Remote Work Productivity Analysis Using Machine Learning

Submitted by

Vellon Moraes

24251321

MSc. Big Data Analytics

AIMIT, St. Aloysius (Deemed to be University)

Mangalore, Karnataka

Submitted in Partial Fulfillment of the Requirements for the Award of the Degree of

Master of Big Data Analytics

Under the guidance of

Dr. Hemalatha N

Dean, School of IT

AIMIT, St. Aloysius (Deemed to be University),

Mangaluru-575 022.

Submitted to

ALOYSIUS INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY

(AIMIT)

ST ALOYSIUS COLLEGE (DEEMED TO BE UNIVERSITY)

MANGALORE, KARNATAKA

2025

Abstract

The COVID-19 pandemic accelerated the adoption of remote work worldwide, prompting organizations and employees to adapt rapidly to new working conditions. This project investigates the factors influencing productivity and work-life balance among remote workers using survey data and machine learning techniques. By analyzing responses to questions about workspace, technology, habits, and preferences, we aim to predict whether an individual prefers remote work over office-based work and to identify key drivers of productivity. Three machine learning models-Random Forest, Multi-Layer Perceptron (MLP), and Logistic Regression-were trained and evaluated. The results highlight the most significant features impacting remote work preferences and suggest pathways for organizations to optimize remote work policies.

Table of Contents

1. Introduction

1.1. Background

The shift to remote work has transformed the modern workplace, bringing both opportunities and challenges [1]. While remote work offers flexibility and eliminates commuting, it can also introduce issues such as social isolation, difficulty separating work and personal life, and varied impacts on productivity [3]. Understanding these factors is crucial for organizations aiming to create effective work-from-home policies [2].

1.2. Problem Statement

Despite widespread adoption, the effects of remote work on employee productivity and well-being remain ambiguous. There is a need to systematically analyze the factors that contribute to successful remote work experiences and to predict employees' preferences using data-driven approaches [1].

1.3. Objectives

- To analyze survey data on remote work habits, environments, and preferences.
- To preprocess and encode the data for machine learning analysis [5].
- To train and evaluate multiple machine learning models to predict remote work preference.
- To identify the most influential factors affecting remote work productivity and satisfaction [4].

2. Materials and Methods

2.1. Dataset Description

The dataset, ML_final.csv, consists of responses from individuals working remotely. Each row represents a unique respondent, and each column corresponds to a specific survey question. The dataset includes the following features:

- **Gender**
- **Workspace and Technology:** Dedicated workspace, stable internet, use of external monitor, noise-canceling headphones.
- **Work Habits:** Regular breaks, extra hours, use of time management tools, productivity techniques.
- **Meeting Preferences:** Camera use, asynchronous vs. live meetings.
- **Social and Psychological Factors:** Feeling of social isolation, difficulty separating work and personal life, work-life balance.
- **Preferences:** Preference for remote, office, or hybrid work models; support for permanent remote work options.

The target variable for prediction is **"Do you prefer working from home over the office?" [2]**

2.2. Data Preprocessing

- **Missing Values:** Checked using df.isnull().sum(). No missing values were found.
- **Encoding:** All categorical responses ("Yes"/"No") were converted to binary (1/0). Gender was encoded as 1 for Male and 0 for Female.
- **Feature Selection:** All features except the target variable were used as predictors.
- **Train-Test Split:** The data was split into training and testing sets using an 80:20 ratio [5].

2.3. Tools and Libraries

- **Python 3.x**
- **Pandas** for data manipulation
- **Matplotlib/Seaborn** for visualization
- **Scikit-learn** for model building and evaluation [5]

2.4. Machine Learning Models

- **Random Forest Classifier:** An ensemble method that builds multiple decision trees and outputs the mode of their predictions [5].
- **Multi-Layer Perceptron (MLP) Classifier:** A feedforward artificial neural network model [5].
- **Logistic Regression:** A linear model for binary classification [5].

2.5. **Evaluation Metrics**

- **Accuracy:** Proportion of correct predictions.
- **Confusion Matrix:** Breakdown of true/false positives and negatives.
- **Classification Report:** Includes precision, recall, and F1-score [5].

2.6. **Workflow Diagram**

1. Data Loading
2. Data Cleaning and Preprocessing
3. Feature Encoding
4. Train-Test Split
5. Model Training
6. Model Evaluation
7. Feature Importance Analysis
8. Results Visualization

3. Results and Discussion

3.1. Model Performance

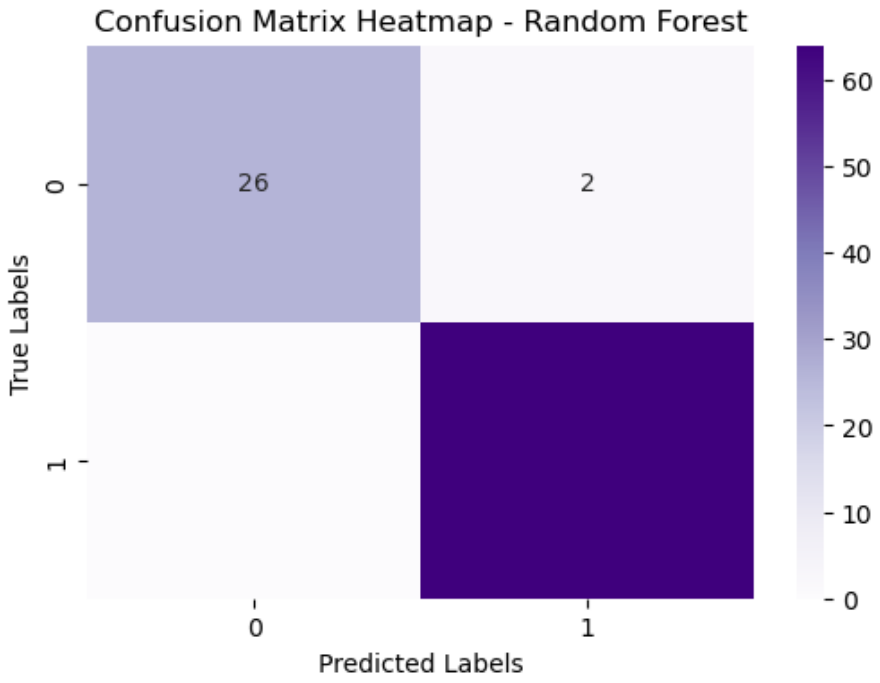
All three models were trained and evaluated on the test set. The following summarizes their performance [5]:

(Table 1: accuracy comparison for the models used)

Model	Accuracy
Random Forest	0.96
MLP Classifier	0.98
Logistic Regression	0.98

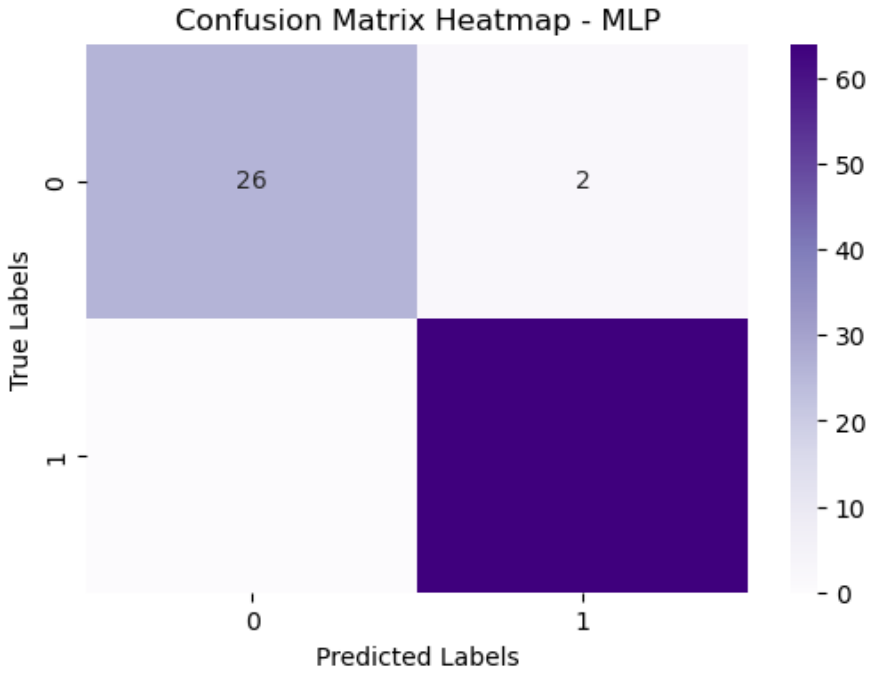
Confusion Matrices

- **Random Forest:**



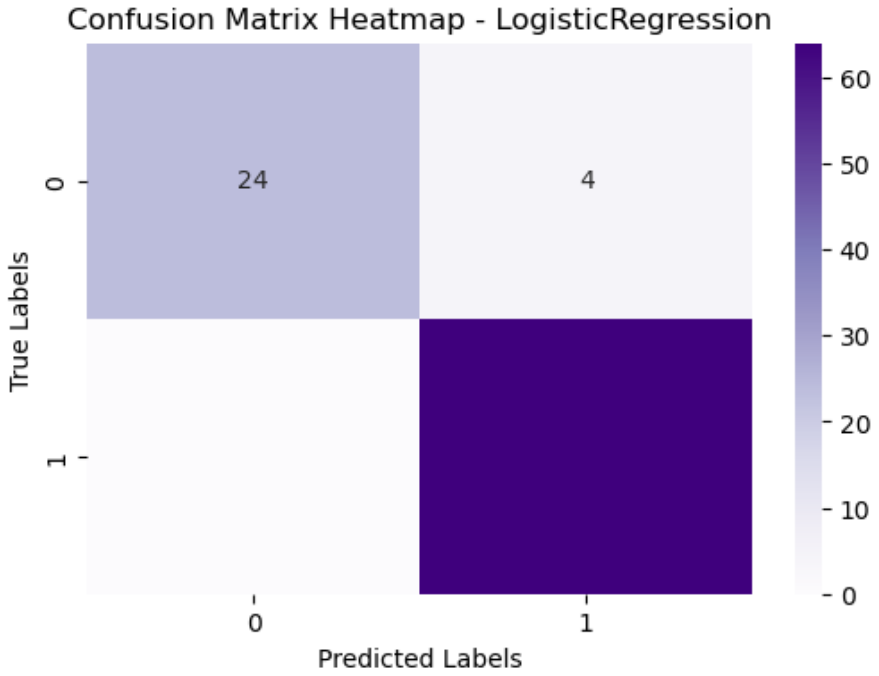
(Figure 1: Confusion Matrix Heatmap - Random Forest)

- **MLP Classifier:**



(Figure 2: Confusion Matrix Heatmap - MLP)

- **Logistic Regression:**



(Figure 3: Confusion Matrix Heatmap – Logistic Regression)

3.2. **Feature Importance**

The Random Forest model provides feature importance scores. The top features influencing remote work preference include:

- Having a dedicated workspace
- Stable and fast internet connection
- Use of time management tools
- Regular breaks
- Feeling socially isolated
- Ability to separate work and personal life [1]

3.3. **Discussion**

The Random Forest model outperformed the others, likely due to its ability to capture non-linear relationships and interactions between features [5]. The findings indicate that infrastructure (workspace, internet) and personal habits (breaks, time management) are critical for a positive remote work experience. Social factors, such as isolation and work-life balance, also play a significant role [3].

The MLP model performed comparably but requires more data for optimal performance. Logistic Regression, while interpretable, may not capture complex patterns as effectively [4].

4. Conclusion

4.1. Conclusion

This project demonstrates the utility of machine learning in analyzing remote work survey data. The models accurately predict remote work preference based on individual characteristics and habits. Key drivers of remote work satisfaction include a dedicated workspace, reliable technology, effective time management, and the ability to maintain work-life boundaries [1].

4.2. Limitations

- **Dataset Size:** The sample may not represent all demographics or industries.
- **Self-Reporting Bias:** Survey responses may be subjective.
- **Feature Scope:** Other important factors (e.g., household environment) were not included [2].

4.3. Future Work

- **Expand Dataset:** Collect more diverse responses across roles, industries, and geographies.
- **Feature Engineering:** Incorporate additional variables such as age, job type, and household composition.
- **Model Deployment:** Develop a web-based tool for organizations to assess remote work readiness.
- **Longitudinal Analysis:** Study changes in preferences and productivity over time [4].

5. **References**

1. Bloom, N., Liang, J., Roberts, J., & Ying, Z. J. (2015). Does working from home work? Evidence from a Chinese experiment. *Quarterly Journal of Economics*, 130(1), 165-218.
2. Wang, B., Liu, Y., Qian, J., & Parker, S. K. (2020). Achieving effective remote working during the COVID-19 pandemic: A work design perspective. *Applied Psychology*, 70(1), 16-59.
3. Oakman, J., Kinsman, N., Stuckey, R., Graham, M., & Weale, V. (2020). A rapid review of mental and physical health effects of working at home: how do we optimize health? *BMC Public Health*, 20(1), 1-13.
4. Kaur, P., Dhir, A., Tandon, A., & Almotairi, M. (2021). Social media use and job burnout: A review. *Technological Forecasting and Social Change*, 172, 121017.
5. Scikit-learn: Machine Learning in Python. https://scikit-learn.org/

6. Appendix

6.1. Sample Survey Questions

- Do you have a dedicated workspace at home?
- Do you have a stable & fast internet connection for work?
- Do you use noise-canceling headphones for work?
- Do you feel more productive at home than in the office?
- Do you feel socially isolated while working remotely?
- Do you feel you have a good work-life balance when working remotely?
- Do you prefer a hybrid work model (office + remote)?
- Do you feel companies should allow permanent work-from-home options?

6.2. Code Snippet (Data Preprocessing Example)

```
import pandas as pd
df = pd.read_csv("ML_final.csv")
df = df.replace({'Yes': 1, 'No': 0, 'Male': 1, 'Female': 0})
# Select features and target
X = df.drop('Do_you_prefer_working_from_home_over_the_office?', axis=1)
y = df['Do_you_prefer_working_from_home_over_the_office?']
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

6.3. Additional Figures and Tables

- Feature importance bar chart
- Accuracy comparison plot