Predicting Employee Attrition Using Weka, Google Sheets, Flourish & Napkin AI

Mini Project for PRP – End-to-End ML Workflow

1. Corporate Story (Narrative)

TechNova Solutions, a large IT company with 25,000 employees, is experiencing a sudden rise in employee attrition. Project managers are losing top-performing employees unexpectedly, HR teams are struggling to understand why people are leaving, and leadership needs accurate predictions to make strategic decisions.

Pain Points in the Corporate World

- Cost of replacing an employee is 30–50% of annual salary
- Major project delays occur when skilled employees resign
- HR must plan promotions, salary hikes, and training investments wisely

Who Faces This Problem?

- HR & Talent Management
- Project Managers
- Finance & Operations
- Leadership & Strategy Teams

What Decisions Depend on This?

- Salary revisions
- Department-level retention strategies

- Identifying employees at high risk of leaving
- Workforce planning & resource allocation

How Data + ML Tools Help

- **Google Sheets** → Clean employee data
- **Weka** → **Predict who is likely to leave**
- **Flourish** → Create visual stories for HR
- **Napkin AI** → Explain patterns & insights

This represents a **real HR Analytics project** used in companies like TCS, Infosys, Deloitte, Amazon, and Accenture.

---

2. Dataset Section

A. Dataset Specification

Dataset Name: TechNova Employee Attrition Dataset

Rows: 500

Columns: 15

Target Variable: Attrition (Yes/No)

Column List (12–18 columns):

1. EmployeeID
2. Age

3. Gender
4. Department
5. JobRole
6. MonthlyIncome
7. YearsAtCompany
8. YearsSinceLastPromotion
9. JobSatisfaction (1–4)
10. WorkLifeBalance (1–4)
11. EducationLevel
12. TrainingHoursLastYear
13. OverTime (Yes/No)
14. PerformanceRating (1–5)
15. Attrition (Yes/No) **TARGET**

Why This Dataset Works

- Represents typical HR data used in real companies
- Supports classification-based ML tasks
- Enables model comparison
- Perfect for demonstrating overfitting & underfitting
- Rich mix of numerical and categorical variables

B. Synthetic Dataset Generation Prompt

Students paste this into AI Tool:

Generate a synthetic CSV dataset with 500 rows for an HR attrition prediction problem.
Columns required:

EmployeeID (1001–1500)
Age (22–58)
Gender (Male/Female)
Department (IT, Sales, HR, Finance, R&D)
JobRole (appropriate job roles)
MonthlyIncome (15000–150000)
YearsAtCompany (0–20)
YearsSinceLastPromotion (0–10)
JobSatisfaction (1–4)
WorkLifeBalance (1–4)
EducationLevel (1–5)
TrainingHoursLastYear (0–80)
OverTime (Yes/No)
PerformanceRating (1–5)
Attrition (Yes/No)

Ensure correlations:
Low job satisfaction, high overtime → higher attrition.
Ensure no missing values.
Output as a clean CSV.

C. Similar Public Dataset Recommendation

Dataset: IBM HR Analytics Attrition Dataset
Link: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
Why Good:

- Widely used in HR analytics
- Clean attributes useful for classification
- Works perfectly with Weka

---

3. Tools Used in This Mini Project

- **Google Sheets** → Data Cleaning
- **Weka** → ML Modeling
- **Flourish** → Visualization
- **Napkin AI** → Insights Summary

Only **primary tools** are used in the steps.

---

4. Step-by-Step Execution Guide

STEP 1: Load the Dataset

Tool: Google Sheets
Click Path:
File → Import → Upload → Replace Spreadsheet

Capture (Screenshot Only):

- First 10–15 rows of dataset in Google Sheets

STEP 2: Clean & Prepare the Dataset

Tool: Google Sheets
Click Path:

- **Data → Data Cleanup → Remove duplicates**

- **Format → Number** (for numeric fields)
- **Data → Data validation** (for categories)
- **File → Download → CSV**

Capture:

- Screenshot of duplicates removal
- Cleaned dataset
- Screenshot of “Download CSV”

STEP 3: Convert CSV to ARFF

Tool: Weka → ARFF Viewer
Click Path:
Weka → Tools → ArffViewer → File → Open → Save As → .arff

Capture:

- ARFF Viewer with attributes visible

STEP 4: Load Dataset in Weka Explorer

Tool: Weka → Explorer
Click Path:
Explorer → Preprocess → Open File

Capture:

- Screenshot of attribute list & class variable

STEP 5: Explore Attributes & Visualize Patterns

Tool: Weka
Click Path:
Preprocess → Select Attribute → Visualize All

Capture:

- Screenshots of:
  - Attrition vs Job Satisfaction
  - Attrition vs OverTime
  - Histograms

STEP 6: Create Visual Narratives in Flourish

Tool: Flourish
Click Path:
New Visualization → Upload Data → Choose Chart

Visuals to Create:

- Attrition by Department (Bar Chart)
- Attrition vs Job Satisfaction (Scatter/Bar)

Capture:

- Screenshot of final rendered charts

STEP 7: Train Baseline Model in Weka

Tool: Weka
Click Path:
Classify → Choose → Trees → J48 → Test Options: Use Training Set → Start

Capture:

- Screenshot of classifier output
- Accuracy & confusion matrix

STEP 8: Overfitting Activity

Goal: Create an overfitted model
Tool: Weka
Click Path:
Classify → Trees → J48 → More Options → minNumObj=1, unpruned=True → Start

Capture:

- Training set accuracy screenshot
- Cross-validation accuracy screenshot

Clue:

Training accuracy very high + CV accuracy significantly lower = Overfitting.

STEP 9: Underfitting Activity

Goal: Create a model too simple to generalize
Tool: Weka
Click Path:
Classify → **Rules** → **ZeroR** → **Start**

Capture:

- ZeroR model output
- Evaluation results (low accuracy)

Clue:

ZeroR predicts only the majority class → Underfitting.

STEP 10: Summarize Findings

Tool: Napkin AI
Task:
Paste model results and ask:
"Summarize key insights from these model outputs and suggest improvements."

Capture:

- Screenshot of Napkin AI summary

---

5. Clues & Hints Section

- High training accuracy but low test accuracy → *Overfitting*
- Low accuracy everywhere → *Underfitting*

- Fewer promotions + high overtime often drive attrition
- Simpler models generalize better
- Compare cross-validation accuracy for model selection

---

6. Expected Learning Outcomes

Students will learn:

- End-to-end ML workflow
- Dataset cleaning & preparation
- Creating visual stories
- Building ML models using Weka
- Diagnosing overfitting & underfitting
- Interpreting attrition predictors
- Preparing a professional analytics report

---

7. Submission Checklist

Students must combine all screenshots into a **single PDF**.

Submit: Combine all below items into a single pdf and submit.

- Google Sheets screenshots (loading, cleaning, exporting)
- ARFF Viewer screenshot
- Weka preprocessing screenshot