

Machine Learning Algorithm for Work from Home Analysis during Epidemic (2022)

Vandana Babshetti¹ and Nihar Ranjan²

^{1,2}RSCOE CSE, Pune, India

Email: Vandana.babshetti@gmail.com, nihar.pune@gmail.com

Abstract—Employee productivity is critical to the company's success in achieving its objectives. Predicting future employee productivity is critical for organizations that work from home to flourish. This article used machine learning to estimate work from home employee productivity in a corporation. The machine learning approach is based on a cross-industry data mining standard (CRIPS-DM). We have analyzed some variables which are important for productivity prediction. The random forest method is used in the machine learning process. The random forest method is employed to analyze the prediction model. This will help organizations to improve work from home by using analyzed results. The results reveal that the random forest method properly identified instances in 91.66 percent of the cases.

Index Terms— Covid-19 epidemic, Random Forest, Decision tree, Productivity, Remote Working, Machine learning.

I. INTRODUCTION

Many occupations rely on computers and telecommunications, many people have reliable home internet connections, and many families have all members working full-time. This method was advanced by the Covid-19 epidemic, which forced a large percentage of the global force to return to WFH as quickly as feasible. WFH has the ability to reduce travel time, allow more flexible operation hours, improve worker happiness, and create a better work-life balance when compared to working from the factory (WFO). Still, little is known about the abecedarian impacts of WFH, as well as its productivity consequences and the factors that influence whether WFH is more or less productive than WFO. We tend to provide an associate degree examination of WFH productivity during a massive IT trade in this paper. In March 2020, the association was forced to return all of its employees to WFO from WFH. The employees are highly skilled experts working in a data technology firm where a substantial percentage of labor has historically been mechanized. They're also some of the most difficult to probe at the same time. Important cerebral point work, communication with groups, working with customers, and invention are all part of the job. For related professions, productivity is difficult to maintain. There has never been a study of WFH for vocations with similar features. This knowledge offers a unique opportunity to gain a productive life for this difficult type of work and to deconstruct the factors of productivity across WFH. Nonetheless, WFH productivity varied depending on worker variables we want to analyze, such as whether children were present and family time. Still, we prefer to dissect the level of collaboration, networking, operation, and training based on the nature of the activity. Our research of the productivity differences between WFH and the UN agency gives critical information about the issues that are likely to be most important when developing future WFH. A classification algorithm is proposed in this paper to predict employee work from home

productivity using machine learning. The CRIPS-DM methodology is used in the machine learning process. The employee's productivity is predicted using a random forest classification technique. The organization can use the forecast to determine which employees are deserving of attention.

II. RELATED WORKS

There have been various studies and many methods have been implemented for the analysis of datasets but the productivity of work from home had not been analyzed using machine learning algorithms. There is a lot of manual and statical analysis present. Here we are going to discuss the various algorithms and methods that have been studied and implemented in earlier studies for analysis.

As machine learning techniques, Decision Tree, Logistic Regression, Random Forest, Nave Bayes, and the neural network approach ANN were used. They came to the conclusion that the ANN model was the most precise and accurate. We used decision trees, logistic regression, and random forests. When all three approaches are tested, the random forest classifier outperforms the logistic regression and decision tree. The focus is primarily on data analysis and prediction in the real world. The proposed technique, which employs a random forest algorithm, can improve the accuracy of identification prediction. The performance of the approaches is assessed using accuracy, sensitivity, specificity, and precision. They test the algorithms SVM, NB, KNN, Logistic Regression, and Random Forest to see which one performs best for analysis. They chose Random Forest and the KNN method based on accuracy or MCC's greatest value.

Various machine learning algorithms for data prediction and analysis have been discussed. It also presents an overview of many types of data analysis and machine learning methods like SVM, Logistic Regression, Random Forest, Neural Network, and Decision Tree. They also stated that random forest prediction has a higher level of accuracy. They have presented a machine learning-based system. They concentrated on logistic regression techniques and offered suggestions for logistic function and data preparation for logistic regression.

III. PROPOSED MODEL

The project's major goal is to analyze work from home productivity. Machine learning algorithms are very helpful for the same as they are more accurate than the traditional methods. Machine learning algorithms are also self-learning and can identify patterns with minimal human intervention. Figure 1 shows proposed the model architecture.

A. Data Collection

We are using a dataset obtained from the real-time collection using google forms for this project. The dataset covers analysis variables that help to predict work from home productivity. The dataset is used 80% for train data and 20 % for test data.

B. Data Exploration

In this project, we will be utilizing a random forest to predict work from home productivity. This dataset contains work hours at the office and home as well as other parameters such as task complete count, and collaboration with teammates, which helps to analyze productivity. Some of these features will be used to train a random forest classifier to predict the productivity of a particular employee based on the total tasks assigned and completed. The data in this demo comes from Google Sheets and below Figure 2 is a preview of what the data looks like.

C. Data Analysis

Clarity of the data being used is very important for working on any project. Tabular illustrations of the data give much clear information about the same. Figure 3 shows the tabular illustrations of the data.

D. Data Analysis

Clarity of the data being used is very important for working on any project. Tabular illustrations of the data give much clear information about the same. Figure 3 shows the tabular illustrations of the data.

E. Data Processing

Data is getting in terms of objects which we have converted in terms of the numerical format. Also, we have filled null data with dummy values which helps for accuracy.

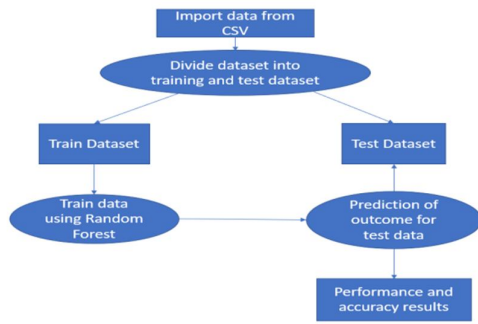


Figure 1. System Architecture

	How much time you will work after working hours To WFH mode	How much time you will work after working hours To WFH mode	How satisfied are you with the frequency of communication from leadership?	How easy or difficult are you finding it to work effectively in the current situation?	How external parameter affects on your work like electricity, internet	How satisfied are you working from home?
count	58.000000	58.000000	58.000000	57.000000	58.000000	58.000000
mean	1.844028	2.705897	3.534403	3.385965	3.327586	3.395552
std	1.151884	1.284300	1.029657	1.081573	1.526238	1.268613
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	1.000000	2.000000	3.000000	3.000000	2.000000	3.000000
50%	1.000000	3.000000	4.000000	3.000000	4.000000	3.000000
75%	2.000000	3.750000	4.000000	4.000000	5.000000	4.750000
max	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000

Figure 2. Data preview

Timestamp	The size of the data set	How much time you will work after working hours To WFH mode	How much time you will work after working hours To WFH mode	How satisfied are you with the frequency of communication from leadership?	How easy or difficult are you finding it to work effectively in the current situation?	How external parameter affects on your work like electricity, internet	How satisfied are you working from home?	How much time you will work after working hours To WFH mode	How much time you will work after working hours To WFH mode	How satisfied are you with the frequency of communication from leadership?	How easy or difficult are you finding it to work effectively in the current situation?	How external parameter affects on your work like electricity, internet	How satisfied are you working from home?
04-06-2022 22:01	Yes	5	5	Yes	3	Yes	Yes	1.0	Yes	5	5	Yes	5
1 4/14/2022 10:57:50	Yes	2	2	No	4	Yes	Yes	5.0	Yes	4	Yes	4	Yes
2 4/14/2022 10:58:52	Yes	1	1	Yes	5	Yes	Yes	5.0	Yes	10	3	Yes	5
3 4/14/2022 11:20:31	No	1	1	No	2	No	No	2.0	Yes	5	2	Maybe	1
4 4/14/2022 11:59:06	Yes	1	3	Yes	5	Yes	Yes	3.0	Yes	8	1	Maybe	5

Figure 3 Data describe

F. Training the model

The Random Forest and decision tree algorithms are used to create the model. The Random Forest technique was chosen because it offers numerous benefits over other classification algorithms. The dataset is divided training and testing data using the Sklearn tool. 80 percent of the data is used to train the model during the training phase, whereas 20% is used to test it subsequently.

G. Model Evaluation

The test dataset is used to put the trained model to the test. The accuracy (the fraction of correct predictions made by the model), precision (the number of correct positive results divided by the number of positive results predicted by the classifier), recall (the number of correct positive results divided by the number of all relevant samples) and other evaluating factors are used to evaluate the results using the metrics module on Sklearn.

IV. METHODOLOGY

The methodology used in this study is CRIPS-DM. CRIPSDM provides a set of defined procedures for completing machine learning projects. It is a well-known methodology that may be applied with a variety of tools and procedures in a variety of industries. Data understanding, modeling, business understanding, data preparation, implementation, and evaluation are the six phases [13]. Risk assessment, success criteria, planning, business objectives, and the cost-benefit of the machine learning assignment are all thoroughly examined in the business understanding step. The goal of this study is to lower the rate of employee turnover at a lower cost. Machine learning would have required some knowledge due to the large amount of data associated with this issue, as well as the targeted company. It is frequently referred to while progressing to earlier phases, and it can improve the overall quality of the dataset in order to balance the chosen model. Comprehension of data entails delving into the definition of data as well as its quality. In data preparation, data cleaning is done to make data preparation easier, more accurate, and representative. Meanwhile, modeling entails the activities of selection, model training, model selection, and parameter setup. Finally, evaluation is used to help with model accuracy and other performance measures, as well as whatever success criteria the model has satisfied. The following section of this research explains data comprehension, data preparation, modeling, and evaluation. The model was not implemented. This study is now complete in the CRISP-DM evaluation phase.

A. Ensemble Learning

Ensemble learning combines individual models to increase the model's predictability and stability. This method improves the model's capacity to forecast. It combines numerous machine learning models into a single prediction model with a low error rate that is a powerful classifier.

B. Algorithm Used

Machine learning algorithms, as previously said, play a significant part in data analysis because they work on both classification and regression problems. These algorithms analyze data exceedingly quickly.

C. Random Forest Classifier

Random Forest, an ensemble learning system, employs the bagging approach. It is one of the most often utilized approaches for data analysis. This is due to the fact that it may be used for both regression problems and classification. The phrase 'forest' refers to a collection of decision trees. Each decision tree comes up with a

unique prediction. The candidate with the most votes is chosen. The bagging approach is used to generate a forest of decision trees using this manner.

Working of Random Forest Algorithm:

1. The algorithm chooses N records at random from the dataset.
2. It then uses these records to build decision trees.
3. According to the number of trees specified, the above steps are repeated for different sets of N random records.
4. Finally, each decision tree is used to predict which category the test record belongs to, and the record is assigned to the category with the most votes.

D. Advantages of Random Forest Algorithm

1. Because Random Forest relies on several decision trees to generate its output, it is unbiased and thus more dependable.
2. Because it only works on a random subset of the dataset at a time, Random Forest can be utilized for a huge dataset.

E. Disadvantages of Random Forest Algorithm

1. Because the Random Forest Algorithm works with many decision trees, the processing time may vary depending on the system's power.

V. COMPARATIVE ANALYSIS

For work from home productivity analysis, various methods have been implemented. Large datasets don't function well with the Support Vector Machine (SVM) Algorithm, which necessitates more pre-processing. As a result, the system becomes inefficient because the model's pace is hampered. For large datasets, Logistic Regression can have over-fitting issues.

TABLE 1. TABLE COMPARING THE ACCURACY AND PRECISION OF VARIOUS MACHINE LEARNING ALGORITHMS

Model	Accuracy	Precision
Logistic Regression	93.84	96.58
Decision Tree	92.88	99.48
Random Forest	97.66	98.85
Naive Bayes	91.62	97.09
Support Vector Machine	94.99	95.98

VI. RESULTS AND DISCUSSION

The random forest classifier is giving an accuracy of 91.66666%. We have also used a decision tree classifier and we are getting an accuracy of 66.666666%. The bigger the number of cases, the better the accuracy Improved.

We can see from the first analytical variable that the frequency of leadership communication appears to be satisfactory. It signifies that it is assisting in the increase of productivity. It is depicted in Figure 4 above. When we see that our communication frequency is lower than it should be, we know we need to work on it.

The second study variable is whether or not the employee takes regular breaks, which will enable them to stay fresh and focused on their work, as well as add value to their production. Figure 5 depicts this.

The next analysis variable to look at is the frequency of leadership communication with density. The density is 3-4-5, which is considered typical. It indicates that a worker is content and productive when working from home. Figure 6 depicts this.

The next analysis variable to look at is the frequency of leadership communication with density. The density is 3-4-5, which is considered typical. It indicates that a worker is content and productive when working from home. Figure 6 depicts this.

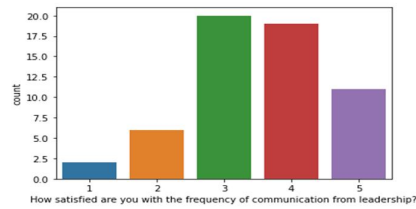


Figure 4. Frequency of Communication

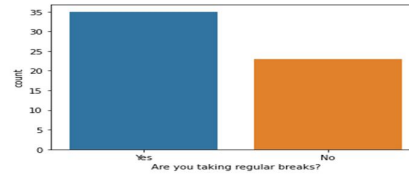


Figure 5. Regular break

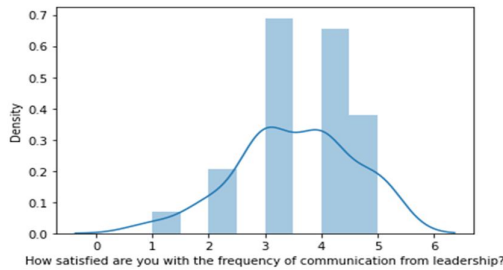


Figure 6 Frequency of communication density

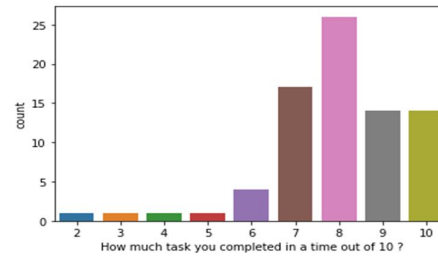


Figure 7. Number of tasks completed

We can also look at the productivity of individual employees. We've gathered a number of jobs performed out of ten to determine an individual employee's productivity. Figure 7 depicts the end result.

We can see from this that the average result is 8. It suggests that overall productivity at home is high. The business is on the right track. We gathered information on work hours in the office and work hours at home. We have data that shows that working from home hours is increasing even after office hours. This also indicates that employees are working longer hours, which will result in increased revenue for the company. If an employee works more hours and takes regular breaks, he or she will produce better work and, as a result, productivity will improve. The violin plot of working hours comparison between WFO and WFH is shown in Figure 8.

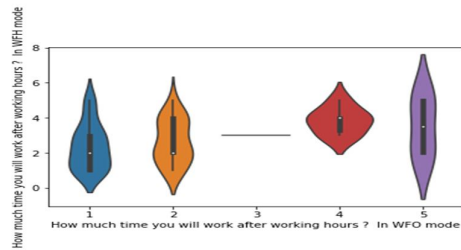


Figure 8. Compare results of WFO and WFH



Figure 9. Working hours of WFO and WFH

In the form of a count plot, we have the results of WFH and WFO's working hours. Figure 9 depicts the situation.

Figure 10 depicts the outcomes of WFH's working hours and tasks completed out of a total of ten.

Finally, contentment with working from home reveals work productivity. Figure 11 depicts this. These results were obtained after a number of regular breaks, and they suggest that there are certain modifications that can be made to assist us analyze and enhancing productivity.

We have also recorded the results of the confusion matrix, which are listed below. The results of the confusion matrix are shown in Figures 12 and 13.

We have analyzed the results of a healthy work-life balance which is positive for working from home. It is shown in Figure 14. We have results of the feature importance tree classifier which helps for better visualization. It is shown in Figure 15.

We have captured correlations of each feature in the dataset using a confusion matrix. It is shown in Figure 16. On the same dataset, we investigated the results with the Decision Tree classifier, but the Random Forest approach had a higher accuracy.

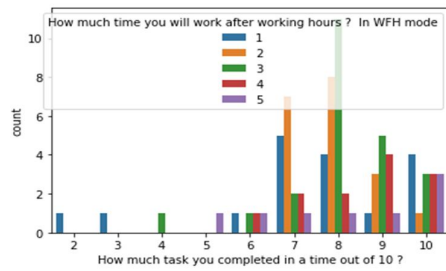


Figure 10. Tasks completed and WFH working hours

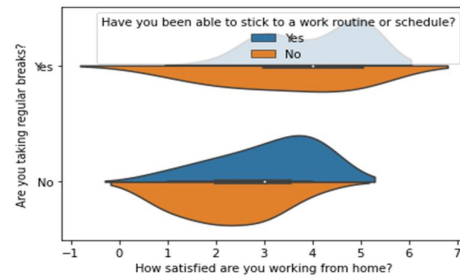


Figure 11. How satisfied WFH

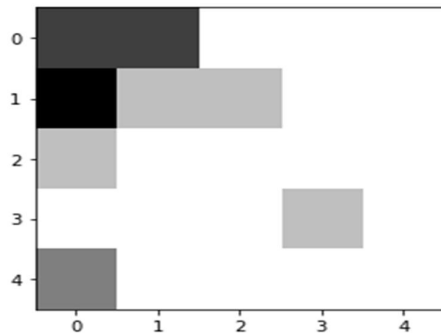


Figure 12. Confusion Matrix

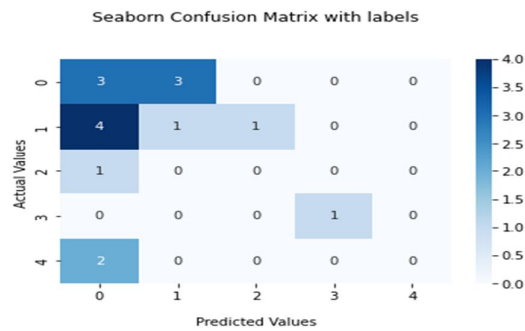


Figure 13: Actual and anticipated values in the Confusion Matrix

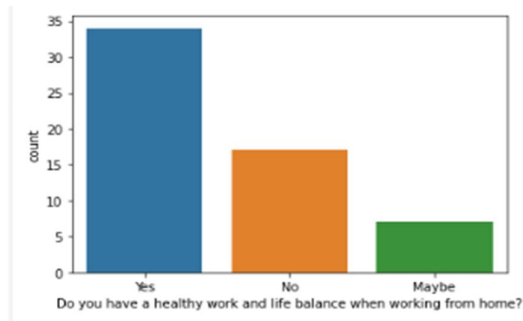


Figure 14. Healthy work-life balance

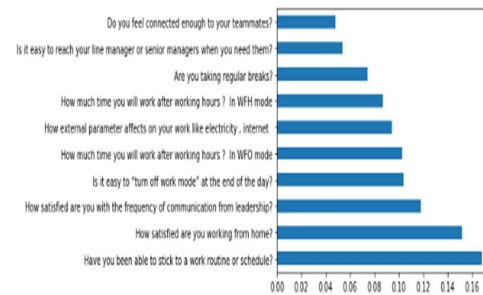


Figure 15. Feature importance tree classifier

A. Decision tree classifier

Accuracy: 0.6666666666666666

B. Random forest classifier

Accuracy of Random Forest: 0.9166666666666666

C. Summarized Result

Accuracy of Random Forest: 0.9166666666666666

Classification Report

VII. CONCLUSIONS

According to this study, employee productivity may be crucial to an organization's growth. They can use machine learning technology to predict employee productivity from remote work before it happens, allowing them to prepare ahead of time for the issue. According to the evaluation, the correctly classified instance using the provided model of regression methods is 91.66 percent. This displays the regression technique's ability to forecast and analyze data. With a big amount of data, the Random Forest Algorithm performs exceptionally well.

	precision	recall	f1-score	support
1	0.500000	1.000000	0.666667	1
2	1.000000	0.909091	0.952381	11
accuracy			0.916667	12
macro avg	0.750000	0.954545	0.809524	12
weighted avg	0.958333	0.916667	0.928571	12

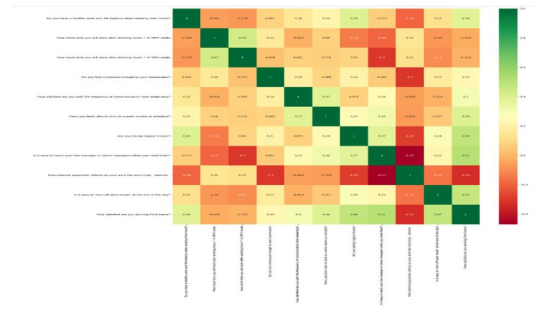


Figure 16. Correlations of each feature in dataset using Confusion matrix

For the same dataset, we have decision classifier findings. With the Random Forest algorithm, we have good accuracy. This model can be used to analyze and predict overall organization and individual employee productivity.

A. Future Scope

The model is solely applicable to the IT industry. This can be further enhanced by including more analytical variables relevant to other industries. In addition, the existing model's dataset can be expanded to improve accuracy. Second, because the technique is Random Forest, it may be merged with a variety of different algorithms to improve performance and accuracy. As a result, the scope of this research is vast in terms of analyzing productivity utilizing various deep learning methods.

ACKNOWLEDGMENT

The authors wish to thank Prof Dr. Nihar Ranjana sir for their valuable guidance and support.

REFERENCES

- [1] Jason Brownlee, "SMOTE for Imbalanced Classification with Python", Machine Learning Mastery.
- [2] <https://www.geeksforgeeks.org/machine-learning/>
- [3] A.Kakade, S. Jain, S. Das, Nihar Ranjan, "A neural network approach for text document classification and semantic text analytics" Journal of data mining and management, Volume 2, Issue 2, pp 1-5, 2017
- [4] W. Fu and P. O. Perry, "Estimating the number of clusters using cross-validation," Journal of Computational & Graphical Statistics, vol. 29, no. 1, pp. 162–173, 2020.
- [5] Nihar Ranjan, Rajesh S. Prasad "Author Identification in Text Mining for used in Forensics" ISSN 2321- 9637, Volume 1, Issue 5, December 2013, pp. 568-571
- [6] F. Goreaud and R. Pelissier, "On explicit formulas of edge ´ effect correction for Ripley's K -function," Journal of Vegetation Science, vol. 10, no. 3, pp. 433–438, 1999.
- [7] W. Fu and P. O. Perry, "Estimating the number of clusters using cross-validation," Journal of Computational & Graphical Statistics, vol. 29, no. 1, pp. 162–173, 2020.
- [8] S. Vassilvitskii and D. Arthur, "K-means++: the advantages of careful seeding," in Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035, New Orleans LA, USA, January 2006.
- [9] M. A. Masud, J. Z. Huang, C. Wei, et al., "I-nice: a new approach for identifying the number of clusters and initial cluster centres," Information Sciences, vol. 466, pp. 129–151, 2018.
- [10] M. Erisoglu, N. Calis, and S. Sakalliolglu, "A new algorithm for initial cluster centers in k-means algorithm," Pattern Recognition Letters, vol. 32, no. 14, pp. 1701–1705, 2011.
- [11] X. Shen, W. Liu, I. Tsang, F. Shen, and Q.-S. Sun, "Compressed k-means for large-scale clustering," in Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco CA USA, February 2017.
- [12] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," in Proceedings of the 2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06), pp. 459–468, IEEE, Berkeley, CA, USA, October 2006.
- [13] A. Clark, The Machine Learning Audit— CRISP-DM Framework, 2018.