

|  |
| --- |
| industry production  Data Science Report Graphical user interface, application  Description automatically generated Presented by:   * Mohamed ibrahim Saad ID: 42010235 * mustafa mohamed kadry ID: 42010410 * tareq fathi ID: 42010245 * kamel mustafa kamel ID: 42010191 * omar mahmoud ID: 41910312 * ali abdelaty mahmoud ID:41910151 |

Introduction

1. Data Science

Data science is the field of data analytics and data visualization in which raw data or the unstructured data is cleaned and made ready for the analysis purpose. Data scientists use this data to get the required information for the future purpose. Data science uses many processes and methods on the big data, the data may be structured or unstructured. Data frames available on the internet is the raw data we get. It may be either in unstructured or semi structured format. This data is further filtered, cleaned and then number of required task are performed for the analysis with the use of the high programming language. This data is further analyzed and then presented for our better understanding and evaluation. One must be clear that data science is not about making complicated models or making awesome visualization neither it is about writing code but about using the data to create an impact for your company, for this impact we need tools like complicated data models and data visualization.

1. Stages of Data Science

There are many tools used to handle the big data available to us. So, we used programming tools such as Python to extract knowledge from prepared data. Data scientists use many algorithms and mathematical models on the data.

1. Use of Data Science in industry production area of a factory

This data is the production area of a factory, information about the teams/classes working in the factory and their status. Here is an overview of some of the variables found in this data:

date (date): The date on which the action was made.

Quarter (Quarter): To fly to the quadrant to which the date belongs.

Today (Today): Reference to the day the action was done.

Team (Team): the working factor in the worker.

product target product (target\_product): main product main product / target product main product / main product product

SMV: Example to Standard Constants of Work (Standard Minute Value), which is a measure of the time a given operation takes.

Work in progress in the factory: work in progress in the factory (work in progress), products or processes that have not yet been completed.

Overtime (overtime): Image to the number of hours of work done.

Motivation (Incentive): An indication of the reward or incentives that were offered / the first team.

idle time (idle\_time): Tells when the new color was created.

Workers, Crude Oil Workers (idle\_men) ...

Styles style count (no\_of\_style\_change): Change to the style count of production styles.

Number of workers (no\_of\_workers): The workers to the workers involved in the work.

Producer Main producer (actual\_production): Worker to main producer/chapter.

final section (department\_finishing): A section to the section where the loop was completed.

Sewing department (department\_sweing): Sew to the section where the sewing was done.

Problem definition :

Handling missing values: Missing values can be handled in different ways depending on the nature and extent of the missingness. One option is to impute the missing values with appropriate values such as mean, median, or mode of the column. Another option is to remove rows with missing values if the percentage of missing values is high. However, it is important to ensure that the missingness is handled appropriately to avoid biasing the results.

Handling duplicates: Duplicate rows can be identified by checking if all the values in two or more rows are the same. If there are duplicate rows, they could be removed or merged based on the specific task at hand. For example, if we are interested in analyzing the performance of each team, we may want to remove the duplicate rows to avoid counting the same team's performance multiple times.

Data type conversion: Data type conversion is important to ensure that the data is in the appropriate format for analysis. For example, the date column could be converted to a datetime data type so that we can extract information such as day, month, or year from the date. Similarly, the numerical columns could be converted to float or integer data types depending on the precision needed.

Feature engineering: Feature engineering is the process of creating new features from the existing ones to improve the performance of machine learning models or to gain insights about the data. For example, we could create a new feature by calculating the ratio of actual productivity to targeted productivity to measure how close each team/section is to meeting their productivity targets. Another example could be creating a new feature that represents the total idle time by adding the idle time and idle men columns.

Outlier detection and removal: Outliers are data points that lie far from the typical range of values for a given feature. Outliers can be detected using various statistical methods such as z-score or IQR (interquartile range) and can be handled by removing them or replacing them with appropriate values.

Scaling and normalization: Scaling and normalization are used to ensure that the features are on the same scale and to improve the performance of machine learning algorithms. For example, min-max scaling can be used to scale features to a range between 0 and 1, while standardization can be used to scale features to have a mean of 0 and a standard deviation of 1.

Encoding categorical variables: Categorical variables are variables that take on a finite set of values such as team names or department names. These variables need to be encoded before they can be used in machine learning algorithms. One-hot encoding is a popular method for encoding categorical variables, where each value in the categorical variable is converted to a binary feature.

Libraries we used :

* from keras.utils import to\_categorical
* from sklearn.preprocessing import LabelEncoder
* from sklearn.preprocessing import OneHotEncoder
* import pandas as pd
* from sklearn import preprocessing
* from sklearn.model\_selection import train\_test\_split
* from sklearn.ensemble import RandomForestRegressor
* from sklearn.metrics import mean\_absolute\_error
* import numpy as np
* from sklearn.linear\_model import Logistic Regression
* from sklearn.metrics import mean\_squared\_error
* from sklearn.metrics import mean\_absolute\_error
* from sklearn.metrics import mean\_squared\_log\_error
* from sklearn.metrics import median\_absolute\_error
* from sklearn.metrics import max\_error
* import tensorflow as tf
* from tensorflow import keras
* import tensorflow.python.keras
* from tensorflow.python.keras.models import Sequential
* from tensorflow.python.keras.layers import Dense
* from sklearn.preprocessing import LabelEncoder
* from sklearn.preprocessing import OneHotEncoder
* from sklearn.metrics import accuracy\_score

question :

1-How many records are there in the dataset?

2-What are the different departments in the dataset?

3-How many unique workers are there in each department?

4-What is the average productivity for each department?

5-What is the distribution of idle time across all workers?

6-How does productivity vary over different quarters?

7-Is there any correlation between incentive and productivity?

8-Which day of the week has the highest average productivity?

9-What is the distribution of overtime hours across all workers?

10-How does productivity vary with respect to the number of idle men?

11-What is the relationship between target productivity and actual productivity?

12-What is the distribution of work-in-progress (WIP) across all workers?

13-How does the productivity vary for each worker?

14-What is the distribution of incentive amounts across all workers?

15-How does productivity vary across different months?

16-How does productivity change over time?

17-What is the average idle time for each department?

18-How does the distribution of productivity vary for each quarter?

19-What is the correlation between different variables in the dataset?

20-What is the average productivity for workers in each department on different days?

Reference :

https://www.kaggle.com/datasets

<https://poe.com/chatgpt>

<https://youtube.com/@HeshamAsem>