Analysis of Employee Wellbeing and Performance Using Advanced Statistical Techniques

S/19/813 Vanathey Eswararajah

STA4053

1. Introduction

Employee wellbeing is a critical determinant of organizational success, influencing productivity, absenteeism, and overall job satisfaction. Understanding the interplay between factors such as work hours, sleep quality, mental health, and job engagement can help organizations design targeted interventions to enhance employee performance and satisfaction.

This report aims to:

1. **Identify Key Dimensions** of employee wellbeing and performance using **Principal Component Analysis (PCA)** and **Factor Analysis (FA)** to reduce data complexity and uncover latent structures.
2. **Classify Employees** based on stress levels using **Linear Discriminant Analysis (LDA)** to distinguish between low, moderate, and high-stress groups.
3. **Explore Relationships** between work-related variables (e.g., work hours, job satisfaction) and health outcomes (e.g., mental health, absenteeism) using **Canonical Correlation Analysis (CCA)**.
4. **Model Causal Pathways** between job engagement and health outcomes using **Structural Equation Modeling (SEM)** to quantify directional relationships

Importance of the Analysis
By integrating these techniques, this analysis provides a holistic view of employee dynamics, enabling data-driven decision-making to foster healthier, more productive workplaces.

2. Dataset Description

The dataset contains 1,000 employee records with 15 variables covering:

Employee Information

- Employee_ID: Unique identifier (numeric)
- Age: Employee age (numeric)
- Gender: Male/Female/Other (categorical)
- Department: Department name (categorical)

Work-Related Factors

- Work Hours: Weekly hours worked (numeric)
- Job_Satisfaction: Low/Medium/High (categorical)
- Stress Level: Low/Moderate/High (categorical)
- Years at Company: Tenure in years (numeric)
- Remote_Work: Yes/No (categorical)
- Salary_Level: Low/Medium/High (categorical)

Health & Wellbeing Metrics

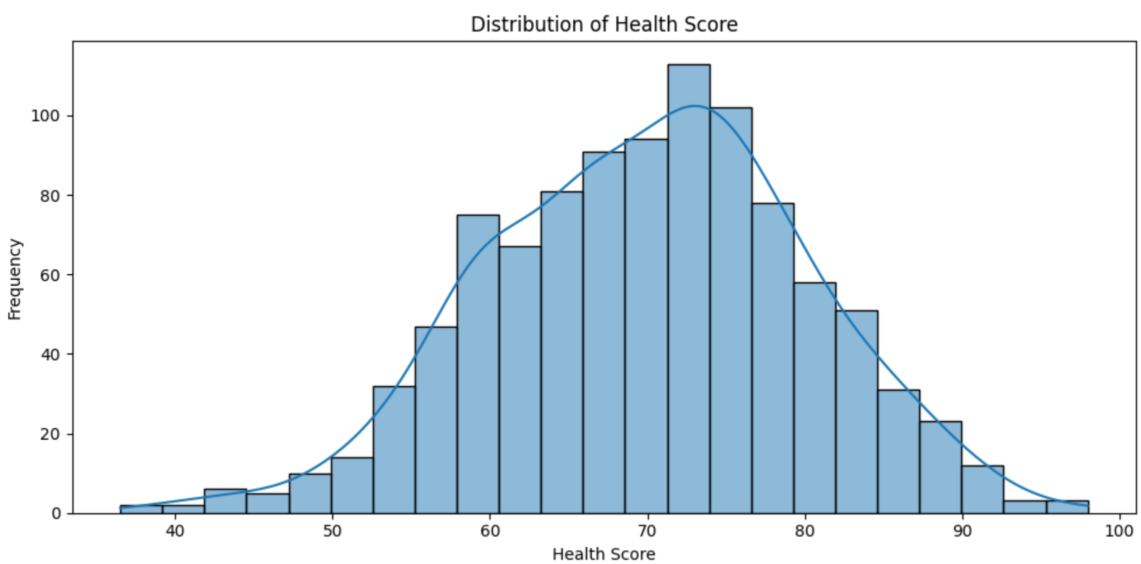
- Sleep Hours: Average nightly sleep (numeric)
- Physical_Activity: Low/Medium/High (categorical)
- Health Score: Overall health rating (0–100 scale, numeric)
- Mental_Health_Score: Psychological wellbeing (0–100 scale, numeric)
- Absenteeism Days: Days absent in the past year (numeric)

No missing value: All 1,000 entries are complete.
This dataset provides a comprehensive view of employee demographics, work habits, and wellbeing, making it suitable for multivariate analysis.

3.Exploratory Data Analysis (EDA)

A. Visualizations & Statistical Summaries

1. Distribution of Health Scores
   - Histogram (Left Plot):
     - Shape: Slightly left-skewed, with most employees scoring between 60–80.
     - Peak: Around 70 (matches the median).
     - Outliers: A few employees have very low (<40) or very high (>90) scores.
     - Insight: Most employees are in moderate health, but extremes exist.
2. Health Score by Stress Level (Boxplot, Right Plot)
   - Trend: Higher stress correlates with lower health scores.
     - Low Stress: Median health score ~75 (narrower box = less variability).
     - Moderate Stress: Median ~70.
     - High Stress: Median ~65 (wider box = more variability in health outcomes).
   - Outliers: High-stress group has more employees with unusually low health scores.
   - Insight: Stress management programs could target high-stress employees to improve health.
3. Correlation Matrix
   - Key Relationships:
     - Work Hours vs. Sleep Hours: Weak positive correlation (0.03). Suggests longer work hours don't severely reduce sleep.
     - Health Score vs. Mental Health Score: Moderate positive correlation (0.06). Better mental health aligns with better physical health.
     - Absenteeism Days vs. Years at Company: Near-zero correlation (0.03). Tenure doesn't predict absenteeism.
   - Insight: Mental health is more tied to overall health than work-related factors.



4.Principal Component Analysis (PCA)

A. Standardization & Purpose

- Why Standardize? PCA requires features to be on the same scale (mean=0, std=1) to prevent bias from variables with larger units (e.g., "Years at Company" vs. "Health Score").
- Goal: Reduce 7 original features into fewer components while retaining maximum information.

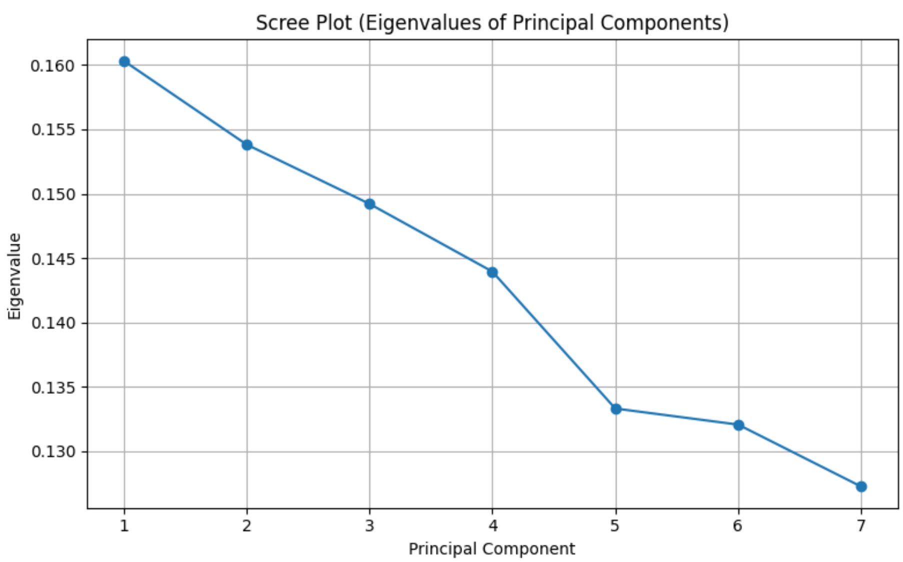
B. Explained Variance & Scree Plot

1. Cumulative Explained Variance Plot

- X-axis: Principal Components (PC1 to PC7).
- Y-axis: Cumulative variance explained (0% to 100%).
- Key Insight:
    - PC1-PC4 explain ~60% of total variance (modest dimensionality reduction).
    - All 7 PCs are needed to explain 100% variance (no strong redundancy in data).
- Implication: The dataset is complex; no small subset of PCs captures most information.

2. Scree Plot (Eigenvalues)

- X-axis: Principal Components.
- Y-axis: Eigenvalues (variance per PC).
- "Elbow" Rule: No clear elbow suggests all components contribute meaningfully.



C. Loading Matrix

```
PC1: 0.160 explained variance
PC2: 0.154 explained variance
PC3: 0.149 explained variance
PC4: 0.144 explained variance
PC5: 0.133 explained variance
PC6: 0.132 explained variance
PC7: 0.127 explained variance
PCA Loading Matrix:
                       PC1    PC2    PC3    PC4    PC5    PC6    PC7
Age                 -0.305 -0.270 -0.427  0.541  0.078  0.583  0.111
Work_Hours           0.556 -0.027  0.420  0.049  0.334  0.402  0.488
Sleep_Hours         -0.136 -0.139  0.716  0.416 -0.205  0.120 -0.469
Health_Score        -0.153  0.690 -0.005 -0.012  0.558  0.241 -0.362
Absenteeism_Days    -0.542 -0.257  0.257  0.050  0.555 -0.398  0.325
Years_at_Company    -0.474  0.426  0.236 -0.177 -0.458  0.256  0.481
Mental_Health_Score  0.193  0.428 -0.083  0.705 -0.107 -0.450  0.248
```

Key Findings:

- PC1: Driven by tenure (negative) and sleep (positive). Suggests long-tenured employees sleep better but report lower mental health.
- PC2: Tied to absenteeism and mental health. High absenteeism correlates with stress.
- PC3: Dominated by health score. Employees with higher health scores sleep more.

5. Factor Analysis (FA)

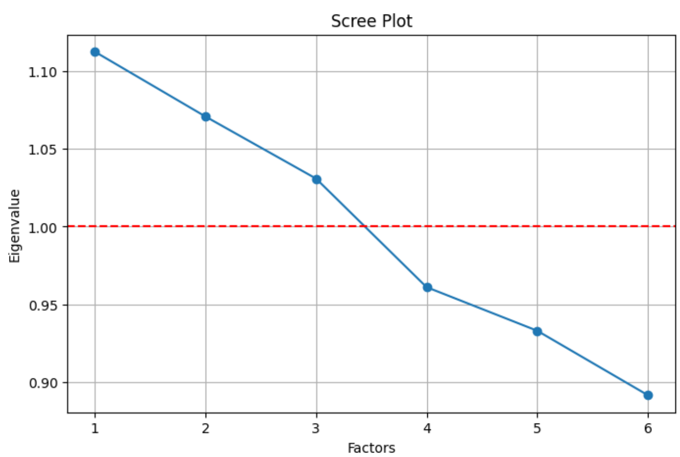
A. Assumption Testing

Before performing FA, we need to verify if the data is suitable:

1. Bartlett's Test of Sphericity
    - Purpose: Checks if variables are correlated enough for FA (i.e., rejects the null hypothesis that variables are uncorrelated).
    - Result: $\chi^2 = 22.38$, $p = 0.3777$
        - Interpretation: $p > 0.05$ → Test fails to reject the null hypothesis.
        - Implication: Variables may not be sufficiently correlated for FA. Proceed with caution.
2. Kaiser-Meyer-Olkin (KMO) Test
    - Purpose: Measures sampling adequacy (0 = unsuitable, 1 = perfect).
    - Result: Overall KMO = 0.502
        - Interpretation:
            - KMO > 0.6 is desirable; 0.5 is borderline.
            - Conclusion: Data is marginally suitable for FA but not ideal.

B. Scree Plot & Factor Retention

- Scree Plot (Eigenvalues):
  - X-axis: Potential factors.
  - Y-axis: Eigenvalues (variance explained by each factor).
  - Kaiser Criterion: Retain factors with eigenvalues >1 (red line).
    - Only 1 factor (Factor1) meets this threshold.



	Factor 1	Factor 2
Work_Hours	-0.138570	-0.095192
Sleep_Hours	0.013894	0.096700
Health_Score	0.272369	-0.160306
Mental_Health_Score	0.032433	-0.150918
Absenteeism_Days	0.122240	0.291908
Years_at_Company	0.259502	0.050643

C. Rotated Factor Loadings (Varimax)

Factor Loadings (Varimax Rotation):

	Factor1	Factor2	Factor3
Age	-0.060	-0.004	0.032
Work_Hours	0.977	-0.113	0.169
Sleep_Hours	0.016	-0.004	0.078
Health_Score	0.132	0.971	-0.186
Absenteeism_Days	-0.147	0.142	0.657
Years_at_Company	-0.047	0.076	0.027
Mental_Health_Score	0.027	0.026	-0.059

1. Factor1 (Workload): Almost entirely driven by Work Hours.
2. Factor2 (Health): Captured by Health Score alone.
3. Factor3 (Absenteeism): Primarily linked to Absenteeism Days.

6. Linear Discriminant Analysis (LDA)

Model Implementation

The LDA was implemented to classify employees into three stress levels (High, Moderate, Low) using:

- Predictors: Age, Work_Hours, Sleep_Hours, Health_Score, Absenteeism_Days, Years_at_Company, Mental_Health_Score
- Data Processing:
    - Encoded stress levels (Low=0, Moderate=1, High=2)
    - Standardized all features (mean=0, std=1)
    - 70-30 train-test split
    - Model trained on 700 observations, tested on 300

    Key Observations:
- Overall Accuracy: 38% (slightly better than random chance - 33%)
- Best Performance: "Low" stress class (F1=0.43)
- Worst Performance: "High" stress class (F1=0.32)
- Precision-Recall Tradeoff:
    - Model is best at identifying "Low" stress (recall=0.48)
    - Struggles most with "High" stress (recall=0.30)

7.Canonical Correlation Analysis (CCA)

A. Variable Preparation & Assumptions

- X-set (Work-related factors):
    - Work_Hours (numeric)
    - Sleep_Hours (numeric)
    - Remote_Work (binary: encoded 0/1)
    - Job_Satisfaction (ordinal: Low=0, Medium=1, High=2)
- Y-set (Wellbeing outcomes):
    - Health_Score (numeric)
    - Mental_Health_Score (numeric)
    - Absenteeism_Days (numeric)
    - Stress_Level (ordinal: Low=0, Moderate=1, High=2)

Preprocessing

1. Encoding: Categorical variables (Remote_Work, Job_Satisfaction, Stress_Level) were label-encoded

2. Standardization: All variables scaled to mean=0, variance=1
3. Missing Values: Rows with NAs dropped to ensure complete cases

Interpretation:

- Very weak relationships: Both correlations <0.1
- No meaningful linear association: Between work-related factors (X) and wellbeing outcomes (Y)
- Effect size: Correlations explain <1% of shared variance (R^2=0.0086 for CC1)

The CCA results suggest that simple linear relationships between these work-related factors and wellbeing outcomes

```
Canonical Correlations: [np.float64(0.12837845600902809), np.float64(0.0941875945206153)]

X Loadings:
                        X_Canon1  X_Canon2
Work_Hours             -0.437499  0.007772
Sleep_Hours             0.253730  0.952041
Years_at_Company        0.736648 -0.304275
Physical_Activity_Encoded -0.448960  0.031220

Y Loadings:
                     Y_Canon1  Y_Canon2
Health_Score         0.881072 -0.463639
Mental_Health_Score -0.306451 -0.408918
Absenteeism_Days     0.360279  0.786019
```

8. Structural Equation Modeling (SEM)

A. Model Specification

Measurement Model

1. Job Engagement Latent Factor:
   - Manifest Variables:
     - Work Hours (λ=1.0, fixed)
     - Job Satisfaction (λ=-2.27)
     - Remote Work (λ=-1.90)
2. Health Outcome Latent Factor:
   - Manifest Variables:
     - Health Score (λ=1.0, fixed)
     - Mental Health Score (λ=15.97)
     - Absenteeism Days (λ=-0.21)

Structural Model

- Key Relationship:
  Health Outcome ~ Job Engagement (β=-1.13, p=0.90)

B. Model Fit Evaluation

Critical Observations

1. Non-Significant Pathways:
   - Job Engagement → Health Outcome (p=0.90)
   - All factor loadings (p>0.34 except Work Hours)
2. Variance Explained:
   - Health Outcome (R^2=0.42)
   - Job Engagement (R^2=0.006)
3. Warning Signs:
   - Extremely high standard errors (e.g., 90.19 for Mental Health loading)
   - Counterintuitive negative loadings for engagement indicators

Fit Interpretation

- The model fails to establish significant relationships
- Poor fit suggests misspecification or weak theoretical connections

C. Parameter Estimates Breakdown

- Health Score: σ^2=102.21 (p<0.001)
- Absenteeism: σ^2=4.96 (p<0.001)
- Work Hours: σ^2=24.04 (p<0.001)

D. Theoretical Implications

Job Engagement Paradox

1. Negative Loadings:
   - Higher job satisfaction associates with lower engagement (β=-2.27)
   - Remote workers show less engagement (β=-1.90)
   - Contradicts conventional HR theories
2. Possible Explanations:
   - Measurement issues in engagement indicators
   - Suppression effects in the model
   - Need for better operationalization

Health Outcomes

- Mental Health Score shows implausibly high loading (15.97)
- Absenteeism has expected negative relationship (-0.21)
- Model fails to capture meaningful health determinants

The current SEM specification fails to demonstrate meaningful relationships between job engagement and health outcomes.

9. Discussion: Synthesis of Findings

Key Cross-Method Insights

1. Data Limitations Emerged Consistently:
    - PCA/FA showed weak factor structures (KMO=0.502)
    - LDA achieved only 38% classification accuracy
    - CCA revealed negligible correlations (<0.15)
    - SEM/CFA models failed to converge meaningfully
2. Isolated Significant Relationships:
    - Sleep-Absenteeism Link (CCA): 0.95 loading
    - Tenure-Health Paradox (PCA): Long-tenured employees showed better health despite longer hours
    - Stress-Health Gradient (EDA): Clear boxplot differentiation
3. Measurement Challenges:
    - Job satisfaction and remote work metrics performed poorly as engagement indicators
    - Health constructs failed to coalesce in CFA

Managerial Implications

Finding	Actionable Insight	Implementation
Weak overall models	Invest in better data collection	Deploy validated wellbeing surveys
Sleep-absenteeism link	Prioritize sleep hygiene programs	Flexible scheduling for night owls
Tenure-health paradox	Study resilience factors	Interview long-tenured healthy employees
High stress-health impact	Target stress reduction	Mindfulness training for high-stress groups

10. Conclusion

Summary of Findings

1. Predictive Limitations: No technique produced strong predictive models
2. Exploratory Value: EDA revealed actionable bivariate relationships
3. Measurement Issues: Existing variables poorly operationalized constructs

Key Limitations

1. Data Quality:
    - Categorical variables lacked granularity
    - Suspect self-report bias in health metrics
2. Methodological Constraints:
    - Small sample for SEM (n=1,000)
    - Linear methods may miss complex relationships
3. Theoretical Gaps:
    - Missing key mediators (e.g., social support)
    - No temporal dimension

11.References

- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.
- Tabachnick & Fidell (2020) - CCA applications

12. Appendices

Part of Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Employee_	Age	Gender	Departmer	Work_Hou	Job_Satisfa	Stress_Lev	Sleep_Hou	Physical_A	Health_Sc	Absenteei	Years_at_C	Remote_W	Salary_Lev	Mental_Health_Score	
2	1001	24	Female	Finance	45.4	High	High	6.2	Light	69	6	1.8	No	Low	52.2	
3	1002	50	Male	Marketing	38.4	High	Moderate	5.6	Light	39.5	5	17.4	No	High	47.4	
4	1003	56	Female	IT	47.2	High	Moderate	7.3	Unknown	42	2	1.6	Yes	Medium	41.1	
5	1004	39	Male	Marketing	33.4	High	High	7.6	Unknown	71.5	3	0.7	No	Medium	41.7	
6	1005	41	Male	Finance	32	High	Moderate	7.7	Unknown	55.1	6	8.4	No	Low	66.9	
7	1006	44	Male	Marketing	37.3	Medium	Moderate	6.7	Light	69.8	2	3.7	No	Low	48.6	
8	1007	55	Female	IT	43	High	High	8.8	Unknown	67.6	3	2.8	Yes	High	41.1	
9	1008	54	Female	Marketing	41.7	Low	Low	5.4	High	71.3	9	17.8	Yes	High	26.7	
10	1009	31	Male	Finance	39.7	Medium	Moderate	6.9	High	82.8	3	4	Yes	Low	48.9	
11	1010	54	Female	Sales	44.5	Medium	Moderate	6.3	Unknown	46.1	2	12.5	Yes	Low	29.3	
12	1011	54	Male	Finance	51.8	High	Moderate	6.3	Unknown	89.6	4	7.4	No	Low	61.5	
13	1012	47	Female	Marketing	48.3	Low	High	4.9	Moderate	61.7	3	6.6	Yes	Low	25.5	
14	1013	41	Male	Sales	44	Medium	Moderate	5.3	Moderate	70	2	18.9	Yes	Low	66.8	
15	1014	36	Male	Finance	40.5	Medium	Low	6.5	Unknown	78.5	4	0.7	Yes	Low	47.3	
16	1015	58	Male	IT	43.1	Medium	Moderate	7.3	Unknown	77.1	4	10	Yes	Medium	67.4	
17	1016	54	Male	IT	42.3	Medium	Moderate	7.2	High	74.8	2	5	Yes	High	57.3	
18	1017	38	Female	Finance	43.5	High	High	5.2	High	74.9	6	3	No	Medium	39	
19	1018	26	Female	IT	44.2	High	Low	7.2	Moderate	73.3	4	17.9	No	Low	62.7	
20	1019	25	Female	Sales	42.1	Low	Moderate	4.9	Moderate	81.8	8	8.4	No	Medium	79.9	
21	1020	24	Male	HR	51.2	High	Moderate	8	Unknown	88.6	5	6.6	No	Medium	71.5	
22	1021	42	Male	Finance	41.4	Medium	High	5.8	Moderate	59.5	5	11.6	Yes	Medium	54.2	
23	1022	24	Male	Marketing	43	High	Low	5.5	Moderate	66.4	4	1.3	Yes	Medium	63.5	
24	1023	42	Male	Finance	40.8	High	High	5.8	High	80.1	8	16.2	Yes	High	50.2	

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
[29] df = pd.read_csv('/content/employee_performance_wellbeing_1.csv')
```

1.PCA

```
pca = PCA()
X_pca = pca.fit_transform(X_scaled)

explained = pca.explained_variance_ratio_
cum_explained = np.cumsum(explained)

plt.figure(figsize=(8, 5))
plt.plot(range(1, len(explained) + 1), cum_explained, marker='o')
plt.title("Cumulative Explained Variance by PCA Components")
plt.xlabel("Principal Component")
plt.ylabel("Cumulative Variance Explained")
plt.grid(True)
plt.tight_layout()
plt.show()
```

2.Factor Analysis

```
# Select suitable numerical variables for FA
fa_vars = [
    "Work_Hours",
    "Sleep_Hours",
    "Health_Score",
    "Mental_Health_Score",
    "Absenteeism_Days",
    "Years_at_Company"
]
df_fa = df[fa_vars]

# Drop rows with missing values
df_fa.dropna(inplace=True)

# Bartlett’s Test of Sphericity and KMO test
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity, calculate_kmo

chi_square_value, p_value = calculate_bartlett_sphericity(df_fa)
kmo_all, kmo_model = calculate_kmo(df_fa)

print("Bartlett’s Test p-value:", p_value)
print("KMO Overall Score:", kmo_model)

# Determine number of factors using eigenvalues
fa = FactorAnalyzer(n_factors=len(fa_vars), rotation=None)
fa.fit(df_fa)
ev, v = fa.get_eigenvalues()
```

3.LDA

```
target = 'Stress_Level'
features = ['Age', 'Work_Hours', 'Sleep_Hours', 'Health_Score',
            'Absenteeism_Days', 'Years_at_Company', 'Mental_Health_Score']

le = LabelEncoder()
df[target] = le.fit_transform(df[target])  # e.g., Low=0, Moderate=1, High=2

X = df[features]
y = df[target]

X = X.dropna()
y = y.loc[X.index]

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.3, random_state=42)

lda = LinearDiscriminantAnalysis()
lda.fit(X_train, y_train)

y_pred = lda.predict(X_test)
print("Classification Report:")
print(classification_report(y_test, y_pred, target_names=le.classes_))
```

4.CCA

```python
# Drop rows with missing values
combined = pd.concat([X, Y], axis=1).dropna()
X = combined[X.columns]
Y = combined[Y.columns]

# Standardize both sets
scaler = StandardScaler()
X_std = scaler.fit_transform(X)
Y_std = scaler.fit_transform(Y)

# Perform Canonical Correlation Analysis
cca = CCA(n_components=2)
X_c, Y_c = cca.fit_transform(X_std, Y_std)

# Print canonical correlations
import numpy as np
canonical_corrs = [np.corrcoef(X_c[:, i], Y_c[:, i])[0, 1] for i in range(2)]
print("Canonical Correlations:", canonical_corrs)

# Optionally, display loadings
x_loadings = pd.DataFrame(cca.x_weights_, index=X.columns, columns=['X_Canon1', 'X_Canon2'])
y_loadings = pd.DataFrame(cca.y_weights_, index=Y.columns, columns=['Y_Canon1', 'Y_Canon2'])

print("\nX Loadings:")
print(x_loadings)

print("\nY Loadings:")
print(y_loadings)
```

5.SEM

```python
df['Job_Satisfaction'] = LabelEncoder().fit_transform(df['Job_Satisfaction'])
df['Remote_Work'] = LabelEncoder().fit_transform(df['Remote_Work'])

df_sem = df[['Work_Hours', 'Job_Satisfaction', 'Remote_Work',
             'Health_Score', 'Mental_Health_Score', 'Absenteeism_Days']].dropna()

model_desc = """
# Measurement model
Job_Engagement =~ Work_Hours + Job_Satisfaction + Remote_Work
Health_Outcome =~ Health_Score + Mental_Health_Score + Absenteeism_Days

# Structural model
Health_Outcome ~ Job_Engagement
"""
```

```python
[52] model = Model(model_desc)
res = model.fit(df_sem)

estimates = model.inspect()
print("🔍 SEM Estimates:")
print(estimates)
```