Estudo de caso de mineração de dados com ambientes de monitoramento e atendimento ao cliente

Artigo do projeto referente à disciplina de Mineração de Dados

João Santos [2]
Laila Barros Campos [1, 2]
Matheus Andreoni [1]
Renan Albuquerque Villarim de Siqueira [1]

[1] Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil,
[2] Pós-graduação em Engenharia da Computação – PPGEC, Escola Politécnica de Pernambuco, Pernambuco, Brasil,

E-mail dos autores: João Santos jpbs@ecomp.poli.br, **Laila Barros Campos** lbc@ecomp.poli.br,
Matheus Andreoni mha@ecomp.poli.br, **Renan Albuquerque Villarim de Siqueira** ravs@ecomp.poli.br.

Resumo

Na área da tecnologia, é importante disponibilizar serviços de TI para usuários internos e externos. Entre esses serviços podemos citar o serviço de internet, e-mail, CRM, data analytics, atendimento aos clientes, entre outros. Neste trabalho, foi abordado uma análise acerca da base de dados referente ao P2D2 um sistema capaz de fazer um monitoramento das execuções dos Jobs em ambientes SAP a todo momento. A fim de descobrir os principais ofensores do sistema, aplicou-se o estudo da Mineração de Dados. O método empregado foi o CRISP-DM, utilizando algoritmos de agrupamento (Silhueta e Davies-Bouldin) com auxílio do software Pentaho para o pré-processamento dos dados. Como resultado, foi mostrado que os principais ofensores se localizam nos horários da noite. Os algoritmos de Silhueta e Davies-Bouldin apresentaram um melhor desempenho em um agrupamento de 3 clusters, no intuito de direcionar uma identificação mais precisa dos Jobs do P2D2.

Abstract

In the area of technology, it is important to make IT services available to internal and external users. Among these services we can mention the internet service, e-mail, CRM, data analytics, customer service, among others. In this work, an analysis about the P2D2 database was approached, a system capable of monitoring Jobs' executions in SAP environments at all times. In order to discover the main offenders of the system, the study of Data Mining was applied. The method used was CRISP-DM, using clustering algorithms (Silhouette and Davies-Bouldin) with the aid of the Pentaho software for data pre-processing. As a result, it was shown that the main offenders are located in the evening hours. The Silhouette and Davies-Bouldin algorithms performed better in a grouping of 3 clusters, in order to direct a more accurate identification of P2D2 Jobs.

1 Introdução

Um dos objetivos da área de tecnologia é disponibilizar serviços de TI para usuários internos e externos. Entre esses serviços podemos citar o serviço de internet, e-mail, CRM, data analytics, atendimento aos clientes, entre outros. Para que isso tudo funcione de maneira adequada, é importante realizar o monitoramento de aplicações e da infraestrutura que suporta esses sistemas e os serviços entregues pela área de TI. Ao monitorar aplicações, é possível obter uma visão global sobre o funcionamento dos softwares corporativos, que auxiliam a empresa a executar tarefas rotineiras e fundamentais

para o sucesso de seu negócio, como o atendimento ao cliente, por exemplo.

Segundo Davenport: "Um sistema corporativo impõe sua própria lógica sobre a estratégia, a cultura e a organização da empresa". São sistemas desenvolvidos para atender a gestão de toda e qualquer organização, de forma integrada, trazendo transparência, rapidez e confiabilidade para as informações corporativas. Um exemplo são os sistemas ERP, que existem várias opções no mercado, dentre elas o SAP ERP.

Os ambientes SAP são compostos por vários elementos diferentes, que são interdependentes. Um problema detectado por um usuário pode ser proveniente de componentes do SAP, do hardware, software ou elementos da infraestrutura subjacente. Para começar a obter uma maior performance, os administradores precisam criar linhas de base efetivas. Assim, é possível identificar rapidamente o surgimento de um problema.

1.1 Contextualização

Na empresa Accenture, a área de atendimento ao cliente é bastante ativa e requer monitoramento constante. Sérgio é o stakeholder envolvido no projeto e gerente de um time que dá suporte aos clientes através de atendimento por chamado.

Para auxiliar em sua rotina de trabalho, a equipe desenvolveu o P2D2, um robô capaz de fazer um monitoramento das execuções dos Jobs em ambientes SAP a todo momento. Essas execuções são automatizadas e existe uma meta de tempo para cada Job. Quando ocorre algum erro ou problema durante a execução de jobs, o robô envia e-mails e mensagens no Telegram para o time, solicitando um acompanhamento ou resolução daquele Job antes que provoque problemas maiores para o ambiente do cliente.

Certamente o P2D2 no momento tem a capacidade de acelerar bastante o trabalho da equipe com esse tipo de monitoramento, mas ainda assim é preciso realizar mais análises desses dados de monitoramento para melhorar a eficiência do uso do P2D2.

1.2 Descrição do Problema

O problema principal que a equipe da Accenture precisa lidar no momento com o atendimento ao cliente são os ofensores existentes no banco de dados que o P2D2 opera, que nada mais são do que os registros que possuem o status "Erro".

Apesar de ter o registro de todos os atendimentos realizados, o P2D2 não traz uma visão analítica dos principais ofensores e de como seria a melhor sequência para resolver esses problemas, além disso também é possível explorar soluções que combinem o horário dos erros com a performance do atendimento dos analistas da equipe.

Com a falta dessa análise fica mais difícil da equipe passar para os clientes uma visão mais precisa dos resultados da operação na resolução dos chamados.

1.3 Objetivo

O objetivo deste trabalho é oferecer uma visão analítica sobre os principais ofensores do sistema.

1.4 Justificativa

Atualmente a equipe de gerenciamento dos atendimentos se divide em turnos para monitorar esses Jobs juntamente com o P2D2, 24 horas por dia. O aprimoramento do robô é de extrema importância para que cada vez mais o trabalho mecânico da equipe se reduza ao ponto em que o esforço será totalmente concentrado em análise de dados e resolução dos problemas que os Jobs representam.

Além disso, os erros resultantes da execução do Job, se não forem tratados com certa agilidade, podem provocar problemas mais críticos aos ambientes disponíveis, influenciando na disponibilidade do mesmo e assim prejudicando os negócios dos clientes da Accenture.

1.5 Escopo Negativo

Não é objetivo deste estudo descobrir novas técnicas de mineração de dados. Este estudo também não visa criar uma interface gráfica para análise das informações por administradores e funcionários que trabalhem com o P2D2.

2 Fundamentação Teórica

2.1 Aplicações de Monitoramento de ambientes

Apesar das afirmações que contradizem o proposto, ninguém pode fazer multi tarefas manualmente o suficiente para acompanhar milhares de dados de uma só vez. Pelo menos não sem ajuda.

No caso do monitoramento de ambientes, precisamos saber se as coisas estão funcionando

como pretendemos o tempo todo. Ser capaz de ter certeza de que todos os ambientes estão funcionando corretamente, todos estão operando conforme o esperado e assim por diante.

Além de proporcionar uma visão clara do desempenho das suas aplicações, ao realizar o monitoramento, você reduz significamente os períodos de indisponibilidade dos serviços, pois saberá, em tempo real, quando ocorrer algum problema (downtime). Caso aconteça alguma falha, você já poderá atuar para solucioná-la no exato momento em que esta ocorrer.

Com os dados em mãos você terá capacidade de descobrir qual a real causa do problema e entender o porquê do mau funcionamento. De modo geral, quando ocorre algum problema, os usuários costumam reclamar em instantes, mas os profissionais de TI muitas vezes ainda estão "cegos" quanto à sua causa.

Caso você receba um alerta indicando essa causa raiz você terá capacidade para alocar o profissional certo para o problema em questão. Esse é um benefício que tem impacto direto nos custos da empresa, pois mobilizar pessoas desnecessariamente é um custo que muitas vezes não é contabilizado.

2.2 Mineração de dados

Mineração de dados é o processo feito de forma automática utilizado para extrair informações de grandes bases de dados [3]. Em uma época onde um número incontável de dados é gerado todos os dias, a mineração de dados pode ser considerada como uma evolução natural da tecnologia de informação [4].

A mineração de dados tem como objetivo identificar padrões úteis que não são possíveis de se encontrar, ou são ignorados por seres humanos, além de poderem fornecer previsões estatísticas do resultado de uma observação futura [3]. Esse tipo de análise pode ser utilizada em diversos meios, oferecendo vantagens aos administradores que o utilizam.

Para compreensão dos dados é possível utilizar diversas técnicas e algoritmos de mineração de dados, mas é necessário entender como os dados estão expostos e qual o tipo de aprendizado de máquina essas técnicas utilizam. Tradicionalmente é utilizado dois tipos de aprendizado de máquina: o aprendizado supervisionado e o não supervisionado [7]. O aprendizado supervisionado utiliza os dados conhecidos de entrada e de saída e tenta encontrar a melhor forma de chegar àquele resultado, esse tipo de técnica é utilizado para realizar predições em valores futuros. Já o aprendizado não supervisionado tentará encontrar um resultado a partir apenas dos dados de entrada, sem a existência de um rótulo presente na base, procurando por semelhanças e padrões para obter maior qualidade de resultados. Ambos os tipos de aprendizado de máquina possuem vantagens e desvantagens [7]. Cada técnica gera tipos de resultados diferentes, como é o caso da classificação que é utilizada para organizar grupo de dados em classes [8], sendo algoritmos que utiliza aprendizado supervisionado, ou a técnica de agrupamento que tem como objetivo agrupar os dados de acordo com suas semelhanças [9], com bases que não apresentam rótulos, sendo algoritmos que utilizam aprendizado não supervisionado. A escolha de qual técnica utilizar está fortemente conectada ao tipo de problema que será resolvido [5].

2.3 Métodos de Agrupamento

As técnicas de agrupamento dividem os dados em grupos. Cada objeto membro de um grupo apresenta similaridades entre os outros objetos daquele grupo e diferenças entre objetos de outros grupos. [9]. Encontrar o que define a semelhança entre os objetos é a chave para um bom agrupamento[2]. As técnicas de análise de grupos, ou os métodos de agrupamento se diferem de outras técnicas que rotulam dados por utilizar apenas as características dos próprios dados para agrupá-los [3]. Então, não existe um rótulo pré-determinado para agrupar os dados, esses rótulos são gerados de acordo com a análise feita na base. A análise e entendimento dos grupos de dados exerce um papel importante para diversas áreas. As técnicas de agrupamento podem ser utilizadas tanto para análise, verificando semelhanças na estrutura natural dos dados, quanto como um ponto inicial útil para outros tipos de análise, gerando resumo dos dados presentes [3]. Os métodos de agrupamento podem ser divididos em cinco tipos assim descritos: Particionamento, Hierárquico, Baseado em Densidade, Baseado em Grade, Baseado em Modelo[4]. Esse trabalho trabalhará com o Método Particionado. Considerado como o método mais simples de agrupamento, o método particionado é também um dos mais importantes. Esse método separa os dados em uma quantidade específica de grupos exclusivos, esses grupos utilizam funções baseadas em distância para identificar seus integrantes [3]. O K-Means é uma técnica de agrupamento que utiliza o método de particionamento. Nessa técnica um ponto é gerado aleatoriamente e um

grupo é criado com os pontos próximos a este. Esse processo é repetido diversas vezes até encontrar o melhor grupo. O ponto deverá ser o meio desse grupo, por isso o nome K-means [14]. O K-means aceita uma grande quantidade de dados, porém se limita a utilizar apenas dados numéricos [15].

2.4 Avaliação de Grupos

Com o intuito de avaliar os resultados obtidos pelas técnicas de agrupamento, algumas delas podem ser utilizadas dependendo do método escolhido. Essas técnicas podem ser utilizadas antes da mineração de dados, para decidir a quantidade de grupos, e após, para checar a qualidade dos grupos.

O Coeficiente de Silhueta (Silhouette Index) é caracterizado como técnica comumente utilizada para validar e interpretar os grupos gerados pelas técnicas de mineração de dados. Essa é uma técnica utilizada para verificar a qualidade de grupos gerados por Algoritmos de Particionamento. Essa técnica tem como objetivo demonstrar o quão alocado um objeto está em seu grupo [13]. Cada silhueta depende apenas dos grupos apresentados e de seus objetos, sendo assim, é possível utilizar diferentes técnicas de agrupamento para medir a qualidade do grupo [13]. Cada objeto de um grupo apresenta uma silhueta de valor que varia de -1 a 1. O Coeficiente de silhueta pode ser calculado com a fórmula [3]:

$$Si = \frac{Bi - Ai}{max(Ai, Bi)}$$

Onde: i é o objeto escolhido.Ai é a distância média de um objeto até todos os outros objetos do seu próprio grupo. Bi é a menor distância média de um objeto para qualquer grupo que não contenha esse objeto.

Um valor negativo no coeficiente de silhueta significa que o objeto está mais próximo de grupos ao qual não pertence que do próprio grupo, por isso, é desejado um valor maior que 0 e, principalmente, próximo a 1 [3]. A partir do valor encontrado no coeficiente de silhueta é possível detectar a qualidade do agrupamento. A Tabela a seguir apresenta a avaliação sugerida por Câmara ([1]) para a análise da qualidade do grupo de acordo com o coeficiente de silhueta.

Valor do Coeficiente	Interpretação
0,71 - 1	Grupos possuem uma estrutura muito robusta
0,51 - 0,7	Grupos possuem uma estrutura razoável
0,26 - 0,5	Grupos possuem uma estrutura fraca e pode ser artificial
-1 - 0,25	Nenhuma estrutura foi descoberta

Neste trabalho, também foi utilizado o índice de Davies Bouldin. O índice de Davies-Bouldin é uma medida de similaridade entre agrupamentos. O mesmo não depende do método de partição que foi utilizado e é comumente utilizado para a avaliação dos clusters formados. Esse índice é dado pela equação:

$$I_{DB} = \frac{1}{c}\sum_{k=1}^{c} max\{\frac{S_c(Q_k)+S_c(Q_i)}{d_{ce}(Q_k,Q_l)}\}$$

onde Q é um cluster, C é o número de clusters, e Sc é uma medida de similaridade intra cluster, dado pela equação:

$$S_c = \frac{1}{N_k}\sum_{i=1}^{N_k}\left|x_i - c_k\right|$$

Sendo, Nk o número de eventos pertencentes ao cluster de centróide Ck. O termo dce é a distância entre os clusters,

$$d_{ce} = \left|c_k - c_l\right|$$

Quanto menor for o índice de Davies-Bouldin, melhor foi o agrupamento do mapa obtido, e mais definidos e separados entre si se encontram os clusters formados.

3 Materiais e Métodos

A metodologia empregada neste trabalho foi o CRISP-DM. Esta metodologia foi desenvolvida por um consórcio que consiste nas empresas: NCR System Engineering (EUA, Dinamarca), Daimler-Chrysler AG (Alemanha), SPSS Inc. (EUA) e OHRA Verzekeringen e Bank Groep B.V (Holanda).

Esta metodologia foi escolhida por possuir um processo genérico que independe do setor e da tecnologia que esteja sendo utilizada, com isso o CRISP-DM fornece uma excelente base para o desenvolvimento de um modelo especializado

para o problema no qual ele está sendo utilizado [Referência da Bibliografia].

O CRISP-DM define uma sequência não rígida de seis fases, que permite a construção e implementação de um modelo de mineração para ser usado em um problema real, ajudando as decisões de negócios [Referência da Bibliografia]. Portanto, o desenvolvimento deste trabalho seguiu as fases do CRISP-DM, apresentadas na Figura 1.

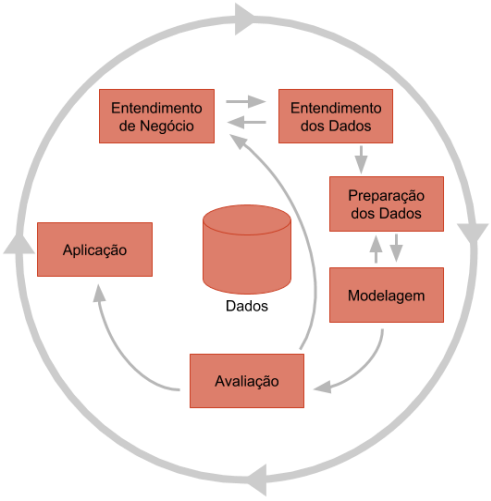

Figura 1: Sequência CRISP-DM.

Entendimento do negócio

Esta fase se concentra na compreensão dos objetivos e requisitos do projeto, em seguida, converter esse conhecimento em uma definição de problema de mineração de dados e um plano preliminar projetado para atingir os objetivos.

Entendimento dos dados

A fase de compreensão dos dados começa com a coleta de dados inicial e prossegue com atividades que permitem familiarizar-se com os dados, identificar problemas de qualidade de dados, descobrir primeiros insights sobre os dados e detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas.

Preparação dos dados

A preparação dos dados inclui registro e seleção de atributos, além de transformação e limpeza de dados para ferramentas de modelagem.

Modelagem

Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para valores ótimos. Representa o desenvolvimento dos modelos para o problema, com base nos dados que já foram adequados para serem utilizados.

Avaliação

Nesta fase do projeto, o modelo desenvolvido é avaliado e são revisadas as etapas executadas para criá-lo, a fim de ter certeza que atinge adequadamente os objetivos definidos. Por fim as informações obtidas sobre a eficiência dos modelos desenvolvidos serão avaliadas e distribuídas.

Aplicação

Todo conhecimento obtido por meio do trabalho de mineração e detecção de anomalias tornam-se subsídios para o desenvolvimento de estratégias que resolvam o problema proposto.

3.1 Descrição da Base de Dados Data Lake

Nesse projeto, foi fornecido uma planilha contendo dados coletados durante o monitoramento da execução de jobs nos ambientes. Nesta base de dados existem colunas de datas e tempos em que os jobs foram executados e quando foram finalizados, status do job quando a captura dos dados foi realizada e por fim a qual cliente e ambiente se trata o registro descrito até então.

Dicionário de dados

O dicionário desenvolvido na Tabela 1 faz uma descrição dos metadados do projeto, apresentando uma visão mais detalhada do tipo de dado que será trabalhado neste projeto, além de também informar o tamanho e valores permitidos nos campos.

Campo	Descrição	Tipo	Tam.	Val. permitidos
NOME DO CLIENTE	Cliente em que o job foi executado	TEXTO	9	Alfanuméricos
AMBIENTE	Ambiente em que o job foi executado	TEXTO	4	Alfanuméricos

JOB	Tipo do Job que foi executado	TEXTO	50	Alfanuméricos
DATA EXECUÇÃO	Data em que o job foi executado	TEXTO	8	Numéricos
HORA INÍCIO	Momento em que o job inicia sua execução	TEXTO	8	Numéricos
HORA FIM	Data em que o job finalizou sua execução	TEXTO	8	Numéricos
TEMPO EXECUCAO	Tempo que o job durou em execução	NÚMERO	8	Numéricos
META_ATUAL DE EXECUÇÃO	Meta calculada para o tempo de execução do job	NÚMERO	8	Numéricos
STATUS	Status do job	TEXTO	15	OK, EM ANDAMENTO, PLANEJADO, ERRO

Tabela 1: Dicionário dos metadados.

3.2 Análise Descritiva dos Dados

Visualização dos Dados

Com o entendimento prévio do tipo de dado que lidaremos neste trabalho é importante apresentar análises estatísticas para um melhor entendimento e visualização dos dados.

Na Figura 3 o histograma mostra a distribuição da quantidade de Jobs por ambiente para uma visualização geral de quais ambientes registram mais Jobs e pode-se notar uma prevalência de Jobs no AMB 9.

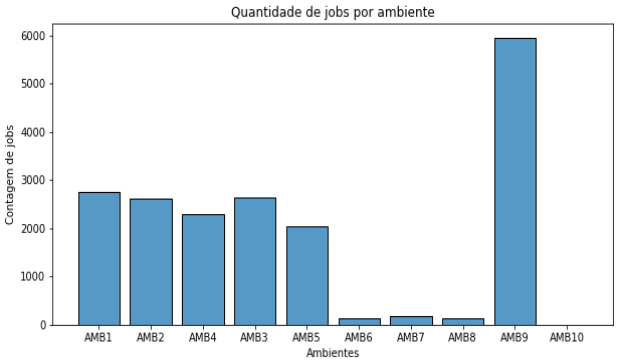

Figura 3: Quantidade de Jobs por ambiente.

Outra visualização importante dos Jobs por ambiente envolve a relação dos status de cada Job. A Figura 4 a seguir mostra com mais detalhes as informações apresentadas no histograma da Figura 3. Com esse gráfico de dispersão podemos visualizar quantos Jobs existem em cada ambiente de acordo com seu status. Nota-se na maioria dos ambientes uma prevalência de Jobs com status OK, principalmente no AMB 9.

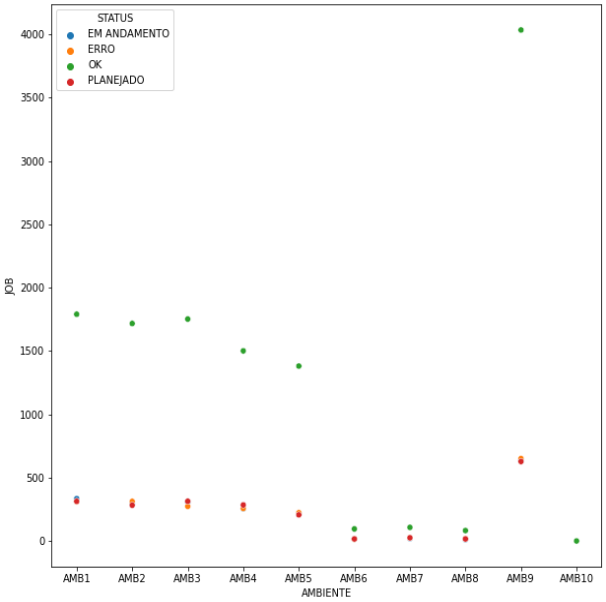

Figura 4: Quantidade de Jobs por ambiente de acordo com seus respectivos status.

Distribuição de frequência dos Jobs

Através de manipulações dos dados pudemos identificar que existem 10 tipos de ambientes diferentes, no qual/onde 8 deles pertencem ao que foi definido como Cliente 1 e os outros 2 (AMB4 e AMB5) pertencem ao Cliente 2, e um total de 18.735 Jobs registrados. Cada Job tem seu status definido como OK, Erro, Em Andamento e Planejado.

Na Tabela 2 a seguir pode-se visualizar numericamente a distribuição de frequência dos status desses Jobs por cada um dos 10 ambientes. É notável a maior frequência de Jobs classificados como OK, com uma frequência relativa de 66,52% e frequência praticamente equivalente dos outros status, registrando frequências relativas de 11,16%.

Status	Limite inferior	Ponto médio	Limite superior	Frequência absoluta	Frequência relativa
Em andament	12	323,5	635	2.091	11,16%
Erro	15	333	651	2.091	11,16%
OK	1	2.017	4.033	12.462	66,52%
Planejado	17	322,5	628	2.091	11,16%

Tabela 2: Distribuição de frequência dos status dos Jobs por ambiente.

Resumo dos Dados

Para efeito de resumo a Figura 5 ilustra a distribuição de frequência dos Jobs agrupados por status e por ambiente através de um gráfico box plot, que apresenta informações estatísticas importantes para uma visualização rápida do comportamento desses dados. Podemos ver novamente a prevalência de Jobs com status OK nos ambientes pela caixa desse status ser bem maior ao ser comparada com as dos outros status. É importante perceber também, como pôde ser visto na distribuição de frequência do tópico anterior, a equivalência da quantidade de Jobs para os status "Em andamento", "Planejado" e "Erro", mostrando suas caixas com tamanhos e alturas bem próximas.

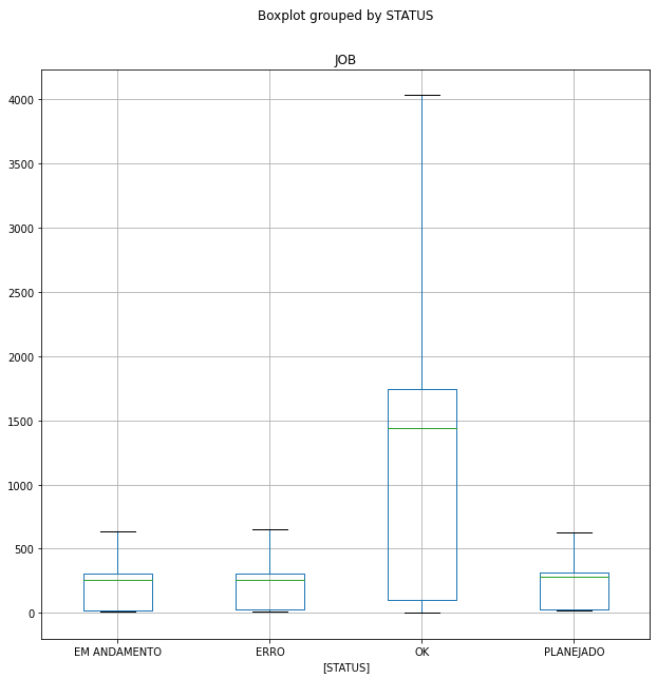

Figura 5: Gráfico boxplot da quantidade de dados por status.

3.3 Pré-Processamento dos Dados

Na etapa de pré-processamento de dados foi considerado que o algoritmo utilizado para agrupamento seria o k-means, este que foi criado para trabalhar com dados contínuos. Com este fato em mente, foi priorizado utilizar as colunas que possuíam dados contínuos, que envolviam data de execução dos jobs e os tempos de duração da execução dos mesmos. O pré processamento dos novos campos utilizados está especificado na tabela a seguir:

Antes	Depois	Processado
HORA INÍCIO	HORA_INICIO	Hora início antes estava no formato HH:mm:ss e na nova coluna ficou apenas HH
HORA FIM	HORA_FIM	Hora fim antes estava no formato HH:mm:ss, na nova coluna ficou apenas HH
-	VARIACAO_TEMPO_EXEC	Coluna gerada a partir da diferença

	UCAO	entre META_ATUAL DE EXECUÇÃO e TEMPO_EXECUCAO
STATUS	STATUS	"ERRO" foi convertido para 0 e "OK" foi convertido para 1
DATA EXECUÇÃO	ANO_DATA_ EXECUCAO	Data execução estava no formato yyMMdd, na nova coluna ficou registrado apenas yy
DATA EXECUÇÃO	MES_DATA_ EXECUCAO	Data execução estava no formato yyMMdd, na nova coluna ficou registrado apenas MM
DATA EXECUÇÃO	DIA_DATA EXECUCAO	Data execução estava no formato yyMMdd, na nova coluna ficou registrado apenas dd

Após serem selecionadas as colunas da base de dados que irão ser utilizadas no algoritmo, foi realizado uma normalização em cada coluna, no qual/onde o resultado foi uma nova tabela onde os valores de cada coluna se encontram no intervalo fechado entre 0 e 1 onde 0 representa o valor mínimo e 1 representa o valor máximo da respectiva coluna.

3.4 Metodologia Experimental

Agrupamento

Através do método de agrupamento k-means (método de Clustering que objetiva particionar n observações dentre k grupos onde cada observação pertence ao grupo mais próximo da média) foi realizada uma análise das colunas de **Hora de Início**, **Hora Fim e Tempo de Execução**, utilizando as seguintes métricas para avaliar a qualidade do agrupamento:

- Silhouette Score
- Davies-Bouldin

Experimentalmente houve uma variação do número de clusters do k-means utilizando respectivamente o valor de 2, 3, 4, 5, 6 e 15 clusters, com o intuito de encontrar uma divisão de clusters que, dada as métricas para este cálculo, encontrasse um valor satisfatório. Após esse experimento, foi constatado que o valor de 3 clusters era o mais apropriado para a progressão da análise, pois as métricas apresentavam os melhores resultados.

4 Análise e Discussão dos Resultados

4.1 Resultados

Após a definição do uso de 3 clusters no k-means, através dos algoritmos de agrupamento conseguiu-se ter um resultado satisfatório e obteve-se o seguinte resultado para as respectivas métricas:

- Silhouette = 0.86257
- Davies-Bouldin = 0.27760

Com isso foi possível gerar o gráfico de dispersão abaixo, concluindo que é possível agrupar os jobs em 3 grupos.

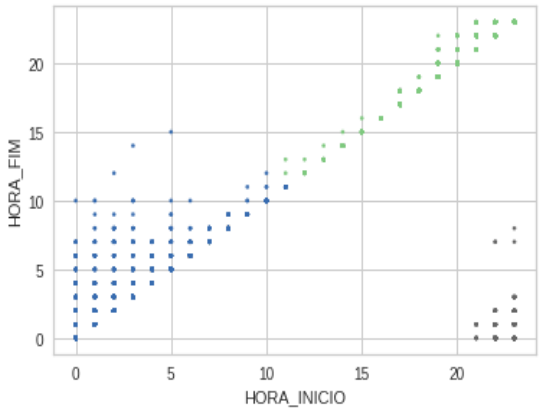

Figura 6: Gráfico de dispersão resultado de algoritmos de agrupamento.

Inicialmente observa-se que o agrupamento realizado gerou 3 grupos cujo suas características são:

- Grupo Azul: Jobs que iniciam entre as 00:00 até 10:00 e encerram sua execução em 00:00 até no máximo 15:00
- Grupo Cinza: Jobs que iniciam no final de um dia e finalizam no início do dia seguinte.
- Grupo Verde: Jobs que iniciam no intervalo de 15:00 até 23:00 e possuem

seu horário de finalização nesse mesmo intervalo.

Observando por outros pontos de vista os resultados obtidos é possível extrair informações sobre horário em que os jobs são ativados.

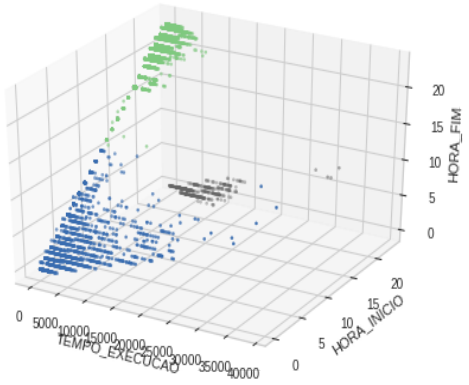

Figura 7: Gráfico de dispersão resultado de algoritmos de agrupamento comparando hora início, hora fim e tempo de execução

É possível notar que os jobs do grupo azul possuem várias ocorrências de jobs com tempo de execução maior que os demais.

Após verificar tais características, foi necessário observar os resultados a partir de outras variáveis do sistema a fim de obter informações sobre os principais ofensores.

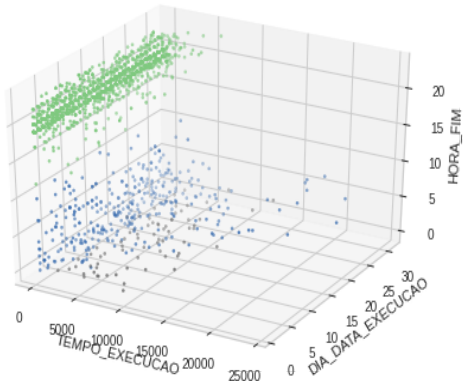

Figura 8: Gráfico de dispersão resultado de algoritmos de agrupamento comparando hora fim x tempo de execução x dia data execução

O gráfico acima foi gerado considerando-se apenas jobs com o status de erro. Foi possível observar que a maioria dos jobs que possuem status de erro estão localizados na região em que os valores de hora de finalização do job está compreendida entre 15:00 e 20:00.

Buscando compreender mais sobre os horários de execução dos jobs, foi feito um agrupamento considerando apenas suas horas de início e em seguida exibir os resultados em um gráfico de barras, o que trouxe o seguinte resultado:

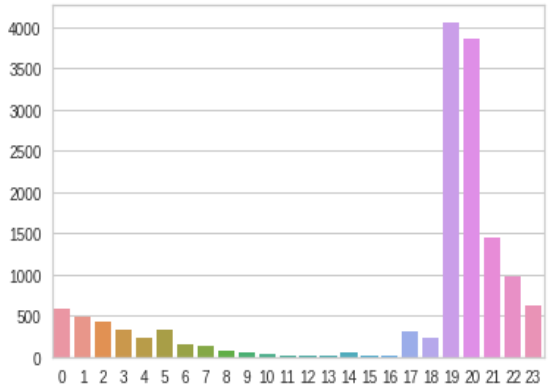

Figura 10: Gráfico de barras comparando hora de início X frequência

Com esse resultado em mente, foi levantada a seguinte hipótese: "O horário de início do job afeta, de alguma forma, a proporção de erros causados nesse determinado intervalo de tempo?". Buscando respostas para esse questionamento então foram realizadas análises descritivas dos dados a fim de obter métricas de tendência central para cada hora de início presente na base de dados. Foram obtidos os seguintes valores para as médias de jobs que possuíam status OK.

Intervalo de Hora de início	Média
00:00 até 06:00	0,857
07:00 até 18:00	0,856
19:00 até 20:00	0,856
21:00 até 23:00	0,855

Percebeu-se que a proporção de jobs que possuem erro possuem valores muito similares independente dos intervalos observados na tabela acima.

4.2 Discussão dos Resultados

A partir da análise dos resultados foi possível concluir que há uma distribuição não equilibrada dos Jobs durante os horários do dia, onde grande parte das ocorrências dos ofensores ficam concentradas no horário da noite.

Possivelmente estudos sobre novas alternativas de distribuições dos acionamento desses jobs devem ser estudados visando diminuir a concorrência de execução entre os jobs.

5 Trabalhos Futuros

Visando futuramente uma melhor definição e monitoramento das metas de execução dos Jobs, a aplicação de uma Rede Neural Artificial simples que seja capaz de treinar e prever essas metas de forma mais realista à medida que mais dados vão sendo adicionados ao banco de dados, ou então testes de predição dessas metas com algoritmos de regressão, visando reduzir sinalizações desnecessárias dos atrasos dos jobs.

6 Referências

[1] CÂMARA, G.; MACIEL, A.; VINHAS, L. Algoritmos de clustering para separação de culturas agrícolas e tipos de uso e cobertura da terra utilizando dados de sensoriamento remoto.2015.

[2] DUTT, A.; ISMAIL, M. A.; HERAWAN, T. A systematic review on educational data mining. Ieee Access, IEEE, v. 5, p. 15991–16005, 2017.

[3] TAN, P.-N.; STEINBACH, M.; KUMAR, V. Introdução ao datamining: mineração de dados. [S.l.]: Ciência Moderna, 2009.

[4] HAN, J.; PEI, J.; KAMBER, M. Data mining: concepts and techniques. [S.l.]: Elsevier,2011.

[5] CHANDRA, E.; NANDHINI, K. Knowledge mining from student data. European journal of scientific research, Citeseer, v. 47, n. 1, p. 156–163, 2010.

[6] COSTA, E.; BAKER, R. S.; AMORIM, L.; MAGALHÃES, J.; MARINHO, T. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. Jornada de Atualização em Informática na Educação, v. 1, n. 1, p. 1–29, 2013.

[7] YE, S.; LI, Z.; MCCANN, M. T.; LONG, Y.; RAVISHANKAR, S. Unifiedsupervised-unsupervised (super) learning for x-ray ct image reconstruction. arXiv preprintarXiv:2010.02761, 2020.

[8] UMADEVI, S.; MARSELINE, K. J. A survey on data mining classification algorithms. In:IEEE. 2017 International Conference on Signal Processing and Communication (ICSPC). [S.l.],2017. p. 264–268.

[9] BERKHIN, P. A survey of clustering data mining techniques. In: Grouping Multidimensional data. [S.l.]: Springer, 2006. p. 25–71.

[10] MARUTHO, D.; HANDAKA, S. H.; WIJAYA, E. et al. The determination of clusternumber at k-mean using elbow method and purity evaluation on headline news. In: IEEE. 2018International Seminar on Application for Technology of Information and Communication. [S.l.],2018. p. 533–538.

[11] HLAB. Error Sum of Squares (SSE). Disponível em: <https://hlab.stanford.edu/brian/error_sum_of_squares.html>.

[12] SYAKUR, M.; KHOTIMAH, B.; ROCHMAN, E.; SATOTO, B. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In:IOP PUBLISHING. IOP Conference Series: Materials Science and Engineering. [S.l.], 2018.v. 336, n. 1, p. 012017.

[13] ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, North-Holland, v. 20, p.53–65, 1987.

[14] MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. [S.l.], 1967. v. 1, n. 14, p. 281–297.

[15] HUANG, Z. A fast clustering algorithm to cluster very large categorical datasets in data mining. DMKD, Citeseer, v. 3, n. 8, p. 34–39, 1997.