A. Project Overview

A1. Research Question or Organizational Need

Research Question:
Are ensuring that employees are well compensated and find satisfaction in their jobs the leading factors in preventing employee turnover?

Organizational Need:
High employee turnover incurs significant costs, including recruitment expenses, training costs, and loss of institutional knowledge. This project seeks to identify key drivers of employee attrition and develop predictive models to assist HR departments in implementing targeted retention strategies.

A2. Scope of the Project

- **Data Sources:** The IBM HR Analytics Employee Attrition & Performance dataset from Kaggle.
    - **Three related published works**
        1. **"A Systemic Review on Important of Employee Turnover with Special Reference to Turnover Strategies"**
        2. **"How to build a Winning people strategy"**
        3. **"Rewriting Employee Engagement"**
- **Data Volume:** 1,470 employee records with 35 features related to demographics, job roles, salaries, and performance indicators.

Scope of Analysis:

- Identifying key factors influencing turnover.
- Using machine learning models to predict employee attrition risk.
- Providing HR leadership with actionable retention strategies.

A3. Overview of the Solution

- **Tools Used:**
    - **Programming and Analytics:** Python with Pandas, NumPy, Sci-kit-learn, Matplotlib, Seaborn libraries.
    - **Data Storage:** CSV Files
    - **Development Environment:** Jupyter Notebook
- **Tools Used:**
    - **Exploratory Data Analysis:** Identified correlations, outliers, and distributions.

- **Predictive Modeling:** Developed and evaluated Logistic Regression and Naïve Bayes models.
- **Evaluation Metrics:** Accuracy, Precision, Recall, ROC-AUC
- **Visualization:** Heatmaps, Feature Importance Charts, and ROC Curves for model performance insights.

Outcome:
Key predictors of employee turnover such as compensation and job satisfaction were identified, and actionable HR retention strategies were formulated.

B. Project Execution

B1. Project Plan Execution

Phase	Planned	Actual Execution	Differences
Data Collection	Used IBM Dataset	No external data was added	The original dataset was sufficient
Data Cleaning	Minimal missing values expected	Addressed minor inconsistencies	Some categorical variables required manual encoding
EDA	Correlation analysis and visualizations	Completed as planned	No major changes
Modeling	Logistic Regression and Naïve Bayes	Successfully Implemented	Considered random forest but excluded for simplicity
Evaluation	Accuracy, Precision, Recall, ROC-AUC	Metrics Calculated	Slight adjustments in hyperparameter tuning
Reporting	Final report and presentation	Delivered as planned	No major deviations

B2. Project Planning Methodology

Methodology Used: CRISP-DM (Cross-Industry Standard Process for Data Mining)

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment (Actionable Insights & Retention Recommendations)

B3. Project Timeline and Milestones

Milestone	Planned Duration	Planned Start	Planned End	Actual Duration	Actual Start	Actual End	Differences from Original
Data Collection and Cleaning	3 Days	1/27/25	1/30/25	3 Days	1/27/25	1/30/25	None
EDA	3 Days	1/31/25	2/2/25	3 Days	1/31/25	2/4/25	None
Model Development and Evaluation	3 Days	2/2/25	2/5/25	3 Days	2/2/25	2/5/25	None
Strategy Formulation and Recommendations	3 Days	2/6/25	2/9/25	3 Days	2/6/25	2/9/25	None
Report Writing and Presentation Prep	5 Days	2/10/25	2/15/25	5 Days	2/9/25	2/11/25	4 Days Early
Final Review and Submission	3 Days	2/15/25	2/18/25	3 Days	2/11/25	2/11/25	7 Days Early

Milestones were able to be completed earlier than expected due to more time per day being allocated to the project and working in parallel to complete multiple tasks at once.

C. Methodology

C1. Data Selection and Collection

- **Differences from the Plan:**
  - **Planned vs. Actual:** No external datasets were incorporated as the IBM HR dataset was sufficient.
  - **Challenges:** Some categorical variables required manual encoding.
  - **Data Governance Issues:** No personally identifiable information (PII) was included in the dataset, ensuring compliance with ethical and privacy standards.

C1a. Advantages and Limitations of the Dataset

- **Advantages:**
  - Rich feature set covering compensation, job roles, satisfaction, and tenure.
  - Pre-cleaned dataset with minimal missing values.
- **Limitations:**
  - Static dataset does not capture real-time employee behaviors.
  - Imbalanced classes with fewer employees who left vs. those who stayed

D. Data Extraction and Preparation

The dataset used for this project was the IBM HR Analytics Employee Attrition & Performance dataset, which contained 1,470 employee records with 35 features covering demographics, job roles, compensation, performance metrics, and engagement indicators. The dataset was relatively clean, with no missing values, making it suitable for analysis. The target variable, "Attrition", indicated whether an employee left the company (*Yes/No*). Before analysis, the dataset was inspected for duplicates, inconsistencies, and categorical variables that required encoding.

To ensure compatibility with machine learning models, categorical variables such as Job Role, Department, and Marital Status were converted into numerical representations using one-hot encoding and label encoding. Additionally, unnecessary columns, such as unique employee IDs, were removed to avoid redundancy. Feature scaling was applied to continuous numerical variables, such as Monthly Income, Age, and Years at Company, using standardization to prevent any one feature from disproportionately influencing model performance.

Finally, the dataset was split into training and testing sets, with 80% allocated for training and 20% for testing, ensuring the models could be evaluated on unseen data. This structured data preparation process helped eliminate inconsistencies, improve model accuracy, and optimize computational efficiency. By applying appropriate preprocessing techniques, the dataset was transformed into a structured and machine-readable format, setting a strong foundation for predictive analysis.

E. Data Analysis Process

E1. Methods Used to Analyze the Data

1. **Exploratory Data Analysis (EDA)**:
    - Used visualizations to assess distributions, correlations, and outliers.
2. **Predictive Modeling**:
    - **Logistic Regression**: Interpretable coefficients, well-suited for binary classification and easy to communicate to HR stakeholders.
    - **Naïve Bayes**: Simple probabilistic model that provides a quick baseline comparison.

E2. Advantages and Limitations of the Tools & Techniques

- **Advantages:**
    - **Python**: Rich ecosystem for data cleaning and visualization.
    - **Logistic Regression**: Interpretable, standard for binary classification.
    - **Naïve Bayes**: Simple, fast, good for baseline checks.

- **Limitations:**
  - **Naïve Bayes**: Assumes conditional independence among features, which may not fully hold.
  - **Logistic Regression**: May not capture complex patterns if relationships are highly non-linear.

E3. Step-by-Step Application of Analytical Methods

Exploratory Data Analysis (EDA)

EDA was performed to understand the dataset, detect patterns, and check for potential issues. Summary statistics and visualizations, such as heatmaps and boxplots, helped identify key predictors like compensation, job satisfaction, and career growth opportunities. The class distribution showed an imbalance, with more employees staying than leaving, influencing model selection and evaluation strategies.

Feature Selection & Preprocessing

To prepare the data for modeling, categorical variables were encoded using one-hot encoding and label encoding, while continuous numerical variables were standardized to ensure equal contribution to model training. The dataset was split into 80% training and 20% testing to evaluate model performance on unseen data.

Model Training & Evaluation

Two models were trained: Logistic Regression and Naïve Bayes. Logistic Regression was chosen for its interpretability, while Naïve Bayes served as a simple, probabilistic baseline. Both models were assessed using accuracy, precision, recall, and ROC-AUC scores. A confusion matrix was used to analyze correct and incorrect classifications.

Business Implications

The models identified compensation and job satisfaction as the strongest predictors of turnover. Logistic Regression provided clear, interpretable coefficients, while Naïve Bayes offered probability estimates for identifying at-risk employees. Findings were translated into HR retention strategies, such as salary adjustments, engagement initiatives, and career development programs.

This structured analytical approach ensured data-driven insights were used to develop actionable solutions for reducing employee turnover.

F. Results

F1. Evaluating Model Output

Model	**Accuracy**	**Precision**	**Recall**	**ROC-AUC**
Logistic Regression	**88%**	**59%**	**41%**	**79%**
Naïve Bayes	**69%**	**25%**	**67%**	**73%**

F2. Practical Significance

- Identified key turnover predictors: compensation and job satisfaction.
- HR can now proactively adjust salaries and retention initiatives.

F3. Overall Success & Effectiveness

- **Success Indicators:**
  - Achieved stable accuracy and precision.
  - Model insights confirmed initial hypothesis: Compensation and job satisfaction were key predictors.
  - The results were interpretable, enabling straightforward recommendations for HR policy improvements.

G. Key Takeaways

G1. Conclusions from the Analysis

- Employees with lower compensation and lower job satisfaction are at notably higher risk of turnover.
- Promotion opportunities and tenure also play roles but appear secondary to the top two factors.

G2. Why the Chosen Tools and Visuals Support Effective Storytelling

- **Logistic Regression Coefficients**: Provide a clear view of how each predictor affects turnover odds, facilitating stakeholder understanding.
- **Bar Charts, Heatmaps, and ROC Curves**: Deliver accessible visual summaries of correlation, feature importance, and model performance.

G3. Two Recommended Courses of Action

1. **Review and Adjust Compensation Structures**
   - Align salaries with market standards to reduce attrition related to pay dissatisfaction.
2. **Enhance Job Satisfaction Initiatives**
   - Introduce continuous feedback loops, mentorship programs, and clear career pathways to boost day-to-day satisfaction.

H. Project Summary Presentation (Panopto Recording)

- **Panopto Link**: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=efb5d480-1e8a-454a-85e0-b28101818332

I. Appendices

1. **Code Used for Analysis**
   - See attached file "Wesley_Hilton_Employee_Attrition_Capstone.pdf" for all code and workflow for the analysis.
2. **Data Sources**
   - IBM HR Analytics Employee Attrition & Performance: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

J. References

1. **IBM.** (n.d.). *HR Analytics Employee Attrition & Performance* [Data set]. Kaggle. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
2. **Shaba, R., & Chakrabarty, P.** (2023). *A systematic review on importance of employee turnover with special reference to turnover strategies* [Conference paper]. ResearchGate. https://www.researchgate.net/publication/367344941_A_Systematic_Review_on_Importance_of_Employee_Turnover_with_Special_Reference_to_Turnover_Strategies
3. **BetterUp.** (n.d.). *How to build a winning people strategy.* https://www.betterup.com/blog/people-strategy
4. **Society for Human Resource Management (SHRM).** (n.d.). *Rewriting employee engagement.* SHRM Executive Network. https://www.shrm.org/executive-network/insights/rewriting-employee-engagement