**Data Mining and Predictive Analytics**

**Final Project**

Vishwak Balaji Jayasankar (Z23765142)

1. Explain the problem and purpose of the data mining task in at least 3 sentences.

This project aims to predict employee burnout based on various demographic and work-related factors. By applying data mining techniques, I strive to understand key factors contributing to employee burnout and develop predictive models that can help organizations take preventive actions. The goal is to create models that accurately predict burnout scores for new, unseen employees.

2. Obtain the dataset for analysis: Paste the link (URL) from where you obtained the dataset

URL: <https://www.kaggle.com/code/mazharjamdar/employee-burnout/input?select=train.csv>

I obtained this dataset from Kaggle. In the original dataset, the size is around 22k, but I can’t use this much size in the model because of the restriction of the licence where I can you only less than or equal to 10k. So, I edited it to 10k dataset and created two sheets of one with ‘original data’ and another with ‘New Data’, where it has only 50 entries for the scoring.

3. Explore, clean, and preprocess data:

a. Cleanup any column that is not a predictor:

* Dropped **Employee\_ID** column as it is just an identifier and not a predictor.
* Dropped **Date\_of\_Joining** as it did not provide direct predictive value without heavy feature engineering (e.g., tenure).

b. Perform ‘Missing Data Handling’: (Common for both the algorithms)

A computer screen with a white box

AI-generated content may be incorrect.

* For categorical variables like **Gender**, **Company\_Type**, Designation, and **WFH\_Setup\_Available**, missing values were filled using the **mode.**
* For numerical variables like **Resource\_Allocation**, and **Mental\_Fatigue\_Score**, missing values were filled using the **mean.**
* **For the target numerical variable like Burn\_Rate, missing values are deleted from the record.**

c. For continuous numerical outcome: Explore using Scatter Plot Matrix between continuous numerical predictors and the outcome (and remove predictors that are not correlated to the outcome). If you have categorical outcome then you do not need to perform Scatter Plot.

A screenshot of a graph

AI-generated content may be incorrect.

d. Do you have any categorical variables ? Do you need to convert them to dummies ? Yes or No ...explain why ?

A screenshot of a computer

AI-generated content may be incorrect.

* Yes, there were categorical variables: **Gender**, **Company\_Type**, and **WFH\_Setup\_Available**.
* These were **converted into dummy variables** because machine learning models require numerical inputs, and categorical variables must be encoded appropriately to retain their meaning.

4. Reduction of data dimension – Use PCA technique or Classification/Regression Tree technique on any one of the models in both the algorithms to reduce the dimensions. Select top 5 dimensions. Explain the technique you used.

A screenshot of a computer

AI-generated content may be incorrect.

* **Principal Component Analysis (PCA)** was applied to reduce the dimensionality of the dataset.
* PCA helped to identify the top 5 principal components that explain most of the variance in the data.
* The Top 5 selected are Mental\_Fatigue\_Score, WFH\_Setup\_Available\_Yes, Gender\_Female, Company\_Type\_Service, Designation and it cover 98% of variance.

5. Partition the data accordingly

* When using Linear Regression, I did a partition of 6 different models with various train and validation split methods to test the model accuracy.
* When using KNN, I used 3 NN and 5 NN with 60/40, 70/30, 80/20 split on training and validation. Which helps me get the best accuracy.
* Also, here in the partition, I only put the top 5 pca features as an input, and the target variable was added as well.

6. Choose the data mining techniques/algorithm to apply Classify or Predict and explain  
the reason for selecting the two algorithms

Two different data mining algorithms were selected:

* **K-Nearest Neighbors (KNN) Regression**: A non-parametric method useful for capturing complex patterns.
* **Linear Regression**: A parametric method that models the relationship between predictors and the target linearly.
* These algorithms were chosen to compare the performance of a **non-linear** model (KNN) and a **linear** model (Linear Regression) in predicting employee burnout.

7. Try at least 6 models for ‘each’ of the two algorithms and select the best model by interpreting results of algorithm and explain your model selection criteria. (Provide Screenshot of Model Table)

A screenshot of a computer

AI-generated content may be incorrect.

I tried six models for both Linear Regression and KNN by changing partition ratios and checking performance. I selected the best model for each algorithm based on the highest Validation R² and the lowest Validation RMSE.

8. Deploy the two models (best one from each algorithm) which means you perform ‘Score’ on New Data using both the algorithms (one by one) and compare and explain the predicted results

Linear Regression:

A screenshot of a computer

AI-generated content may be incorrect.

KNN:

A screenshot of a computer

AI-generated content may be incorrect.

Comparing the results of both result accuracy, which was calculated by me, shows Linear Regression as 86% and KNN as 83%. This proves that Linear Regression performed well.