment, duplication removal, feature scaling, and categorical encoding. A Train-Test Split prepares data for fair model evaluation.
Model Training	Ensemble ML algorithms such as Random Forest (RF) and Extra Trees Classifier (ETC) are trained to study historical patterns associated with employees who stay versus those who leave, enabling accurate risk learning.
Model Evaluation	Each trained model is assessed using standard evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The top-performing model is selected for deployment to ensure dependable prediction output.
Prediction & Risk Scoring	The selected model automatically generates real-time attrition risk scores and classifies employees into High/Medium/Low attrition risk groups. These predications help HR teams prioritize intervention efforts effectively.
CI-Based Knowledge Integration	Model predictions are cross validated with collective intelligence inputs such as HR insights, resignation feedback records, and managerial assessments. This combination offers richer context and increases the reliability of prediction-based decisions.

Decision Support & Intervention Planning	The system presents interpretable findings through the PERS interface, such as key drivers influencing attrition and suggested preventive actions, helping HR teams plan and apply tailored retention measures.
Continuous Learning Feedback Loop	As HR implements interventions and records new outcomes, updated data is fed back into the model. Retraining ensures that the system continually adapts to organizational changes, improving prediction accuracy over time.

2.4 Inter-stage Dependencies

Every stage in the PERS pipeline is connected through a strict sequence in which the output of one stage becomes the direct input for the next.

Firstly, the data preprocessing stage is dependent on the output from data collection stage. The cleaning, normalisation, and encoding task cannot begin until diverse data points, such as HRIS records and engagement surveys are aggregated into the single, raw dataset produced in the first stage.

Next, the model training stage depends on the clean, well-structured data produced during the data preprocessing stage. Machine learning algorithms require properly encoded inputs and cannot work with missing or inconsistent values. High-quality preprocessing is essential, as the model's performance and accuracy are directly tied to the completeness and reliability of previous stage.

Following model training, the evaluation stage relies on the trained models and the predefined test set to measure predictive quality. Only after evaluation and identification the best-performing model can risk scoring occur.

The prediction and risk-scoring stage relies entirely on the trained model produced during the model training stage. The system employs the patterns, relationships, and mathematical rules the model has learned from historical data to estimate the likelihood of risk for current employees. Without this trained model, the system would have no basis for interpreting new inputs or generating meaningful probability scores. These predictions then feed directly into the CI-based integration stage, which enriches the model's output using HR feedback, managerial insights, and organisational knowledge.

Then, the decision-support stage depends on the outputs of prediction and CI integration. The HR dashboard needs the risk scores and enriched insights to highlight high-risk employees with the explanation of the contributing factors. This will guide HR in planning targeted interventions.

Lastly, the continuous learning stage depends on the HR feedback and the intervention outcomes recorded during the decision support stage. This new data regarding whether and employee stayed or left serves as the necessary input to trigger the retraining cycle. This allows the system to adapt to new trends.

2.5 Flexibility and Adaptability

The PERS pipeline is designed to remain flexible and adaptive system as organisational conditions, workforce behaviours, and data environments evolve. The modular data architecture of the pipeline allows new information sources to be integrated smoothly. For instance, if the organisation can introduce new data sources, such as remote work frequency, or new engagement metrics without requiring a complete system redesign. This is because the pre-processing engine can be updated to encode these features seamlessly. The selected algorithm enhances this adaptability, since its computational efficiency enables frequent and fast retraining cycles even as the dataset grows. This allows the system to scale seamlessly with the organisation's growth and operational complexity.

Furthermore, the adaptability of the system is achieved through a continuous learning feedback loop, where the outcomes of HR interventions and actual employee departures will be used to improve the application. This new data triggers periodic retraining of the model, enabling it to learn new attrition patterns. For instance, new attrition patterns that are caused by economic changes and workplace dynamic shifts. Through this mechanism, the pipeline continually updates its decision logic and preserves relevance over time without manual reconfiguration.

2.6 Visual Representation

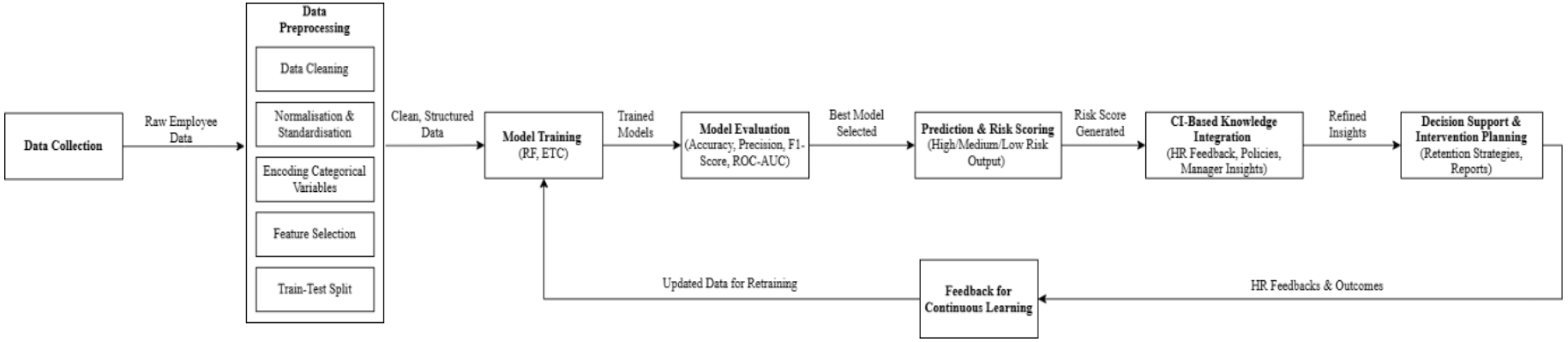

Figure 1 – Pipeline Flow of the Proposed Employee Attrition Prediction System (PERS

References

Sim, N. C. (2024). Employees turnover intention of service industry in Malaysia. *Selangor Business Review June 2024, 9*(1), 99-112.