

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**

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).

Figura 1 - Campos da base de dados

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

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).

Figura 2 - Descrição de dados

Texto

Descrição gerada automaticamente com confiança média

Fonte: Elaboração Própria

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

Figura : valores ausentes wip

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura :Substituição de dados ausentes

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Fonte: Elaboração Própria

1. **Redução e normalização**

Normalização com Z-Score, obtendo os resultados da figura a seguir.

Figura : Normalização Z-Score

Uma imagem contendo Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Projeção PCA utilizando Z-Score para normalização de dados

Figura : PCA com normalização Z-Score

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura : Dados PCA redução Z-Score

Uma imagem contendo Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Normalização utilizando o Min-Max, mostrado a imagem a seguir

Figura : Normalização Min-Max

Texto

Descrição gerada automaticamente com confiança média

Fonte: Elaboração Própria

Projeção PCA utilizando mínimos e máximos para a normalização

Figura :PCA com normalização min-máx

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura : Dados PCA com redução min-máx.

Uma imagem contendo Tabela

Descrição gerada automaticamente

Fonte: Elaboração Própria

1. **Visualização**

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 : Produtividade com base na frequência

Gráfico, Histograma

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura : Número de pessoas no time durante a semana

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Fonte: Elaboração Própria

A figura 13 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 : produtividade dias da semana

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Fonte: Elaboração Própria

A figura 14 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 : 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 15. Onde na quinta-feira é apresentado o pior índice de produtividade e o sábado seu maior índice de produtividade.

Figura : Produtividade geral nos dias da semana

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Fonte: Elaboração Própria

Podemos também analisar com a figura 16, comparando o valor da produtividade real com o valor de produtividade que foi atribuído pela autoridade do setor.

Figura : Comparação da produtividade real e a atribuída pela Autoridade

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Fonte: Elaboração Própria

Na figura 17 temos a produtividade com base em todos os dias que foram coletados para a base, conseguimos analisar a produtividade em dias específicos e que dia teve melhor produtividade, conseguimos perceber também que a quantidade de informações coletadas poderia ser maior

Figura : Produtividade ao decorrer dos dias

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Fonte: Elaboração Própria

Comparando o valor real de produtividade com o valor atribuído pela autoridade do setor.

Figura : Comparação da produtividade real e a atribuída pela Autoridade (decorrer dos dias)

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Fonte: Elaboração Própria

1. **Descrição**

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.

Cálculo de média: , resultados figura 17.

Figura : Média

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

A mediana é valor de centro do conjunto de dados, é necessário ordenar os dados, resultados figura 18.

Figura : Mediana

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Cálculo da amplitude é dados por, como mostrado na figura 19.

Figura : Amplitude

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

A moda é o valor que mais se repete na base de dados, resultados figura 20.

Figura : Moda

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Cálculo do desvio padrão é dado por, , resultado obtidos na figura 21.

Figura : Desvio Padrão

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Cálculo da variância é dado por, , resultado figura 22.

Figura : Variância

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Cálculo do desvio padrão é dado por,

Figura : Desvio Absoluto

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura : Covariância e Correlação

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Fonte: Elaboração Própria

1. **Análise de grupos**

Utilizando dois cluster como mostrado na figura 27, e os dados na figura 28.

Figura : Kmens com 2 clusters

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura : Dados Kmens com 2 clusters

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Utilizando quatro cluster como mostrado na figura 29, e os dados na figura 30.

Figura : Kmens com 4 clusters

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura : Dados Kmens com 4 clusters

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Utilizando oito cluster como mostrado na figura 31 e os dados na figura 32.

Figura : Kmens com 8 clusters

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura : Dados Kmens com 8 clusters

Texto

Descrição gerada automaticamente

Fonte: Elaboração Própria

1. **Regressão**

Regressão é a técnica mais antiga e bem conhecida estatística de que a comunidade de mineração de dados utiliza. Basicamente, a regressão tem um conjunto de dados numéricos e desenvolve uma fórmula matemática que se ajusta aos dados. Quando você está pronto para usar os resultados para prever o comportamento futuro, você simplesmente pega seus novos dados, conecte-o a fórmula desenvolvida e você tem uma previsão! A principal limitação desta técnica é que ela só funciona bem com contínua de dados quantitativos

O R2 determina se o modelo ajusta bem os dados. Quanto mais alto o valor de R2 melhor o modelo ajusta seus dados.

* 1. **Linear**

Regressão linear é o processo de traçar uma reta através dos dados em um diagrama de dispersão. A reta resume esses dados, o que é útil quando fazemos previsões.

Figura : Linear

Tela preta com letras brancas

Descrição gerada automaticamente

Fonte: Elaboração Própria

* 1. **SVM (máquina de vetores de suporte)**

SVM padrão toma como entrada um conjunto de dados e prediz, para cada entrada dada, qual de duas possíveis classes a entrada faz parte, o que faz do SVM um classificador linear binário não probabilístico. Dados um conjunto de exemplos de treinamento, cada um marcado como pertencente a uma de duas categorias, um algoritmo de treinamento do SVM constrói um modelo que atribui novos exemplos a uma categoria ou outra. Um modelo SVM é uma representação de exemplos como pontos no espaço, mapeados de maneira que os exemplos de cada categoria sejam divididos por um espaço claro que seja tão amplo quanto possível.

Figura : SVM

Texto branco sobre fundo preto

Descrição gerada automaticamente

1. Fonte: Elaboração Própria
   1. **Bayesian**

No ponto de vista bayesiano, formulamos regressão linear usando distribuições de probabilidade em vez de estimativas pontuais. A resposta, y, não é estimada como um valor único, mas é assumida como sendo derivada de uma distribuição de probabilidade.

As técnicas de regressão bayesiana podem ser usadas para incluir parâmetros de regularização no procedimento de estimativa: o parâmetro de regularização não é definido em sentido difícil, mas sintonizado com os dados em mãos.

Figura : Bayesian

Texto branco sobre fundo preto

Descrição gerada automaticamente

Fonte: Elaboração Própria

* 1. **Nearest Neighbors**

K-Nearest Neighbor é um método de predição que utiliza da distância entre a amostra atual e seus ​k vizinhos mais próximos no conjunto de treinamento para definir qual será o resultado de sua predição.

A regressão baseada em vizinhos pode ser usada nos casos em que os rótulos de dados são contínuos e não variáveis discretas. O rótulo atribuído a um ponto de consulta é computado com base na média dos rótulos de seus vizinhos mais próximos.

Figura : Nearest Neighbors

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Fonte: Elaboração Própria

* 1. **Neural Network**

O Class MLPRegressor implementa um *perceptron* de várias camadas (MLP) que treina usando a retropagação sem função de ativação na camada de saída, que também pode ser visto como usando a função de identidade como função de ativação. Portanto, ele usa o erro quadrado como função de perda, e a saída é um conjunto de valores contínuos.

Valores abaixo foram obtidos após executar algumas vezes o algoritmo, apresentando variação entre os valores como mostrado na figura abaixo.

Figura : Neural Network

Texto branco sobre fundo preto

Descrição gerada automaticamente

Fonte: Elaboração Própria

Figura : Neural Network

Texto

Descrição gerada automaticamente com confiança média

Fonte: Elaboração Própria

Erro absoluto () compara o ajuste do modelo escolhido com o de uma linha reta horizontal (a hipótese nula). Se o modelo escolhido se encaixar pior que uma linha horizontal, então é negativo. É negativo apenas quando o modelo escolhido não segue a tendência dos dados; portanto, se encaixa pior que uma linha horizontal.

* 1. **Decision Trees**

O nó raiz ou superior da árvore (e há apenas uma raiz) é o nó de decisão que divide o conjunto de dados usando uma variável ou recurso que resulta na melhor métrica de divisão avaliada para cada subconjunto ou classe no conjunto de dados resultante do Dividido. A árvore de decisão aprende dividindo recursivamente o conjunto de dados da raiz em diante (de maneira gananciosa, nó por nó) de acordo com a métrica de divisão em cada nó de decisão. Os nós terminais são alcançados quando a métrica de divisão está em um extremo global.

Figura : Decision Trees

Texto branco sobre fundo preto

Descrição gerada automaticamente

Fonte: Elaboração Própria