Predictive Analytics of Productivity Prediction of Garment Employees

Professor: Ceni Babaoglu

Student Name: Maryam Mortezapourdehkordi

Student Number: 501143277

Date of submission: Dec 5, 2022

Table of Contents

[Abstract 3](#_Toc452457337)

**Introduction......................................................................................................................... 4**

**literature review ………………………………………………………………………….……5**

**Goal ………………………………………………………………………………………...……7**

**Machin Learning Metric ……………………………………………………………………...7**

**Dataset ……………………………………………………………………………………….....9**

**Graph Methodology …………………………………………………………………………14**

**Steps to implement …………………………..……………………………………………15**

**Result ………………………………………………………………………….………………22**

**Reference………………………………………………………………………………………24Appendix……………………………………………………………………………………….30**

# Abstract

As a result of industrial globalization, the garment industry has significantly affected contemporary society. It is a field which employs a vast number of manual procedures and requires much effort. The workers’ performance in garment manufacturing enterprises in terms of production and delivery is mainly responsible for meeting the enormous worldwide demand for clothing items. The profit margin of a business rises in accordance with the increase in productivity. some research has been conducted by professionals and academics continuously to determine what factors may impact employee productivity or lead to increased productivity. On the basis of the data at hand, this study's primary objective is to identify performance-enhancing elements and anticipate how much efficiency will occur.

Our dataset includes essential attributes of the garment manufacturing process and the employees’ productivity which had been collected manually and validated by industry experts and published in the UC Irvine Machine Learning Repository. We are curious about the following questions:

* which variables are significant and directly have a positive effect?
* How can work efficiency be predicted in a corporation given specific factors?
* Has time affected the efficiency or productivity of employees?

This research aims to discover the answers to the abovementioned questions, and we will apply the classification problem, which determines whether actual productivity reached the target productivity or not. The research will use supervised learning (Regression analysis, SVM, decision trees, k-Nearest Neighbour, Random Forest, …) with Python programming and essential libraries.

# Introduction

In many industries, assessing, monitoring, and forecasting employee productivity is crucial since businesses depend on their employees' output and performance. Additionally, several elements, such as incentives given, the industry in which they operate, working hours, day as people frequently think it has a significant impact, the team they work in, and many other features, play an essential role in affecting the productivity of employees. Companies must evaluate and take care of these aspects since they require their employees to be productive.

A common problem in the industry in this regard is that the actual productivity of garment employees sometimes does not meet the targeted productivity set by the authorities to meet production objectives in time, resulting in enormous losses. Before increasing the productivity of employee performance, it is necessary to know in advance what factors affect and how to predict employee productivity, especially garment employees that are being discussed [1].

According to much research, many key factors affect employees’ productivity. Some of these factors include employee training, employee empowerment, and teamwork skills [2]. Other vital factors have been found in research that studied a Bangladesh factory. It has been summarized into nine key elements: working hours, wages and benefits, holidays, discrimination, harassment and abuse, workplace conditions, forced labour, welfare and employment relations [3].

This study aims to solve the problem mentioned by predicting the actual productivity of the employees. To achieve this aim, we applied Machine Learning Algorithms to the dataset and compared the results to determine the appropriate algorithm with the best accuracy to predict the actual productivity of the employees.

# Literature review

Machine learning algorithms such as decision tree, Naïve Bayes, Random Forest and SVM were applied by study Bhatia, Arora, and Tomar 2016 [4] for the presence of diabetic retinopathy, and the results proved that the model could help in detecting symptoms earlier. Outperformed results were found in a study conducted by Kruppa et al. 2013 [5] for credit risk prediction using the framework of machine learning algorithms such as random forests (RF), k-nearest neighbours (KNN) and bagged K nearest neighbours (BKN). Furthermore, a study by Balla, Rahayu, and Purnama 2021 [6] proved a promising result in predicting employee productivity which is one of the most substantial factors in any organization. The study applied three classification algorithms, namely, Neural Network (NN), Random Forest (RF) and Regression Linear (RL). Random forest showed minimal values of the correlation coefficient, MAE, and RMSE, which reflect that RF is very appropriate in predicting employee productivity.

Decision tree classification algorithms utilized by Attygalle and Abhayawardana 2021 [7] for investigating and visualizing employee productivity and any other social phenomenon with evidence.

On the other hand, a prediction model has been built by study Sorostinean, Gellert, and Pirvu 2021 [8] using decision tree methods and data mining tools to investigate the effect of decision tree methods and ensemble learning for improving performance prediction in assembly assistance system. The results demonstrated that the gradient-boosted decision trees were the best through all the decision-tree-based methods. Some studies evaluated workers’ performance in textile companies by using ML and ensemble learning algorithms, such as the study Saad 2020. [9] which applied different Machine learning algorithms, including decision trees and bagging algorithms ,to achieve the highest accuracy. The CHAID model produced high-level specificity and sensitivity.

Four different ML algorithms, including support vector machine, optimized support vector machine (using a genetic algorithm), random forest, XGBoost and Deep Learning, were used by El Hassani, El Mazgualdi, and Masrour 2019 [10] for predicting the overall equipment effectiveness (OEE) which is a performance measurement of the manufacturing industry. Deep learning and random forest with cross-validation manifest the best results for predicting OEE.

Additionally, an approach built-in study De Lucia, Pazienza, and Bartlett 2020 [11] of ML and logistic regression used for financial performance prediction by focusing on predicting the accuracy of main financial indicators such as Return of Equity (ROE) and Return of Assets (ROA).

Ruba Obiedat and Sara Toubasi released a study that focused on predicting garment employee productivity using different machine learning algorithms such as J48, RF, SVM, NB, and RBF with and without ensemble learning algorithms, including bagging and Adaboost. Their proposed approach succeeds in enhancing almost all classifiers’ performance. J48 was superior compared with all other applied algorithms. The best results were obtained by J48 combined with Adaboost on 20th iterations with 0.9916 accuracies, 0.0083 MAE and 0.0908 RSME. Consequently, J48 with the Adaboost algorithm was found to be the best for garment employee productivity prediction. [12]

# Goal

My goal is to build models to classify and predict the productivity of garment employees for my dataset. Meanwhile, I would like to compare performance, Regression Models (Linear, Lasso, Ridge, Poisson, decision tree, KNN regression, SVR, Random Forest) and boosting algorithms (xgboost, Gradient boosting).

# Machin Learning metric

#### Evaluation Metrics

To evaluate our model performance and measure the errors, we have considered three evaluation metrics named Mean Squared Error, Mean Absolute Error, and Mean Absolute Percentage Error. All the metrics have been chosen by considering their benefits and interpretation. Details about the metrics are discussed below.

#### Mean Squared Error (MSE)

MSE basically measures the average squared error of our predictions. For each point, it calculates the square difference between the predictions and the target and then averages those values. The higher this value, the worse the model is. MSE can be formulated as:

1N∑i=1N(yi−y^i)2(8\*)

##### **Mean Absolute Error (MAE)**

MAE is calculated as an average of absolute differences between the target values and the predictions. The MAE is a linear score which means that all the individual differences are weighted equally in the average. What is important about this metric is that it penalizes huge errors that are not as that badly as MSE does. Thus, it is not as sensitive to outliers as mean square error. MAE can be formulated as:

1N∑i=1N|yi−y^i|(9\*)

##### **Mean Absolute Percentage Error (MAPE)**

MAPE is basically expressed as the relative error preference. For each instance, the absolute error is divided by the target value, giving a relative error. The primary advantage of using MAPE is the clear interpretability of its results. MAPE provides a single value of percentage for the error. Therefore, when the average range of the prediction is known, it can be simply estimated what the predictions are going to look like. MAPE can be formulated as:

100%N∑i=1N|yi−y^i|yi(10\*)

**Accuracy and Precision**

We use the following formulae to evaluate, accuracy and precision. The Mathews correlation coefficient (MCC) is a machine learning measure which is used to check the balance of the binary (two class) classifiers. It considers all the true and false values, which is why it is generally regarded as a balanced measure that can be used even if there are different classes.

****

# Dataset

In this study, the dataset used is garments worker productivity, which is a public dataset because it is taken from the UCI repository website. The dataset used in this study was published in 2020 with 15 attributes and has 1197 instances. All attributes’ details are shown in the below tables.

**A brief descriptive statistics of the selected dataset**

Each attribute for the training dataset is described in the following two tables. The first table, which includes a description of each attribute, was taken from the original UCI Repository. The second table, which shows which characteristics are numerical and whether they are continuous or discrete, and which features are categorical and whether they are nominal or ordinal, is obtained by looking at each attribute individually. It also offers some preliminary findings regarding the ranges and typical values of the attributes.

**Table 1: Attribute description**

| **ID** | **Attribute** | **Description** |
| --- | --- | --- |
| **1** | **Date** | Date in MM-DD-YYYY |
| **2** | **Quarter** | A portion of the month. A month was divided into four quarters |
| **3** | **Department** | Associated department with the instance |
| **4** | **Day** | Day of the Week |
| **5** | **Team** | Associated team number with the instance |
| **6** | **Targeted\_productivity** | Targeted productivity set by the Authority for each team for each day. |
| **7** | **Smv** | The Standard Minute Value, is the allocated time for a task |
| **8** | **Wip** | Work in progress. Includes the number of unfinished items for products |
| **9** | **Over\_time** | Represents the amount of overtime by each team in minutes |
| **10** | **Incentive** | Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action |
| **11** | **Idle\_time** | The amount of time when the production was interrupted due to several reasons |
| **12** | **Idle\_men** | The number of workers who were idle due to production interruption |
| **13** | **No\_of\_style\_change** | Number of changes in the style of a particular product |
| **14** | **No\_of\_workers** | Number of workers in each team |
| **15** | **Actual\_productivity** | The actual % of productivity that was delivered by the workers. It ranges from 0-1 |

**Table 2: Attribute type**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Attribute | Type | Sub-type | Comments |
| 1 | date | Categorical | Ordinal | Contains duplicates. Range from 2015-01-01 to 2015-03-11, which is 59 days. Can drop |
| 2 | quarter | Categorical | Ordinal | Has 5 values. Quarter 1-5 |
| 3 | department | Categorical | Nominal | Has 2 values. ‘sewing’ and ‘finishing’ |
| 4 | day | Categorical | Ordinal | Has 6 values. Friday does not appear. Day off? |
| 5 | team | Categorical | Nominal | Has 12 values, for 12 teams |
| 6 | targeted\_productivity | Numerical | Continuous | Has 9 unique values - range is 0.07 - 0.8. Very likely to have outliers |
| 7 | smv | Numerical | Continuous | Values range from 2.9 to 54.56 |
| 8 | wip | Numerical | Discrete | Values range from 7 to 23122. Mistake? Extreme outliers |
| 9 | over\_time | Numerical | Discrete | Values range from 0 to 25920. Extreme outliers. The majority are 0, 960, 1440 |
| 10 | incentive | Numerical | Continuous | Values range from 0 to 3600. Extreme outliers. The majority is 0 |
| 11 | idle\_time | Numerical | Continuous | Values range from 0 to 300. Extreme outliers. The majority is 0 |
| 12 | idle\_men | Numerical | Discrete | Values range from 0 to 45. The majority is 0 |
| 13 | no\_of\_style\_change | Numerical | Discrete | Only has 3 values - 0,1,2. The majority is 0 |
| 14 | no\_of\_workers | Numerical | Discrete | Ranges from 2 to 89. The majority is 8. There are some numbers that include decimals i.e. 51.5. Mistake? |
| 15 | actual\_productivity | Numerical | Continuous | Values range from 0.2337 to 1.1204 |

**Graph methodology**

Data Garment Worker Productivity

Abstract

Literature review

Set Parameters

Cross-Validation

Best Parameters

Retrain Model

Test data

Training data

Prepare Data | Normalize | Attribute selection

Figure 1. Research Method

Final Represent

Final Result

**Steps to implement**

* **Data Processing**

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labelling raw data into a form suitable for machine learning algorithms and then exploring and visualizing the data.

In this research, I found missing data in the “Wip” feature. It has a 42 percent null value, and replaced the missing data with the use of interpolate function also; I Separated categorical and numerical data for simplicity in analysis. Department Attribute had space present in one of the values, which needs modification. Few features had numeric values but, in fact, were categorical variables, and I Performed One Hot encoding. Scaling of the data makes it easy for a model to learn and understand the problem. Scaling applied for 'smv', 'wip,’ 'over\_time,’ 'incentive', 'no\_of\_workers', 'idle\_time', 'idle\_men' features.

* **Feature Selection**

Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached.

Creating an estimator, Creating an RFE object and fitting the training data into our model are all steps that I applied for feature selection, then I prepared data for data modelling.

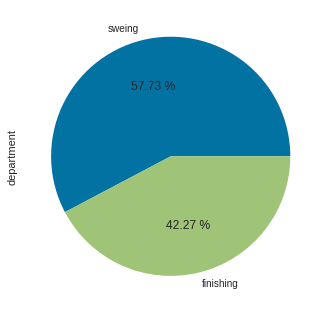
* **Model Development**

Nine models were trained: Linear Regression, Lasso Regression, Ridge Regression Decision Trees & Random Forest, Support Vector Regression, KNeighbour Regression, XGBoost, Gradient Boost. We developed these models using Python’s libraries. By scatter plot, we can see Actual Productivity with Predicted Productivity in each model.

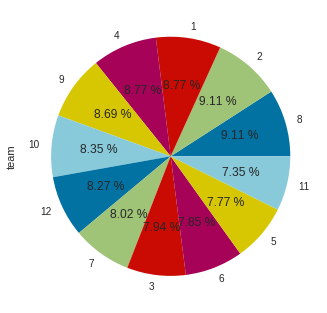
In this research, we obtained some charts with different types for the best understanding of data and feature impact

* **pie charts for some categorical feature**

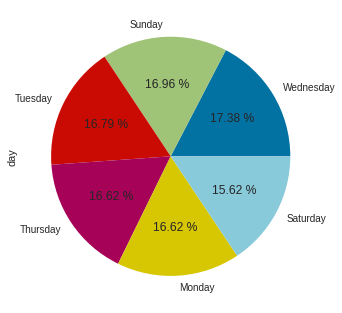
1. department



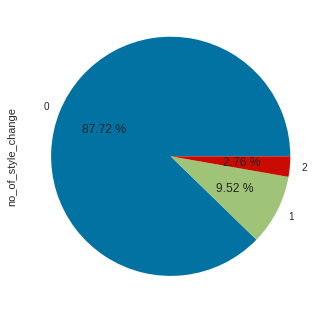
1. team



1. day

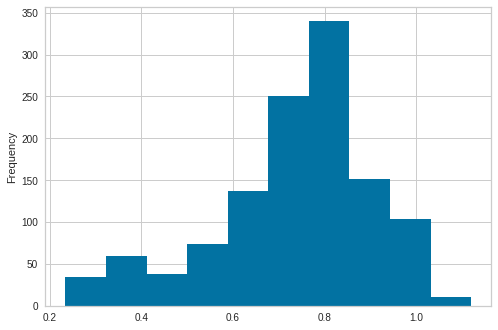


4-no-of-style-change

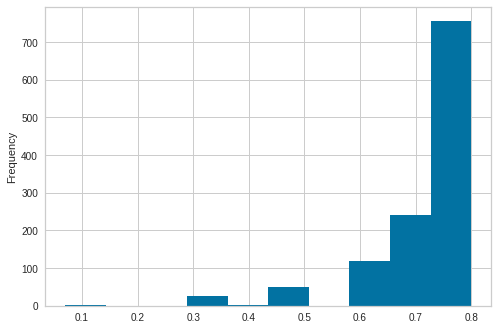


* **Bar charts for some numerical feature**

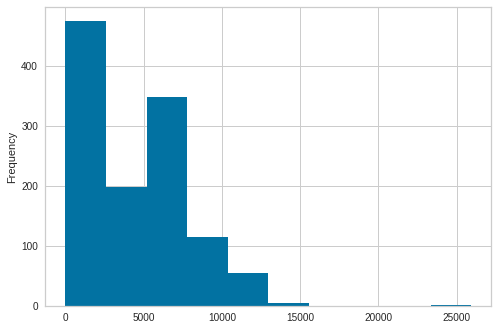
1. **Actual productivity**

****

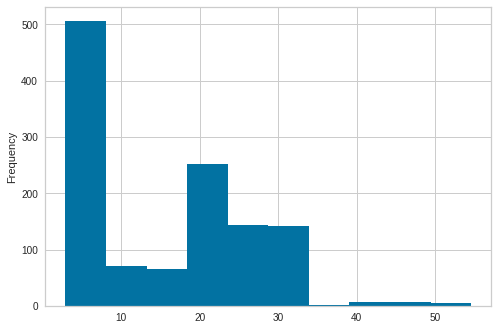
1. **Target productivity**

****

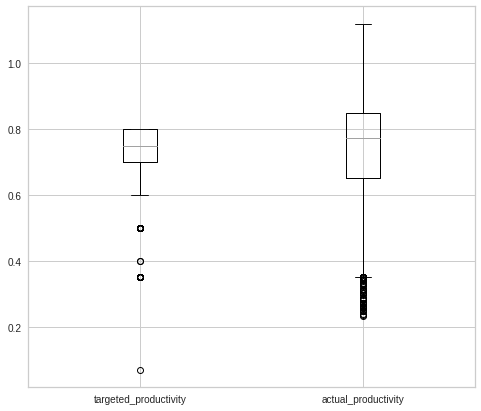
1. **Overt time**

****

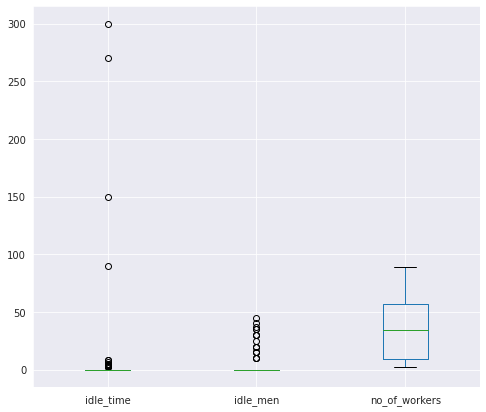
1. **Smv**

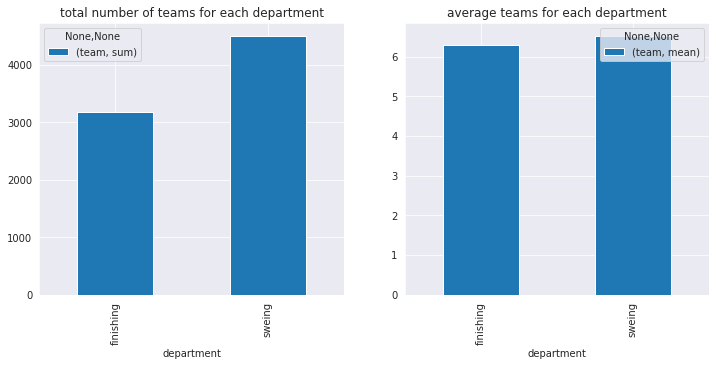
****

* **Boxplot for some numerical feature**

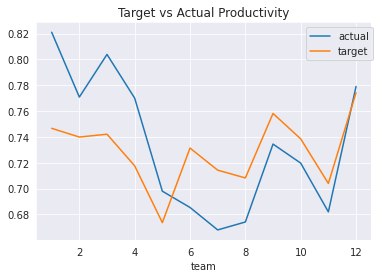


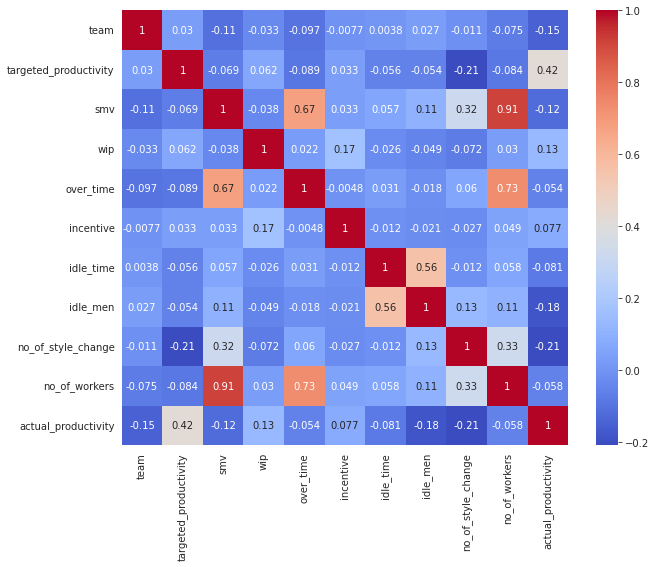
One can see that the targeted productivity is capped at 0.8 while actual productivity is over 1

****

****

In the above plot, one can see the number of teams for each department

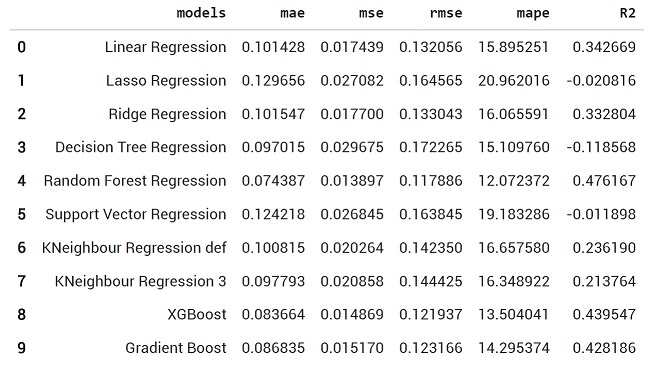


****From the above heatmap, we can see that there is a very high correlation between smv and no\_of\_workers.

over\_time & smv as well as over\_time and no\_of\_workers are correlated

**Result**

I could see that there is no significant difference between ensemble methods and traditional methods for prediction productivity in my data set. All the methods had good metrics for classification and prediction. Although random forest, Xboost and Gradient Boost had the highest accuracy. Based on the result, it appears that targerted\_productivity can have a strong positive impact on actual\_productivity, some sort of a self-fulfilling prophecy effect. Quarter is another factor that can influence actual productivity, with Quarters 5, 1, and 2 being the most productive, respectively. Team can also affect actual productivity, with teams 1, 12, and 3 being the most productive, respectively.no\_of\_workers can also play a role in influencing actual\_productivity, even though its effect is more nuanced and not necessarily linear, as demonstrated through EDA and model development. Incentive is also another important factor that can be a meaningful predictor of actual\_productivity. I understand that date and time features have the least impact on productivity.

****

**Reference**

[1] Imran2019 Deep NN, Deep Neural Network Approach for Predicting the Productivity of Garment Employees, author={Abdullah Al Imran and Md Nur Amin and Md. Rifatul Islam Rifat and Shamprikta Mehreen},

[2] Hanaysha, J. (2016) Examining the Effects of Employee Empowerment, Teamwork, and Employee Training on Organizational Commitment. Procedia-Social and Behavioral Sciences, 229, 298-306.

[3]. Alam, Mohammad & Alias, Rosima & Azim, Mohammad. (2018). Social Compliance Factors (SCF) Affecting Employee Productivity (EP): An Empirical Study on RMG Industry in Bangladesh. 10. 87-96.

[4] Alam, Mohammad, Rosima Alias, and Mohammad Azim. 2018. 'Social Compliance Factors (SCF) Affecting Employee Productivity (EP): An Empirical Study on RMG Industry in Bangladesh', 10: 87-96. https://www.researchgate.net/publication/326733299

[5] Bhatia, Karan, Shikhar Arora, and Ravi Tomar. 2016. "Diagnosis of diabetic retinopathy using machine learning classification algorithm." In 2016 2nd international conference on next generation computing technologies (NGCT), 347-51. IEEE DOI: 10.1109/NGCT.2016.7877439.

[6 ] Kruppa, Jochen, Alexandra Schwarz, Gerhard Arminger, and Andreas Ziegler. 2013. 'Consumer credit risk: Individual probability estimates using machine learning', Expert systems with applications, 40:5125-31 DOI:https://doi.org/10.1016/j.eswa.2013.03.019. [https://www.sciencedirect.com/science/article/pii/S0 957417413001693](https://www.sciencedirect.com/science/article/pii/S0%20957417413001693)

[7] Balla, Imanuel, Sri Rahayu, and Jajang Jaya Purnama. 2021. 'GARMENT EMPLOYEE PRODUCTIVITY PREDICTION USING RANDOM FOREST', Jurnal Techno Nusa Mandiri, 18: 49-54 DOI: <https://doi.org/10.33480/techno.v18i1.2210>

[8] Attygalle, Dilhari, and Geethanadee Abhayawardana. 2021. 'Employee Productivity Modelling on a Work from Home Scenario During the Covid-19 Pandemic: A Case Study Using Classification Trees', Journal of Business and Management Sciences, 9: 92-100 DOI: 10.12691/jbms-9-3-1

[9] Ďurica, Marek, Jaroslav Frnda, and Lucia Svabova. 2019. 'Decision tree based model of business failure prediction for Polish companies', Oeconomia Copernicana, 10: 453-69 DOI: 10.24136/oc.2019.022

[10] Mahoto, Naeem, Rabia Iftikhar, Asadullah Shaikh, Yousef Asiri, Abdullah Alghamdi, and Khairan Rajab. 2021. 'An Intelligent Business Model for Product Price Prediction Using Machine Learning Approach,’ 30: 147-59 DOI: 10.32604/iasc.2021.018944

[11] Sorostinean, Radu, Arpad Gellert, and BogdanConstantin Pirvu. 2021. 'Assembly Assistance System with Decision Trees and Ensemble Learning,’ Sensors, 21: 3580 DOI: <https://doi.org/10.3390/s21113580>

[12] Saad, Hamza. 2020. 'Use Bagging Algorithm to Improve Prediction Accuracy for Evaluation of Worker Performances at a Production Company, arXiv preprint arXiv:2011.12343 DOI: 10.4172/2169- 0316.1000257

**Appendix**

A link to a repository on GitHub:

<https://github.com/Maryam-Dehkordi>

The link below refers to the source of this data:

<https://archive.ics.uci.edu/ml/machine-learning-databases/00597/garments_worker_productivity.csv>