**Deliverable – 1**

**ISM6136.004F23 – Data Mining**

**Team “We Tried”**

Bapuji Satyala (U61600601)

Ajay Srikar Medidi(U85645387)

Rehaan Ahmed Abdul (U84763124)

Ramyasri Muthineni (U41740753)

Sanjana Anumula (U64639710)

Gayatri Sowbhagya Lakshmi Akkireddi (U37764178)

**Business Question**

One of the most notable instances of this modern era's industrial globalization is the apparel industry. There are many manual operations in this business, which require a lot of labor. Employee performance in terms of production and delivery in clothing manufacturing enterprises is a major factor in meeting the huge demand for textile products around the world. Hence, monitoring, evaluating, and forecasting the productivity performance of the factory work teams is highly desired by the decision-makers in the apparel business.

We would like to know what all factors in would influence the productivity of the workers where we consider all factors that would influence and how the stake holders can use this data to evaluate the productivity trend and what all factors, they need to look at for the productivity of their workers to rise to help the company’s profits.



**Dataset Source**

The dataset we selected was obtained from the UC Irvine repository and provides information on the productivity of employees in the garment manufacturing process.

The link to the dataset is given below-[Link](https://archive.ics.uci.edu/dataset/597/productivity+prediction+of+garment+employees)

**About the Dataset**

The Data set contains different teams of garment workers, including their work hours, incentives, and productivity goals, and it provides insight into factors affecting their productivity, such as style changes and interruptions in production. It is roughly containing 1200 instances with 14 attributes and 1 target variable.

From this we need to Identify what all attributes influence the productivity of the Garment employees and what we can do to improve the productivity of the workers which will simultaneously improve the company’s profit values.

Firstly, we need to Identify what all influence the Productivity of the dataset. The dataset contains.

|  |  |
| --- | --- |
| Dataset Attributes | Description |
| date | Date in date format |
| Day | Day of the week |
| Quarter | Part of the month where month was divided into 4 parts |
| Department | What team is the garment worker part of |
| Team number | Team number of the garment worker |
| Number of Workers | Total number of the workers in that day |
| Number of Style Changes | Number of changes in the style of a particular product |
| Targeted Productivity of the workers | The calculated productivity of the company which will result in company profits. |
| Standard Minute Value | It is the specific time allocated to each task |
| Work in Progress | Number of Unfinished items for products. |
| Over time | It the total time exceeded by the team in getting the work completed. |
| Incentive | It is the total financial input from the company to improve a particular course of action |
| Idle time | Time that got wasted due to several reasons when work was halted |
| Actual Productivity | It is the productivity of the worker that has been achieved from the company calculations. |

**Real-world Impact of solving this problem**

With predictive analysis there are many garment companies out in the real world that can similarly use this kind off model to enhance their company strategies and investments in their respective departments that will improve their worker’s productivity which in turn will yield more profits to the company.



**Analysis**

We plan to analyze the dataset, which contains detailed information about garment employee’s productivity, including work hours, skills, and machine data. Our objective is to predict productivity based on these features.

This is a classification problem. So, we’ll start with Data Exploration, Preprocessing, Model fitting and Visualization. We plan to work on this dataset using the Classification Algorithms which we have learnt in class. We commonly use logistic regression, decision trees, and support vector machines as classification techniques. We’re planning to implement all the mentioned models and find which model best suits to this dataset and also how this will help the decision makers in their business.

We want to classify the productivity of work teams into two distinct groups, such as “Productive” and "Not Productive" based on the difference between actual and target productivity numbers. We’re planning to use suitable metrics and find best model based on it.

**Net Productivity = Targeted Productivity – Actual productivity**

**Net productivity <=0** is said to be **productive** and

**Net productivity >0** is said to be **not Productive.**

**Productive = 1**

**Not Productive = 0**

We have stored these values under ‘**Classification**’ column and made this as our **Target** Variable.

**Model Planning and Techniques Used from Class**

Since it is a classification dataset, we would like to maximize all the classification techniques taught in class - logistic regression, decision tree, and support vector machine - and apply Hyperparameter tuning to the models both Grid search and random search- to find a better fit to the model with better accuracy and help stakeholders with their productivity planning. Process Overview has been listed below.