ANDRÉ LUIZ DE ANDRADE

FABRÍCIO RANGEL DE SOUSA

GUILHERME HENRIQUE CARVALHO

TATIANA HITOMI MIYAZAKI

VINÍCIUS ROBERTO POLO

**USO DE ALGORITMO DE INTELIGÊNCIA ARTIFICIAL NA APLICAÇÃO “UM TREIM DI CUMÊ”**

Trabalho apresentado à Fatec de Franca, como parte da avaliação da matéria Mineração de Dados do Curso de Desenvolvimento de Software Multiplataforma.

Professora: Dra. Jaqueline Brigladori Pugliesi

Franca

2023

**RELATÓRIO**

**INTRODUÇÃO**

O presente trabalho tem a finalidade de relatar os passos realizados e as conclusões de cada etapa, desde a criação do conjunto de dados, seu pré-processamento, extração de padrões, pós-processamento e finalmente utilização na aplicação “Um Treim di Cumê”, criada pelo grupo de alunos para entrega do projeto integrador do sexto semestre do curso de Desenvolvimento de Software Multiplataformas da FATEC de Franca.

A primeira providência para o presente trabalho é explicar, ainda que resumidamente, do que se trata a aplicação “Um Treim di Cumê”, além de expor qual é a intenção do grupo ao fazer a integração com algoritmos de mineração de dados.

O projeto “Um Treim di Cumê” tem como objetivo a criação de aplicação, para plataformas web e mobile, na qual um empresário, proprietário de restaurante ou lanchonete, poderá fazer a inscrição de seu negócio e respectivo menu, oferecendo seus produtos a clientes que deverão fazer um cadastro na aplicação para aquisição dos diversos artigos oferecidos.

A intenção do grupo, ao integrar ao projeto algoritmo de mineração de dados, é que o sistema faça sugestões de produtos aos clientes, baseado em um produto selecionado.

**CONJUNTO DE DADOS**

A fim de obter um conjunto de dados que atendesse aos requisitos necessários para as pretensões do grupo no projeto integrador, foram feitas pesquisas em sites que disponibilizam datasets, porém não foi localizado nenhum que tivesse as propriedades desejadas para os fins almejados.

Entre os conjuntos pesquisados e testados estão as bases “NYC Restaurants Data – Food Ordering and Delivery” (<https://www.kaggle.com/datasets/ahsan81/food-ordering-and-delivery-app-dataset>), “Food Delivery Time Exploratory Data Analysis” (<https://www.kaggle.com/datasets/ankitkalauni/the-food-delivery-time-for-different-cuisines>) e “Food Preferences” (<https://www.kaggle.com/datasets/vijayashreer/food-preferences>).

As imagens abaixo mostram parte do pré-processamento feito na base “NY Restaurants Data – Food Ordering and Delivery”, que ao final mostrou-se inadequada para o projeto, pois faz classificação e indicação com base em nota que o cliente atribuiu aos restaurantes, não havendo esta possibilidade no projeto “Um Treim di Cumê”, ao menos por enquanto.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente

Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamenteUma imagem contendo Site

Descrição gerada automaticamente

Na próxima imagem é possível ver que o dataset “Food Delivery Time Exploratory Data Analysis” mostrou-se insuficiente logo ao serem exibidas suas colunas e tipos de dados existentes, uma vez que também faz utilização de notas atribuídas aos restaurantes pelos clientes, assim como a base anterior.

Tela de computador

Descrição gerada automaticamente

O conjunto de dados “Food Preferences” igualmente deixou de atender às necessidades do projeto “Um Treim di Cumê” logo no início do pré-processamento dos dados, tendo também sido descartado, uma vez que não guardava correlação com o projeto e as informações trazidas seriam pouco proveitosas. Abaixo imagens de algumas etapas do pré-processamento realizadas.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente

Assim, o grupo decidiu criar o próprio dataset, e, inicialmente tentou-se fazer o preenchimento dos dados de forma automática e totalmente aleatória, utilizando para isso fórmulas do programa Excel, porém o resultado não foi satisfatório e esse conjunto de dados também foi descartado pelo grupo, por não ser possível fazer correlações que atendessem às necessidades do projeto, conforme imagens seguintes.

Tela de computador com jogo

Descrição gerada automaticamente

Gráfico

Descrição gerada automaticamenteTexto

Descrição gerada automaticamente

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Interface gráfica do usuário

Descrição gerada automaticamente com confiança média

Texto

Descrição gerada automaticamente

Finalmente, foi criado um novo dataset pelo grupo, com as seguintes colunas: “id\_pedido”, “day\_of\_week”, “id\_cliente”, “gender”, “age”, “id\_restaurante”, “cousine\_type”, “id\_item”, “food\_name”, “type\_of\_product”, “quantity”, “price” e “total\_value”.

Criado um banco de dados PostgreSQL com as colunas já mencionadas, utilizou-se a plataforma online gratuita Clever-cloud para hospedagem do banco de dados. A partir de então cada integrante do grupo, utilizando a API criada para o sistema da aplicação “Um Treim di Cumê” hospedada na plataforma online gratuita Render.com, passou a fazer inserção de dados de restaurantes, menus, clientes e pedidos, de forma a simular a utilização real do sistema.

Ao final, obteve-se um conjunto de dados com 566 registros de pedidos, que foi integrado ao projeto “Um Treim di Cumê”. As fases da mineração de dados estão relatadas a seguir.

**PRÉ-PROCESSAMENTO, EXTRAÇÃO DE PADRÕES E PÓS-PROCESSAMENTO**

Para o pré-processamento de dados foi utilizada a plataforma do Google Colab, utilizando a linguagem de programação Python.

O primeiro passo foi a importação do conjunto de dados e a exibição dos cinco primeiros registros, conforme imagem a seguir.

Texto

Descrição gerada automaticamente

Ao fazer a exibição dos dados do conjunto de dados, obteve-se a informação de que foram feitos 566 registros, com 13 colunas, como mostra a imagem abaixo.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

O passo seguinte foi a busca por atributos com valores ausentes, que não foram localizados no conjunto de dados utilizado, conforme mostra a imagem abaixo.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

Antes da exclusão das colunas que foram consideradas irrelevantes para a indicação de cozinhas ou pratos, foram gerados diversos gráficos, a fim de procurar um melhor entendimento dos dados cadastrados no dataset.

O primeiro gráfico gerado foi o de distribuição de valores de vendas e os tipos de cozinha. De acordo com a imagem seguinte, o tipo de cozinha com maior soma de valores vendidos é a “mineira’, seguida por “churrasco”, “japonesa” e “francesa”.

Gráfico

Descrição gerada automaticamente

A fim de ter uma ideia da distribuição dos pedidos pelo tipo de cozinha, foi gerado um gráfico, conforme figura abaixo, no qual mostra que o tipo “churrasco” é o mais pedido, seguido por “francesa”, enquanto o menos pedido é “fastfood”.

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

Comparando-se os dois gráficos anteriores, é possível concluir que o tipo de cozinha mais pedido não é, necessariamente, o que gera maiores valores. Isso pode ocorrer pela variação dos preços de cada tipo de cozinha.

Assim, por exemplo, no gráfico de valores a cozinha japonesa aparece em terceiro lugar, enquanto no gráfico de quantidade de pedidos aparece em sexto lugar. Isto possivelmente ocorre porque os pratos da cozinha japonesa são, em geral, mais caros do que os pratos das cozinhas mineira, francesa, árabe e italiana.

Também foi gerado um gráfico exibindo a quantidade de vendas dos itens dos menus. Na primeira imagem não é possível identificar os pratos, mas nas imagens seguintes, que são recortes do gráfico original, é possível verificar que o prato mais pedido é “tiramisù”, seguido por “pão”, “ratatouille”, “filé do meio ao alho”, “picanha” etc, enquanto os menos pedidos são “frango gong bao”, “x-salada”, “esfirra aberta de calabresa”, “x-frango bacon”, “simplão” etc.

Gráfico, Histograma

Descrição gerada automaticamente

Gráfico

Descrição gerada automaticamente com confiança baixa

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Outro gráfico que foi gerado foi o que compara o valor total médio dos pedidos feitos em dias da semana ou finais de semana. Observa-se no gráfico a seguir que nos finais de semana são geradas mais receitas do que nos dias da semana.

Uma imagem contendo Texto

Descrição gerada automaticamente

Também foi gerado gráfico da quantidade de pedidos nos finais de semana ou dias da semana, conforme gráfico abaixo, ficando claro que o volume de vendas nos finais de semana é maior do que nos dias da semana, sendo, este, provavelmente o motivo pelo qual nos finais de semana há maior geração de renda.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Ainda foi gerado um gráfico com a distribuição de vendas em relação ao gênero dos clientes, observando-se na imagem a seguir que os clientes do sexo masculino fazem mais pedidos do que os de gênero feminino.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Outro dado que foi buscado, foi a indicação de quais os tipos de produtos mais vendidos, concluindo-se, pelo gráfico abaixo, que comida é o tipo mais pedido, seguido por bebida, sobremesa e bebida alcoólica.

Gráfico

Descrição gerada automaticamente

Para concluir a análise do conjunto de dados, foi gerado gráfico para indicação da quantidade de vendas por grupos etários, ficando evidente que o grupo que mais faz pedidos é o da faixa etária compreendida entre 40 e 60 anos, conforme imagem abaixo.

Gráfico

Descrição gerada automaticamente

Logo após, considerando-se os dados que foram selecionados como mais relevantes para o propósito do sistema, foram eliminadas algumas colunas, consideradas irrelevantes pelo grupo. Neste ponto, vale salientar que foram criados dois subconjuntos diferentes, mais tarde utilizados para que o algoritmo de recomendação fizesse duas sugestões, uma do tipo de cozinha e outra de itens dos menus (pratos).

Primeiro foi criado o subconjunto a ser utilizado para recomendação do tipo de cozinha, mantendo-se apenas as colunas “user\_id”, “age” e “cousine\_type”. Na próxima imagem é possível verificar que apenas os atributos mencionados foram mantidos no dataset.

Texto

Descrição gerada automaticamente

Utilizando outro projeto no Google Colab, foram eliminados atributos diferentes, tendo sido mantidos apenas os atributos “id\_cliente”, “age” e “food\_name”, utilizados para a recomendação de pratos aos clientes, conforme exibido na imagem a seguir.

Texto

Descrição gerada automaticamente

Depois de feita a remoção dos atributos que foram considerados, pelo grupo, irrelevantes o propósito do projeto, foi gerado um gráfico do tipo mapa de calor relacionando os tipos de cozinha com a idade dos clientes, sendo possível observar quais tipos de cozinha são mais apreciados por diferentes grupos etários. Por este gráfico, é possível confirmar que o grupo etário que mais faz pedidos é de 40 a 60 anos e o que menos faz pedidos é o de menores de 18 anos.

Uma imagem contendo Tabela

Descrição gerada automaticamente

Paralelamente, no projeto do Google Colab utilizado para a recomendação de itens do menu, onde também foram feitas as exclusões dos atributos considerados irrelevantes, sendo gerado gráfico do tipo mapa de calor relacionando os itens do menu com a idade dos clientes, possibilitando observar quais itens do menu são mais apreciados por diferentes grupos etários.

Tela de computador com texto preto sobre fundo branco

Descrição gerada automaticamente

É importante salientar que o objetivo da inserção de um algoritmo de mineração de dados ao projeto é bastante definido, ou seja, fazer recomendações aos clientes, o que limitou a escolha do algoritmo a ser utilizado, uma vez que não há, ao menos na atual fase do projeto “Um Treim di Cumê”, um atributo a ser classificado (como verdadeiro/falso, positivo/negativo ou bom/ruim, pertencente a classe A/B/C etc), impossibilitando a aferição de precisão e erro, criação de matriz de confusão e teste de outros algoritmos além do que foi selecionado (recomendação com KNN).

ALGORITMO DE RECOMENDAÇÃO – KNN

Seguindo a proposta do projeto “Um Treim di Cumê” de integrar ao projeto algoritmo de mineração de dados para que o sistema faça sugestões de produtos aos clientes, baseado no produto selecionado, foi implementação um algoritmo de recomendação, utilizando-se para isso o KNN.

Para isso, foi necessária a instalação de duas bibliotecas que são utilizadas para algoritmos de recomendação na linguagem Python, sendo elas “scikit-surprise” e “decouple”.

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Em seguida o algoritmo foi implementado, passando-se pelas etapas de definição do formato do leitor, carregamento do conjunto de dados, criação do conjunto de treinamento, criação do modelo KNN, simulação da escolha de um tipo de cozinha e de um prato pelo cliente e, finalmente a exibição das recomendações.

Texto

Descrição gerada automaticamente

A imagem exibida acima exibe os códigos do algoritmo para a recomendação de um tipo de cozinha, enquanto a imagem abaixo exibe os códigos do algoritmo para recomendação de pratos.

Texto

Descrição gerada automaticamente

Para o processamento do algoritmo, foi feita a simulação de um cliente escolhendo um prato e um tipo de cozinha e, baseado nas escolhas feitas, foram geradas recomendações para um tipo de cozinha e três pratos.

Texto

Descrição gerada automaticamente

No exemplo acima, simulou-se que o cliente pediu um prato chamado Picanha, cujo tipo de cozinha é Churrasco, obtendo-se a recomendação de tipo de cozinha Japonesa e dos pratos ‘combinado sushi salmão’, ‘saquê’ e ‘teppanyaki’.

Texto

Descrição gerada automaticamente Texto

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente Texto

Descrição gerada automaticamente

As imagens acima demonstram que foram feitas outras simulações com pratos e tipos de cozinha diferentes, observando-se que o tipo de comida recomendado é sempre “Japonesa”, havendo alguma variação nos pratos sugeridos.

Esta peculiaridade de repetição da sugestão de comida japonesa pode ocorrer pelo fato de o dataset ter sido construído artificialmente pelos integrantes do grupo, que fizeram inserções de usuários, restaurantes, menus e pedidos. Assim, estas inserções certamente foram influenciadas pelos gostos pessoais de cada integrante do grupo, o que acabou viciando o algoritmo, o que provavelmente não ocorreria com um conjunto de dados criado com dados reais e um número mais expressivo de registros.