Data Mining Final Project Report

Jasmine Lewis

CIS 4750-21 Data Mining

December 13, 2024

1. Introduction

Objective

The objective of this project is to analyze the impact of remote work on employees' mental health and productivity using a comprehensive dataset. The analysis uses data mining techniques to identify patterns, insights, and provide actionable recommendations that can help organizations improve employee well-being.

Summary of Proposal and Progress

The project began with a proposal to examine the relationship between remote work and mental health, focusing on stress levels, productivity changes, and access to mental health resources. Milestones included dataset selection, exploratory data analysis (EDA), and application of data mining techniques. Progress has been consistent, with a focus on ensuring clean data, robust analysis, and actionable insights.

2. Dataset Overview

Dataset Description

The dataset contains information on employees' remote work experiences and mental health conditions.

- **Source**: Remote Work & Mental Health 🌍🧠
- **Key features include**:
  - **Demographics**: Age, Gender
  - **Professional Details**: Job_Role, Industry, Years_of_Experience
  - **Work Environment**: Work_Location, Hours_Worked_Per_Week, Number_of_Virtual_Meetings
  - **Mental Health Indicators**: Stress_Level, Mental_Health_Condition, Access_to_Mental_Health_Resources
  - **Lifestyle Metrics**: Physical_Activity, Sleep_Quality
  - **Geographic Region**: Region

- **Size**: The dataset includes 500 rows and 20 columns

Relevance

The dataset is ideal for analyzing the proposed problem due to its rich feature set, covering both objective metrics (e.g., hours worked) and subjective metrics (e.g., stress levels). Its diversity in roles, industries, and regions makes it suitable for generalizable insights.

3. Data Preparation and EDA

Data Cleaning and Preprocessing

- **Missing Values**: Rows with missing values were deleted for the sake of simplifying the analysis
- **Normalization**: Applied scaling to numeric variables to ensure compatibility with clustering algorithms.
- **Duplicates**: Performed a search of duplicates and deleted if existed.
- **Feature Engineering**: Created derived features, such as Work_Intensity (hours worked per meeting)
-

Exploratory Data Analysis

- **Descriptive Statistics**: The average Work_Life_Balance_Rating was 3.2, with Stress_Level skewed toward "High" in onsite work locations.
- **Visualizations**:
    - Bar charts showing Mental_Health_Condition distribution by Work_Location.
    - Heatmaps highlighting correlations between numerical features (Age, Hours Worked, Social Isolation Rating, and Work Life Balance Rating).
    - Cluster plots depicting segmentation based on work-life balance and productivity.

Insights and Challenges

- Remote workers reported higher satisfaction but also higher isolation scores.
- Disproportionate data in categories like Mental_Health_Condition posed challenges for analysis.

- Data appeared to be synthetic and not completely trustworthy for real world solutions. Found errors such as age 27 and years worked 20. These types of errors were persistent within the dataset and made working with it difficult.

4. Application of Data Mining Techniques

Techniques Used

1. **Clustering**:
    - **Rationale**: To segment employees based on mental health and productivity metrics.
    - **Implementation**: Applied K-means clustering using features like Stress_Level, Work_Life_Balance_Rating, and Satisfaction_with_Remote_Work.
    - **Result**: Identified three distinct clusters: "High Stress, Low Satisfaction," "Balanced," and "Low Stress, High Satisfaction."
2. **Anomaly Detection**:
    - **Rationale**: To identify outliers in productivity changes and stress levels.
    - **Implementation**: Used isolation forests to detect anomalies in Productivity_Change.
    - **Result**: Highlighted a small group of employees with extreme productivity decreases and high stress levels.

5. Model Evaluation and Validation

Evaluation Metrics

- **Clustering**: Evaluated using the silhouette score (average score: 0.65), indicating moderate cluster separation. This score reflects the degree of cohesion within clusters and separation between clusters.
- **Anomaly Detection**: Validated results by cross-referencing anomalies with high-stress employees and extreme hours worked. The isolation forest algorithm flagged employees who significantly deviated from typical patterns, ensuring interpretability and actionability.

Validation Techniques

- **Clustering Validation**:

    - Conducted multiple runs of clustering with different initializations.
    - Used the elbow method to determine the optimal number of clusters, balancing between over-segmentation and meaningful grouping.
- **Anomaly Detection Validation**:
    - Adjusted hyperparameters such as contamination rate in the isolation forest to fine-tune sensitivity.
    - Cross-referenced flagged anomalies with domain knowledge and data trends to ensure they aligned with logical expectations.

Additional Considerations

- **Comparison of Techniques**:
    - Clustering provided broad insights into group behaviors, aiding in the identification of high-risk and low-risk employee segments.
    - Anomaly detection offered granular insights, pinpointing specific employees or scenarios requiring immediate attention.
- **Strengths and Limitations**:
    - **Clustering**: Strength in segmentation and group-level insights, but relies on well-defined feature selection.
    - **Anomaly Detection**: Effective in identifying outliers but susceptible to noise in data.

6. Business Insights and Recommendations

Insights

- Employees in hybrid roles had the best work-life balance, while onsite workers faced higher stress levels.
- Access to mental health resources positively impacted productivity and satisfaction.

Recommendations

1. **Enhance Hybrid Work Models**: Promote hybrid setups where feasible.
2. **Increase Access to Mental Health Resources**: Provide resources to employees reporting high stress levels.

3. **Monitor Productivity Anomalies**: Use automated systems to flag significant changes in productivity and address root causes.

Future Work

- Apply more advanced techniques for predictive modeling.

References:

Kaggle, Google, Youtube