SMIT
SAYLANI MASS IT TRAINING

Artificial Intelligence and Data Science Specialization

Batch - 09
Miss Javeria Hassan
Hajiyani Sakeena Campus
Mid-Hackathon

Task 01:
Time Allowed: 6 hours

AI-based Resume Screening System

You will build a complete machine-learning pipeline to analyze resumes and predict whether a candidate should be shortlisted, and visualize all steps using Streamlit.
Dataset:
https://www.kaggle.com/datasets/mdtalhask/ai-powered-resume-screening-dataset-2025/data

Part 1: Data Understanding & Cleaning

1. What does each column in the dataset represent?
2. How many rows and columns are there?
3. Which columns contain missing values?
4. How did you handle missing values?
5. Are there duplicate records? If yes, what did you do?
6. Are there any obvious outliers? How did you treat them?

Part 2: Basic Analysis

1. What is the overall distribution of shortlisted vs non-shortlisted candidates?
2. What are the key numerical and categorical variables in the dataset?

Part 3: Univariate Analysis

1. Analyze the distribution of all numerical features.
2. Analyze the frequency of all categorical features.

3. What insights can you draw from individual feature distributions?

Part 4: Bivariate Analysis

Answer the following 15 questions using visualizations and comments:

1. How does the AI Score differ between shortlisted and rejected candidates?
2. Is experience higher for shortlisted candidates?
3. Does education level affect shortlisting?
4. How does the number of skills relate to recruiter decisions?
5. Relationship between the number of projects and shortlisting?
6. Does salary expectation influence the decision?
7. Compare AI Score across different job roles.
8. Correlation between experience and AI Score?
9. Are certain certifications more common among shortlisted candidates?
10. Does education level impact AI Score?
11. Compare experiences across job roles.
12. Identify salary outliers and their impact.
13. Compare projects count by recruiter decision.
14. Which numerical features are most correlated with the target?
15. What overall patterns help predict shortlisting?

Part 5: Data Preprocessing

1. How did you prepare categorical variables for modeling?
2. How did you prepare numerical variables for modeling?
3. What transformations were applied before training models?

Part 6: Model Building

- Apply cross-validation on the dataset.
- Train any 5 classification models.
- Compare model performance using:
  - Accuracy
  - Precision
  - Recall
  - F1-Score
- Which model performed best and why?

Part 7: Visualization & Insights

- Create clear visualizations for:
  - Data analysis
  - Model performance comparison
  - Confusion matrices
- Write meaningful comments under each visualization.

Part 8: Streamlit Application

- Build a Streamlit app that displays:
  - Dataset overview
  - All analysis visuals
  - Model comparison results
  - Evaluation metrics
- Add a prediction section where users can input resume details and see the result. Ensure every output and visualization appears on the UI.

Final Requirement

- Briefly explain how your system can help improve resume screening in real-world hiring.

Task 02

Skill-Based Task Assignment

Dataset: https://www.kaggle.com/datasets/umerfarooq09/skill-based-task-assignment/data

Part 1: Data Understanding & Cleaning

- What are all the columns in the dataset, and what do they represent?
- How many rows and columns are in the dataset?
- Which columns have missing values?
- How did you treat missing values?
- Are there duplicate rows? If yes, how did you handle them?
- Are there any incorrect or inconsistent entries? How did you fix them?

Part 2: Basic Data Summary

- Show the count and percentage of each class in the target variable (task assignment outcome).
- Which features are numerical and which are categorical?

Part 3: Univariate Analysis

- Visualize the distribution of each numerical feature.
- Visualize the frequency of each category for all categorical features.
- What are the notable observations from each single-feature analysis?

Part 4: Bivariate Analysis

12. How does the target variable relate to each numerical feature?
13. Do certain skill levels or task types occur more often with specific outcomes?
14. Is there a relationship between experience/skill values and task assignment outcomes?
15. How do two numerical features relate to each other (e.g., skill level vs task completion time)?
16. Does the distribution of outcomes differ across different task categories?
17. Compare average values of one numerical feature across different categorical groups.
18. What patterns emerge when comparing two categorical features?
19. Which features show a strong correlation with the target?
20. Use a heatmap to show correlation among numerical features.
21. Is there a trend between the complexity of the task and the assignment decision?
22. Identify any strong interactions between features that affect the assignment outcome.

Part 5: Preprocessing

- How did you encode categorical variables for modeling?
- What scaling or normalization did you apply to numerical features?
- How did you handle any feature imbalance or skewness in data?

Part 6: Modeling

- Use cross-validation to evaluate model performance before final training.
- Train **any 5 classification models** for predicting task assignment.
- Compare models using:
    - Accuracy
    - Precision
    - Recall
    - F1-Score

Part 7: Visualization & Comparison

- Show confusion matrices for all models.
- Create comparison charts for all evaluation metrics.
- Which model performed best? Explain why based on metrics and visuals.

Part 8: Streamlit UI

- Display dataset overview on Streamlit UI.
- Show all univariate and bivariate visualizations in the app.
- Present model performance comparisons and metrics.
- Add a **prediction interface** where users can input feature values and see the predicted assignment.
- Ensure every output (tables, charts, metrics, and prediction results) appears clearly in the UI.

Final Insight

- Summarize how this task assignment model could be used in a real workplace to improve efficiency or fairness.