

Engenharia da Computação

Carlos Henrique Gomes dos Santos/1760092 /ENG 171i

Tema: Previsão de produtividade dos funcionários de vestuário

Birigui – SP

2021

Carlos Henrique Gomes dos Santos

Previsão de produtividade dos funcionários de vestuário

Relatório apresentado como requisito parcial para avaliação final – disciplina Mineração de Dados, do curso superior em Engenharia da Computação, ministrado pelo Prof Dr. Murilo Varges da Silva

Birigui – SP

2021

1. **Resumo**

Foi escolhido no site *UCI – Machine Learning Repository* a base de dados Previsão de produtividade dos funcionários de vestuário, onde a mesma deveria ter alguns dados ausentes, utilizando o algoritmo em linguagem *Python* fornecido pelo Prof. Dr Murilo para o processamento da base de dado escolhida, gerando dados estatísticos e filtrando dados da base que podem ser utilizadas para gerar informações relevantes para o processo.

Para isso então foi feito o download da base de dados e verificado quais atributos são relevantes, após isso é realizado o estudo do algoritmo fornecido, e a extração das informações com base nos dados fornecidos.

1. **Informações sobre a base de dados**

A Indústria de Vestuário é um dos principais exemplos da globalização industrial desta era moderna. É uma indústria altamente intensiva em mão-de-obra com muitos processos manuais. Satisfazer a enorme demanda global por produtos de vestuário depende principalmente do desempenho de produção e entrega dos funcionários nas empresas de fabricação de vestuário. Por isso, é altamente desejável entre os tomadores de decisão da indústria de vestuário acompanhar, analisar e prever o desempenho da produtividade das equipes de trabalho em suas fábricas.

1. **Informações de atributos**
2. date : Data em MM-DD-YYYY
3. day: Dia da Semana
4. quarter: Uma parte do mês. Um mês foi dividido em quatro partes
5. departament : Departamento associado com a intância
6. team\_no : Número de equipe associado com a intância
7. no\_of\_workers : Número de trabalhadores em cada equipe
8. no\_of\_style\_change : Número de mudanças no estilo de um determinado produto
9. targeted\_productivity : Produtividade direcionada definida pela Autoridade para cada equipe para cada dia.
10. smv : Valor do minuto padrão, é o tempo alocado para uma tarefa
11. wip : Trabalho em andamento. Inclui o número de itens inacabados para produtos
12. over\_time : Representa a quantidade de horas extras por cada equipe em minutos
13. incentive : Representa a quantidade de incentivo financeiro (no BDT) que permite ou motiva um determinado curso de ação.
14. idle\_time : A quantidade de tempo em que a produção foi interrompida por diversos motivos
15. idle\_men : O número de trabalhadores que estavam ociosos devido à interrupção da produção
16. actual\_productivity : O real % de produtividade que foi entregue pelos trabalhadores. Varia de 0-1.
17. Productivity: foi criado para se ter dados binários, actual\_productivity maior que 5, o campo recebe 1, se for menor recebe 0
18. **Pré-processamento**
    1. *Data Cleaning*

A primeira etapa do algoritmo é mostrar as 15 primeiras linhas da base de dados, logo após é mostrado informações gerais referentes a base nome de coluna, tipo de dado (*objetc, float, int*), após é mostrado a descrição de dados, quantidade de dados, é apresentado também a quantidade de dados faltantes em cada uma das colunas (figura 1).

Texto

Descrição gerada automaticamente

Figura 1 - Campos da base de dados

Conseguimos por meio deste algoritmo outras informações, tais como,valor mínimo e máximo, desvio padrão, e alguns dados referente a média relacionadas a parte dos dados ¼ ½ e ¾ (figura 2).

Texto

Descrição gerada automaticamente com confiança média

Figura 2 - Descrição de dados

Podemos também filtrar dados de colunas especificas, gerando informações importantes, que chamamos de *target*, no nosso caso será utilizada a *produtivy,*

Quando se trata da substituição de dados ausentes para completar a base de dados, temos diversas opções, substitui por um número específico de escolha, pela mediana, média ou a moda, foi utilizado em nosso problema a média.

* 1. *Data Descriptive*

Para esta etapa foi escolhida alguns campos numéricos para obter valores que descreva informações da nossa base de dados, usando média, moda, quartil, mediana, amplitude, variância, desvio padrão, desvio absoluto, covariância e correlação.

Figura 3: Média

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura 4: Mediana

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura 5: Amplitude

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura 6: Moda

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura 7: Desvio Padrão

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura 8: Variância

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura 9: Desvio Absoluto

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura 10: Covariância e Correlação

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Fonte: Elaboração Própria

* 1. *Data Visualization*

O objetivo desta etapa é abstrair conhecimento da base de dados e apresentar de forma gráfica, conseguindo desta forma extrair o maior número de informação dos dados fornecidos.

Figura 11: Número de pessoas no time durante a semana

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Fonte: Elaboração Própria

A figura 3 apresenta a produtividade de cada setor ao decorrer dos dias das semanas, para elaboração dessa gráfico foi pegada a variável actual\_productivity, que se trata da produtividade e separada nos setores de costura e acababento, após isso foi correlacionado com a variável day, que representa os dias da semanas, com isso temos o gráfico abaixo onde podemos ver qual setor possui um maior valor de produtividade ao decorrer da semana, levando em consideração todos valores da base de dados, conseguimos perceber que o setor de costura apresenta uma leve superioridade em comparação com os de acabamento, levando em conta que o valor varia de zero a um como explicado anteriormente.

Figura 12: produtividade dias da semana

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Fonte: Elaboração Própria

A figura 4 mostra em um gráfico pizza a produtividade de cada setor de uma forma geral, dessa forma é possível perceber que ambos setores apresenta valor de produtividade bem próximos.

Figura 13 Produtividade Departamentos

Gráfico, Gráfico de pizza

Descrição gerada automaticamente

Fonte: Elaboração Própria

Podemos também analisar qual dia da semana a empresa possui um maior valor de produtividade, com base nos dados que foi gerada e posteriormente saber o que ocasiona está situação, como demonstrado na figura 5. Onde na quinta-feira é apresentado o pior índice de produtividade e o sábado seu maior índice de produtividade.

Figura 14: Produtividade geral nos dias da semana

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura 15: Produtividade ao decorrer dos dias

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Fonte: Elaboração Própria

4.4. *Data Reduction*

1. **Clusterização**
2. **Regressão**
   1. Linear

Tela preta com letras brancas

Descrição gerada automaticamente

* 1. SVM

Texto branco sobre fundo preto

Descrição gerada automaticamente

* 1. Bayesian

Texto branco sobre fundo preto

Descrição gerada automaticamente

* 1. Nearest Neighbors

Tela preta com letras brancas

Descrição gerada automaticamente

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html

* 1. Neural Network

Texto branco sobre fundo preto

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente com confiança média

<https://scikit-learn.org/stable/supervised_learning.html>

R^2 negativo = https://qastack.com.br/stats/12900/when-is-r-squared-negative