INSTITUTO SUPERIOR
DE TECNOLOGIAS
AVANÇADAS DE LISBOA

Privacidade num mundo com crescente *datamining*

Gercia Rosa Bastos Fonseca

Dissertação para obtenção do Grau de Mestre em

INFORMÁTICA

Júri

Presidente: Professora Doutora Andreia Cristina Teles Vieira

Orientador: Professor Doutor Pedro Ramos dos Santos Brandão

Arguente: Professor Doutor Luís Carlos Gonçalves

Abril, 2022

Agradecimentos

Ao meu orientador por todo o apoio e encorajamento, para não desistir nestes tempos difíceis ao longo deste projeto, aos meus professores é colegas ,bem como ao ISTEC pela excelência do ensino.

Resumo

A rápida evolução do tráfego de dados nas redes sociais, que pode ser utilizado em estudos de análise de sentimento sobre determinados produtos, marcas, serviços etc. Além disso, os campos da computação em nuvem foram uma das áreas mais interessantes em estudos de pesquisa. Neste trabalho, recorremos à análise de dados de um dos principais provedores de serviços em nuvem: o *Microsoft Azure*, para analisar a equidade e a privacidade de um processo de recrutamento online. Para o fazer, um conjunto de dados foi extraído de uma base de dados *open source*, os quais consistem em dados pessoais disponibilizados num inquérito. Nós estudamos e analisámos a forma de tornar o processo mais justo, em detrimento da raça ou sexo. A esse respeito, muitas organizações tendem a aplicar a política da diversidade aos seus recursos humanos de forma eficaz para tentar obter um impacto positivo. Os resultados são analisados e explicados, em detalhe, em termos de classificações de polaridade para demonstrar o impacto da análise de equidade e para apoiar as decisões das organizações. Podemos observar nos resultados da classificação de regressão linear que a categoria *race* é melhor para o *Microsoft Azure* com o uso de uma *API*, em comparação com a regressão simples, a percentagem de *sex* é maior para a regressão simples em comparação com o *Fairlearn*.

Palavras-chave: *Fairleran*, privacidade, classificação, regressão, *Python*, *Azure, machine learning.*

ABSTRACT

The rapid evolution of data traffic in social networks can be used in sentiment analysis studies on certain products, brands, services... etc. In addition, cloud computing has been one of the most interesting areas in research studies. This paper, uses data analysis from one of the leading cloud service providers: Microsoft Azure to analyze privacy's fairness in an online recruitment process. To do this, a dataset is extracted in an open-source database, which consists of personal data made available in a survey. We study and analyze how to make the process fairer at the expense of race or gender. In this respect, many organizations tend to apply diversity policy on their human resources effectively and try to get positive impact. The results are analyzed and explained in detail in terms of polarity rankings to show the impact of equity analysis to support organizations' decisions. We can observe from the linear regression ranking results that race category is better for Microsoft Azure with the use of an API compared to simple regression; sex percentage is higher for simple regression as compared to *Fairlearn*.

Keywords: *Fairleran*, privacy, classification, regression, Python, Azure Machine Learning

Índice

Agradecimentos vii

Resumo ix

1. INTRODUÇÃO 1

1.1. Objetivos 5

1.2. Organização do documento 6

2. ESTADO DA ARTE 9

2.1. Revisão sistemática da literatura 9

2.1.1. Data Mining 12

2.1.2. Privacidade dos dados 15

2.1.3. Datasets 21

3. METODOLOGIA 33

3.1. Análise dos dados 34

3.2. Processo de mineração de dados 37

3.3. Implementação de processo de mineração de dados 41

3.3.1. Módulo de mineração em *Azure* com *Fairlearn* 42

4. RESULTADOS E DISCUSSÃO 45

4.1. Testes de modelo de mineração 46

4.2. Testes de desempenho 50

5. CONCLUSÕES 52

5.1. Síntese de discussão de resultados 53

5.2. Principais contribuições 54

5.3. Trabalho futuro 54

6. REFERÊNCIAS 56

Índice de imagens

Figura 1 – Diagrama de Venn identificação de dados 2

Figura 2-Diagrama de CRISP-DM 3

Figura 3 - Esquema de anonimização numa rede social 18

Figura 4 - Esquema geral do processo de mineração de dados 37

Figura 5 – Pré-processamento em Python :Carregar,adicionar,juntar e particionar em paralelo 43

Figura 6 - Implementação do algoritmo SVM..... 43

Figura 7- Diferença de falsos negativos..... 47

Figura 8 - Diferença de falsos positivos 47

Figura 9 - Taxa de desempenho com a métrica de precisão 48

Figura 10 -Taxa de desempenho com Recall..... 48

Figura 11 -Taxa de desempenho com F1-meseaure 49

Figura 12 - Taxa de desempenho do atributo raça 49

Figura 13 - Taxa de desempenho do atributo sexo com os algoritmos de métrica de desempenho..... 50

Figura 14 - Taxa de desempenho do atributo sexo..... 50

Figura 16 – Performance da máquina virtual 52

Índice de tabelas

Tabela 1 - Protocolo de revisão da leitura..... 10

Tabela 2 Caraterísticas dos dados demográficos sensíveis..... 36

Tabela 3 Contagem de instâncias por raça..... 38

1. INTRODUÇÃO

A mineração de dados, consiste no processo de extração de informação útil, interessante e previamente desconhecida de grandes conjuntos de dados. O sucesso da extração de dados depende da disponibilidade de dados, e da partilha eficaz da informação. A recolha de informação digital por parte de governos, empresas e indivíduos, criou um ambiente que facilita a prospeção e análise de dados em larga escala. Além disso, impulsionada por benefícios mútuos, ou por regulamentos, que exigem a publicação de determinados dados, existe uma procura de partilha de dados entre várias partes. Por exemplo, os hospitais licenciados na Califórnia são obrigados a submeter os dados demográficos específicos, acerca de cada paciente que recebe alta das suas instalações. A partilha desta informação tem um longo historial na tecnologia da informação. A partilha tradicional de informação refere-se ao intercâmbio de dados entre um detentor de dados e um recetor de dados. Por exemplo, o intercâmbio eletrónico de dados (EDI) é uma implementação bem-sucedida da transmissão eletrónica de dados entre organizações, com ênfase no sector comercial.

O desenvolvimento do EDI teve início nos finais de 1970 e continua a ser atualmente utilizado, os termos "Privacidade de dados" e "*data mining*" referem-se não só ao modelo tradicional um-a-um, mas também aos modelos mais gerais com múltiplos detentores e recetores de dados. A recente normalização dos protocolos de partilha de informação, tais como a linguagem de marcação extensível (XML), o Protocolo de Acesso a Objetos Simples (SOAP), e a linguagem de descrição de serviços web (WSDL) são catalisadores do recente desenvolvimento da tecnologia na partilha de informação.

A anonimização refere-se à abordagem PDDM que procura ocultar a identidade e/ou dados sensíveis dos proprietários de registos, assumindo que os dados sensíveis devem ser retidos para análise de dados. É evidente que os identificadores explícitos dos proprietários de registos devem ser removidos. Mesmo com todos os identificadores explícitos removidos, Sweeney [1] demonstra uma ameaça à privacidade na vida real sobre William Weld, antigo governador do estado de Massachussetts. No exemplo de Sweeney, o nome de um indivíduo numa lista pública de eleitores foi ligado ao seu registo, numa base de dados publicada através da combinação de código postal, data de nascimento, e sexo, conforme apresentado na figura 0.1, cada um destes atributos não identifica de forma única um proprietário de registo, mas a sua combinação, chamada de quase-identificador [2]

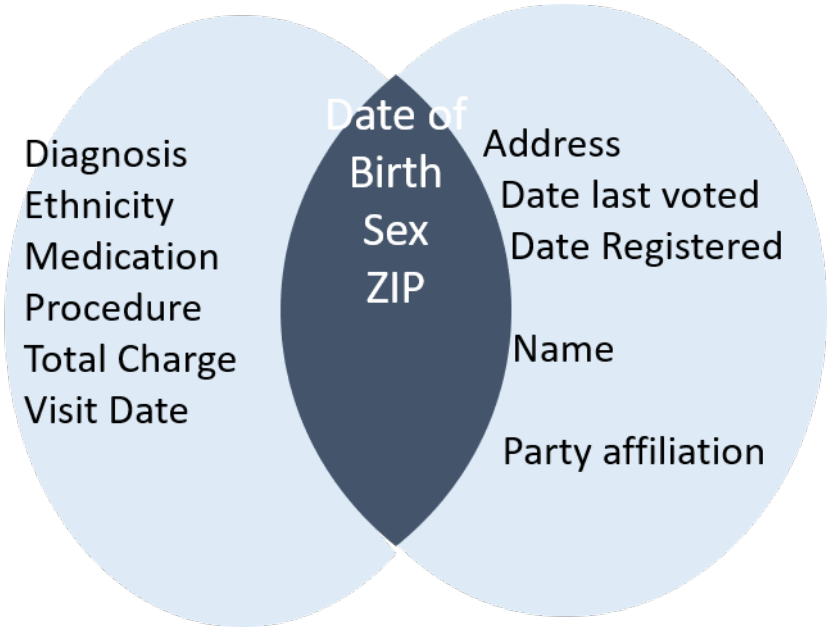

Figura 1 – *Diagrama de Venn identificação de dados*

O CRISP-DM continua a ser a metodologia de topo para projetos de mineração de dados com, essencialmente, a mesma percentagem que em 2007 (43% vs 42%). No entanto, é reportada a sua utilização em pelo menos 50%. O mesmo foi concebido por volta de 1996, por um grupo de empresas [3]. E pode ser descrito por 6 fases (ver Figura 1.1). Esta metodologia não engloba ainda a realidade de *Big data*. O site original crisp-dm.org já não está ativo, e o *IBM SPSS Modeler* é, provavelmente, a única ferramenta que ainda o inclui. Uma justificação para a falta de metodologias modernas é o aumento significativo de pessoas que utilizam as suas próprias metodologias. Talvez seja mais encorajador utilizar ferramentas de trabalho próprias, quando a falta de metodologia é suscetível de produzir descobertas aleatórias e falsas, e nenhuma das pessoas relatou não utilizar qualquer metodologia.

***Figura 2**-Diagrama de CRISP-DM*[1]

[1]Fonte: https://commons:wikimedia:org/w/index:php?curid_24930610 Por Kenneth Jensen CC BY-SA 3.0

Atualmente, o processo padrão do setor para a mineração de dados é o *CRISP-DM [4]*. Trata-se de um modelo de processo amplamente difundido, para todos os tipos de estudos de mineração de dados e aprendizagem de máquina. O seu objetivo consiste em fornecer uma diretriz de desenvolvimento, para todos os tipos de aplicações baseadas em dados. Isto foi realizado por um modelo cíclico que contém as seguintes fases:

1. Pesquisa e estudo (*Business Understanding*) – determinados os objetivos que se pretende atingir, avaliar as variáveis determinantes, pontos chaves do estudo, descrever as espectativas e efetuar uma avaliação da situação, determinar os objetivos da mineração de dados, e produzir o plano do projeto, pesquisar acerca da política de privacidade da fonte de dados.
2. Analise exploratória dos dados (*data understanding*) - Nesta fase, e abordadas as fontes de extração dos dados, descrevem-se os dados, os números de amostras e as suas características, de cada um dos conjuntos de dados. É frequente a exploração dos dados, como por exemplo por via da aplicação de algumas técnicas estatísticas, como a média, para ajudar a entender a complexidade dos mesmos, e verifica-se também a qualidade dos dados, para que o resultado final não seja influenciado por dados incompletos ou eventuais erros.

3. Preparação dos dados (*Data Preparation*)[2] - A terceira fase consiste na preparação dos dados, o que inclui a descrição do conjunto de dados, bem como a seleção dos dados (*feature selection*), para incluir ou excluir casos particulares; e posteriormente a limpeza dos dados, construir os dados, o que significa aspetos tais como: atributos derivados e registos gerados, e depois

[2] https://www:kdnuggets:com/2014/10/crisp-dm-top-methodologyanalytics-data-mining-data-science-projects:html - consultado a 4 de novembro de 2021

integrar os dados para fundi-los com outras fontes. Para a seguir formatar os dados para que se ajustem aos dados existentes e ao software que se vai utilizar.

4. Avaliação - é aqui que se podem avaliar os resultados, ver como funcionam no que respeita aos critérios de sucesso empresarial. E depois pode rever o processo, determinar os próximos passos, elaborar uma lista de ações possíveis ou talvez uma decisão final.
5. Modelagem - é aqui que seleciona a técnica de modelagem e verifica os pressupostos, gera o desenho do teste, constrói o modelo - incluindo a análise das definições dos parâmetros e a descrição do modelo. Depois pode avaliar o modelo, isto é, analisar aspetos tais, como o quão bem ele se ajusta aos dados, e se precisa de rever qualquer um dos parâmetros que utilizou.
6. Implementação - em se aplicam as conclusões ao problema a ser resolvido, a razão pela qual se concebeu o projeto em primeiro lugar, o que inclui planear a implementação e planear a monitorização e a manutenção, verificar até que ponto se mantém, produzir o relatório ou a apresentação final, e rever o projeto, incluindo a documentação de todo o processo, para que possa ser avaliado e possivelmente replicado no futuro.

1.1. Objetivos

Segundo LeFevre [5] os processos de mineração de dados, são, na sua maioria, implementados em ferramentas *open-source,* mas muitas das vezes, as aplicações de técnicas de privacidade neste modelo são triviais. Sendo verdade que a sua implementação é discutível e, frequentemente, ferramentas mais elaboradas acabam por nos apresentar resultados similares, então, cada tipo de tecnologia de proteção de informação, deve ser escolhido de acordo com o problema a ser resolvido.

Esta dissertação tem o objetivo de comparar, dentro de um ambiente de armazenamento distribuído, as técnicas para a preservação de dados, procurando encontrar o ponto, sobre o qual a utilização de uma tecnologia se torna mais vantajosa face a outra. Neste caso, foram realizados testes com diferentes algoritmos para determinar a tecnologia mais vantajosa. O conjunto de dados foi obtido num repositório de censo populacional (ver o Capítulo 3 para mais detalhes).

1.2. Organização do documento

O documento encontra-se organizado da seguinte forma:

Introdução – Primeiramente, são explicadas algumas definições referentes à preservação da privacidade dos dados e a sua evolução, depois é introduzido o contexto de *data mining* e consequentemente, de *data analytics.* Isso assim foi feito para explicar melhor esses dois campos de pesquisa. Essas informações permitem apresentar melhor a questão da pesquisa, que norteia esta dissertação. E é também, apresentado o objetivo principal.

Estado da arte – Trata da fundamentação teórica dos temas desta dissertação. De acordo com Demo [6] esse capítulo é de extrema importância para uma pesquisa científica, uma base teórica afirma o poder para um caráter explicativo. Para a construção da mesma, primeiramente, é apresentada a revisão sistemática da literatura, em seguida, a exploração de conceitos e plataformas atuais onde estes algoritmos são implementados - Capítulo 2.

Metodologia – Após a segmentação da base teórica, e apresentado o método de pesquisa utilizado neste trabalho bem como os resultados de campo. Não há

ciência sem o emprego de métodos científicos, segundo Marconi e Lakatos, para tal, são apresentadas as abordagens de pesquisa assim como as técnicas em que a dissertação se baseia. Nesta dissertação, foi realizado um estudo de caso múltiplo, de caracter exploratório, para poder investigar como as empresas utilizam os dados disponibilizados num processo de recrutamento, com o auxílio da análise de dados. Portanto, também são apresentadas as etapas para a realização dessa mesma pesquisa - Capítulo 3.

Resultados da investigação - São apresentadas a análise individual de cada pesquisa de campo e a análise comparativa entre as mesmas - Capítulo 4.

Conclusões - discussão dos resultados obtidos, trabalhos futuros e sugestões - Capítulo 5.

2. ESTADO DA ARTE

Segundo Demo [6] é crucial que uma pesquisa académica tenha uma base teórica sólida para uma melhor explanação. Foram seguidos alguns fundamentos tais como a definição de conceitos-chave e o estudo da bibliografia existente.

Após a exploração das referências bibliográficas mais relevantes e a definição dos conceitos-chave, foi realizado um mapeamento das publicações existentes em diversas bases de dados, a respeito da temática da privacidade de dados e *data mining.* Este capítulo, também contém uma seção totalmente focada na teoria do tema proposto.

2.1. Revisão sistemática da literatura

Tranfield, Denyer e Smart [7], definem a revisão da literatura como um procedimento, que deve ser utilizado para tratar a diversidade do conhecimento extraído. O objetivo principal é mapear e avaliar o campo intelectual existente. Foi adaptada uma revisão sistemática da literatura, com base nas definições do autor referenciado, extração do conhecimento existente na área de *data mining* e da preservação da privacidade.

A revisão sistemática realizada num primeiro momento, não apresentou resultados contundentes, nesse instante foi utilizado em conjunto as palavras-chave “privacidade de dados” e “privacidade em *data mining*”. Por essa razão, foi desenvolvida uma busca modificada da leitura, a tabela apresenta o protocolo de revisão de leitura do segundo momento, da que foi realizada em janeiro de 2021.

Tabela 1 *- Protocolo de revisão da leitura*

Termos-chave da pesquisa	Sessão 1 - Privacidade dos dados Sessão 2 – *Data mining privacy*, cloud data mining*, cloud mining services**
Operadores booleanos	AND entre busca, OR entre grupos
Base de dados	IEEE *Xplore*
Áreas de pesquisa	*Data privacy, cloud computing, authorization, security of data, social networking, data protection.*
Critérios de exclusão	Conteúdos relacionados com a área da saúde ou medicina; Conteúdos que apresentam, uma única vez, o termo privacidade dos dados, no texto completo - sem considerar o título, resumo ou palavras-chave.
Idioma	Inglês
Tipos de documento	Artigos, conferências
Anos da publicação	Sem filtro[3]

As palavras apresentadas nos termos de busca, foram pesquisadas no título, palavras-chaves e artigos dos conteúdos. Foi utilizado o termo de busca *data mining*, ao invés de uma outra especificação para a área da ciência dos dados, por dois motivos: primeiramente, porque *data mining* engloba todos os conceitos relacionados com a ciência de dados, e em segundo lugar, porque a pesquisa é mais especifica e restringe a quantidade dos resultados da mesma.

O símbolo * na segunda sessão, em alguns termos significa que o sufixo destas palavras pode variar. Dessa forma, foi possível tecer algumas derivações, que

[3] Fonte: Adaptado de Demo

poderiam abranger outras referências das mesmas palavras-chave: *Data mining privacy, cloud data mining, cloud mining services.*

Não foi definida uma linha temporal durante a pesquisa, pois todos os conteúdos relacionados foram publicados a partir de 1991, trata-se de um dado relevante, pois o tema proposto é uma tendência recente, e a quantidade de conteúdo relacionado está em procura crescente.

Pode existir uma variante nos resultados, e isso deve-se ao facto de a revisão sistemática da leitura ser efetuada numa altura diferente. Num primeiro momento a pesquisa obteve um resultado de 574 artigos.

Foi feita uma comparação com as revisões sistemáticas realizadas nos dois momentos, para perceber a relevância dos conteúdos encontrados. Durante a pesquisa notou-se um rápido crescimento dos conteúdos que abordam este tema.

Após a leitura dos resumos de todos os conteúdos obtidos na pesquisa, foram aplicados os critérios de exclusão (detalhados na Tabela) aos 574 artigos, para analisar os mais relevantes. Foram aplicados critérios de exclusão, foram removidos mais de 80% dos dados encontrados. Entre os quais, mais de metade estavam relacionados com a área da saúde, e um pouco mais de um terço não estavam relacionados a ciência da computação - critérios estes apresentados na tabela.

Contudo, foram analisados 47 artigos, este número pode ser justificado por o *data mining* ser um tema relativamente recente, isto em relação à *data science* e a privacidade dos dados nesta área ser um tema pouco explorado.

Dos quarenta e sete artigos analisados, apenas três autores escreveram mais de um artigo, em relação as palavras-chave, foram consideradas as *keywords* fornecidas pela base de conhecimento. Foram consideradas outras palavras-chave diferentes, mas relacionadas. Foi observado que muitas das palavras-chave que

foram citadas estão associadas à ciência de dados, o que tornou a pesquisa mais ampla.

Seguidamente, foram analisadas as definições de privacidade de dados e *data mining*; os algoritmos relativamente à proteção dos dados; o seu relacionamento com a privacidade dos dados e algumas questões relacionadas com a legislação em vigor. Os resultados obtidos encontram-se descritos nas próximas subsecções, intituladas "*Data mining*" e "Privacidade de dados".

2.1.1. Data Mining

Feng, [7] afirma que o *data mining* consiste no agrupamento de características e modelos que estão implícitos num conjunto de dados, com objetivo de encontrar uma correlação, que possa agregar benefícios a um determinado negócio, através de determinados algoritmos e da análise de dados. As técnicas de *data mining* distinguem-se dos tradicionais métodos estatísticos, pela capacidade de serem implementados métodos analíticos mais avançados, tais como: *machine learning*, *business inteligence* e inteligência artificial.

De acordo com Leskovec, Rajaraman e Ullman [8], os modelos de *data mining* são capazes de interpretar, extrair e prever, através de um conhecimento de uma fonte de dados. Com a crescente procura de dados, os métodos tradicionais tornaram-se obsoletos e incompatíveis, dado que os requisitos primordiais para a sua execução é o entendimento da heurística dos dados. A resolução desta problemática consiste na utilização de métodos baseados em dados.

Hand, Mannila e Smyth [9] defendem que, quanto maior for o conjunto de dados, melhor será a precisão dos resultados obtidos, pois o espaço de execução do algoritmo é maior, e mais amostras serão analisadas.

A modelagem de dados é categorizada como: descritiva e preditiva. A descritiva categoriza as propriedades intrínsecas aos dados, para revelar detalhes da informação, o que auxilia na tomada decisão. Este método é utilizado para a encontrar padrões ou tendências, que auxiliam na interpretação dos dados, segundo a análise de Jiawei [10]. A modelação preditiva, prevê um atributo num universo de dados - um atributo em relação aos outros - é utilizado o histórico para desenvolver o modelo de dados, para tal, Santos [11] defendeu que cada objeto do conjunto de testes deve possuir atributos de *input* e *output,* e uma variável *target.*

Ester et al [12] define quatro etapas para a descrição de uma base de conhecimento: a segmentação, as regras de associação, o *link analysis* e a visualização. O *data mining* na sua análise exploratória utiliza algoritmos de *machine learning.*

Existem distintos métodos de aprendizagem dos dados para uma análise explanatória, aprendizagem supervisionada e não supervisionada, como descreveu Bengfort [13]. Estas técnicas oferecem uma gama de algoritmos, que desenvolvem modelos capazes de criar perspetivas de um determinado negócio. Se o processo de classificação for baseado em atributos e classes, é utilizado o modelo da aprendizagem supervisionada, este modelo é baseado numa pré-classificação, para que seja possível supervisionar o modelo de aprendizagem, tendo a possibilidade de avaliar a performance do mesmo, e alterar as variáveis ambientes. A diferença deste modelo com o da aprendizagem não supervisionada segundo Dean [10], é que o atributo de *output* é ignorado. Não existe um atributo para classificação, é e baseado em padrões de semelhanças de dados.

Support Vector Machine (SVM) é definido como um algoritmo que procura mapear pontos, dos atributos dos dados, e procura encontrar um ambiente de decisão entre os mesmos, tendo em conta a sua classe, segundo foi definido por Hand [14]

com isso é possível supor que, novas amostras podem pertencer a um dos conjuntos definidos. O mesmo autor indica que uma tarefa de classificação é baseada em técnicas estatísticas, seja h um classificador H é o conjunto de dados. Durante o processo de mineração de dados, o algoritmo utiliza um conjunto de treino X, constituído por n pares (x_i, y_i) para gerar um classificador único $\alpha \in H$. Os pontos localizados próximos ao hiperplano são denominados *Support Vectors* e a função de decisão que maximiza a separação dos dados é denominada de ótima.

Uma aprendizagem não supervisionada descobre as características explicitas dos dados, de acordo com a teoria de Dean [10]. Na sua abordagem, não existe necessidade de que um conjunto de dados seja pré-classificado, não existindo um processo de formação ajustado ao modelo de dados, em contrapartida à aprendizagem supervisionada, neste cenário não existe a necessidade de ter o conhecimento prévio das categorias e das classes.

O algoritmo *MinGen* de Sweeney [15] examina, exaustivamente, todas as potenciais generalizações de um domínio completo, para identificar a medida perfeita. M. D. Sweeney reconheceu que esta busca exaustiva é impraticável, mesmo para conjuntos de dados de tamanho modesto, motivando a segunda família de algoritmos de *k-anonimity* a ser discutida mais tarde. Samarati [16] propõe um algoritmo de pesquisa binária, que primeiramente identifica todas as generalizações mínimas, e depois encontra a generalização otimizada, e M. D. enumera todas as generalizações mínimas de uma operação dispendiosa e, portanto, não escalável para grandes conjuntos de dados.

LeFevre [5] apresenta um conjunto de algoritmos de generalização de baixo para cima, chamados de Incógnito, para gerar todas as generalizações possíveis de

domínio total do *k-anonimity*. Os algoritmos exploram a pré-prestação de *rollup* para calcular o tamanho dos grupos *qid*.

Outro algoritmo chamado *K-Optimize* [17] reduz eficazmente tabelas anónimas não otimizadas, modelando o espaço de pesquisa, utilizando uma árvore de enumeração definida. Cada nó representa uma solução *k-anonimy*. O algoritmo assume um conjunto totalmente ordenado de valores de atributos, e examina a árvore de cima para baixo a partir da tabela mais geral, reduzindo um nó na árvore quando nenhum dos seus descendentes poderia ser uma solução otimizada, baseada na métrica de discernibilidade DM, e na métrica de classificação CM. Ao contrário dos algoritmos acima referidos, o *K-Optimize* emprega a generalização da subárvore e esquemas de supressão de registos. É o único algoritmo ótimo eficiente, que utiliza a generalização flexível da subárvore. A segunda família de algoritmos produz uma tabela mínima de *k-anonimity,* por via do recurso a uma pesquisa abundante, dividida por métrica de pesquisa. Sendo de natureza heurística, estes algoritmos podem não encontrar a solução anónima ótima, mas são mais escaláveis do que a família anterior.

2.1.2. Privacidade dos dados

Data mining é tudo o que vemos, é importante respeitar as pessoas que nos fornecem os dados, ou que irão beneficiar-se do trabalho, e um dos primeiros aspetos a respeitar é a privacidade. Em 1953[4] entrou em vigor na Europa a Convenção Europeia dos Direitos do Homem, este documento contém o seguinte artigo que reforça a privacidade:

Os Estados-membros da União Europeia têm a obrigação de cumpri-la, em 25 de maio de 2018 entrou em vigor o Regulamento Geral da Proteção de Dados (RGPD)

[4] https://dgpj.justica.gov.pt/Relacoes-Internacionais/Organizacoes-e-redes-internacionais/Conselho-da-Europa/Tribunal-Europeu-dos-Direitos-Humanos - consultado a 4 de novembro de 2021

[18]. O regulamento consolida alguns direitos relacionados com a privacidade, como o de o utilizador poder solicitar que a sua informação seja apagada ou atualizada. Vários estados incluem direitos semelhantes nas suas constituições, pois foram aprovadas leis que garantem esses direitos.

Nos Estados Unidos, uma das leis mais relevantes é provavelmente o *HIPAA(The Health Insurance Portability and Accountability Act)* [19]. Trata-se de uma lei federal que exige a criação de padrões nacionais para proteger informações confidenciais de saúde do paciente, na Califórnia existe a *CCPA* [20] (*Califórnia Consumer Privacy Act*) e o *FERPA* [21] (*Family Educational Rights and Privacy Act*). Quando estivermos a trabalhar com um conjunto de dados provenientes de uma instituição de ensino, este é um dos cuidados que devemos ter em atenção, observar as premissas deste regulamento. Todos estes regulamentos existem para nos clarificar o que pode ou não ser extraído e utilizado por terceiros, com ou sem consentimento.

Backstrom [22] estuda o problema de ataques ativos que visam a divulgação de identidades e reidentificação de nós, com o pressuposto que o adversário tem a capacidade de modificar a liberação anterior da rede. A ideia geral é a de primeiro escolher um conjunto arbitrário de utilizadores-alvo de vítimas a partir de uma rede social, criar um pequeno número de novas contas de utilizador da rede social "sybil", vincular essas novas contas às vítimas-alvo e criar um padrão de ligações entre as novas contas, para que possam ser identificadas de forma única, no anonimato da estrutura da rede social. O ataque envolve a criação de O (logN) número de contas sybil, onde N é o número total de utilizadores na rede social. O problema da anonimização das redes sociais depende do conhecimento de fundo do adversário, e do modelo de ataque à privacidade, que pode ser a identificação de indivíduos na rede da relação entre indivíduos.

Li et al sugere que existem dois tipos de ataques à privacidade nas redes sociais. A seguir são apresentadas essas características, por ordem de frequência de citações, de acordo com o estudo desse autor:

- Divulgação da identidade - Numa rede social, trata-se de ligar os nós da rede anonimizada a indivíduos ou entidades do mundo real. Conceptualmente, o sistema pode escolher um valor de k, e certificar-se de que na rede social divulgada G', que pode ser diferente dos dados originais G', nenhum nó v em G' pode ser desidentificado com uma probabilidade ou graduação superior a 1/k, com base no pressuposto do conhecimento de fundo do adversário. A maioria dos mecanismos existentes tentariam assegurar que, pelo menos k nós diferentes são semelhantes em termos do conhecimento do adversário, sobre o alvo de ataque, de modo a que cada um dos k nós, tenha a mesma hipótese de ser o culpado.
- Divulgação de atributos - Numa rede social, cada nó representa um indivíduo ou um grupo de indivíduos. Pode haver atributos ligados a cada nó, tais como informação pessoal ou de grupo. Por vezes, essa informação pode ser considerada sensível ou privada a alguns indivíduos.

Numa rede social, um indivíduo pode ser identificado com base nos valores dos atributos, e também nas relações com outros indivíduos [23]. Os termos ataques “passivos” versus “ativos” são utilizados para diferenciar os casos em que o adversário apenas observa os dados e não os adultera, dos casos em que o adversário pode alterar os dados para fins de ataque. Hay el. [24] considera uma classe de consultas, de poder crescente, que relatam sobre a estrutura local do gráfico de um nó alvo. O autor também sugere que, para além de uma classe de consulta mais forte e mais realista, a consulta do subgrafo, afirma que a existência de um subgrafo em torno do nó alvo. O poder descritivo da consulta é contado pelo número de arestas no subgrafo.

Note-se que esta classe de conhecimento do adversário abrange também a 1-vizinhança, uma vez que a 1-vizinhança é um possível subgrafo em torno do nó alvo.

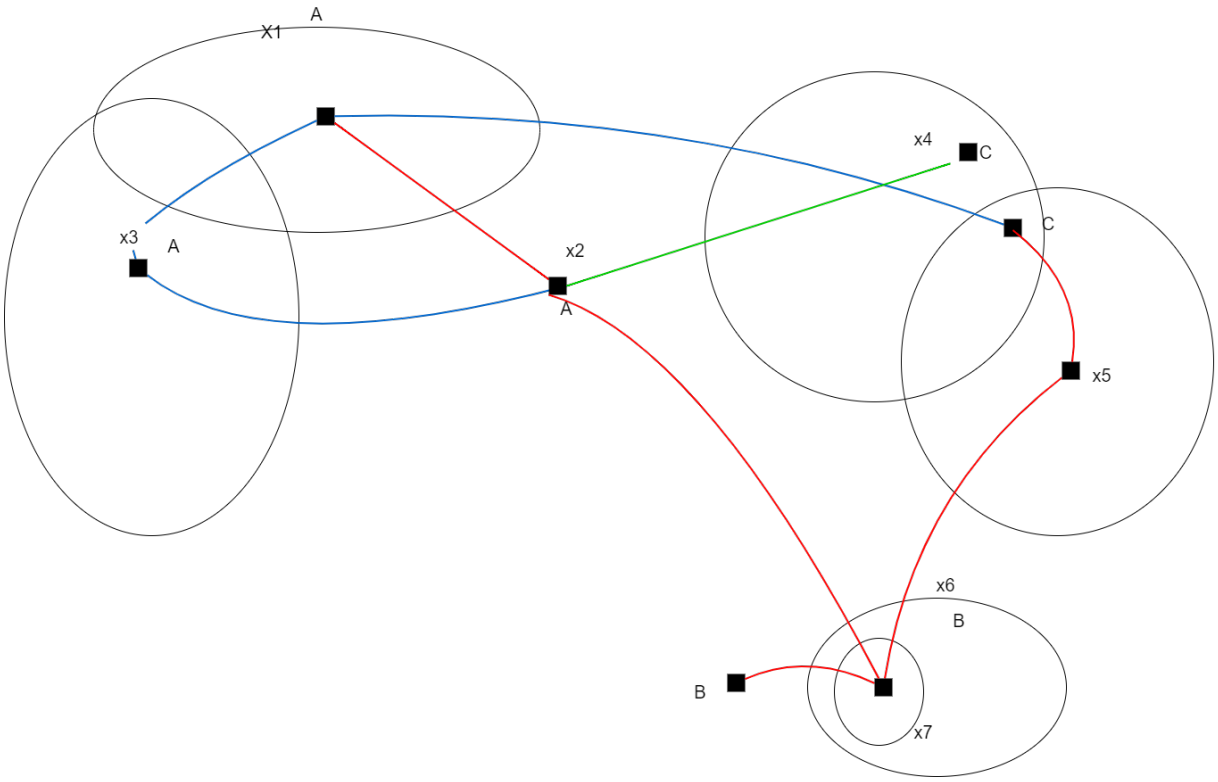

Figura 3 *- Esquema de anonimização numa rede social*[5]

Na figura 0.2, se o adversário souber que o indivíduo alvo está no centro de uma estrela com 4 vizinhos, então o adversário pode apontar o nó x2 na rede como alvo, uma vez que x2 é o único nó com 4 vizinhos.

No entanto, se o adversário souber apenas que o alvo é rotulado como A e tem um vizinho com etiqueta A, então os nós x1, x2 e x3 qualificam-se todos como o possível nó para o alvo, e o adversário não pode ter sucesso num ataque seguro.

[5] Fonte: Adaptado de *Hay privacy protetion in social network*

Para preservar a privacidade num processo de *data mining*, o outro lado da moeda é a utilidade dos dados publicados, a maioria das medidas de utilidade para as redes sociais estão relacionadas com as propriedades gráficas. Seguem-se alguns exemplos:

Propriedades de interesse em dados de rede: Hay et Al [25] usa as seguintes propriedades para a medida de desempenho.

- *Degree*: distribuição dos graus de todos os nós do gráfico;
- Caminho legal: distribuição dos comprimentos dos caminhos mais curtos, entre 500 pares de nós aleatoriamente removidos na rede;
- Transitividade: distribuições do tamanho do componente ligado, a que um nó pertence;
- Resiliência da rede: número de nós no maior componente ligado do gráfico à medida que os nós são removidos;
- Infecciosidade: proporção de nós infetados por uma hipotética *disase*, primeiro escolhe aleatoriamente um nó para a primeira infeção, e depois espalha a *disase* a uma taxa específica.

Korolova [26] considera as consultas da seguinte forma: para duas etiquetas l1, l2, qual é a distância média de um nó com etiqueta l1 ao seu nó mais próximo com etiqueta l2. Para evitar a violação, pode haver diferentes estratégias. O primeiro método, a abordagem mais amplamente adotada, é publicar uma rede anônima. Noutros casos, como o Facebook, o detentor dos dados pode optar por permitir que apenas os subgrafos sejam conhecidos por utilizadores individuais. Por exemplo, o sistema pode permitir que cada utilizador veja um nível dos vizinhos do outro utilizador.

Outra abordagem não é a deixar a rede exposta, mas de retornar uma resposta aproximada a qualquer consulta efetuada na rede.

Para preservar a privacidade, podemos liberar um conjunto de dados que se desvia do conjunto de dados original, mas que ainda contém informações úteis para diferentes usos dos dados. Alternativamente, um adversário iria encontrar ou criar k nós na rede, nomeadamente [x1... x2], e em seguida criar a aresta (xi, xj) independentemente, com probabilidade 1/2. Isso produz um gráfico aleatório H. O gráfico será incluído numa dada rede social, formando um gráfico G. O primeiro requisito de H é tal que não haja nenhum outro subgrafo S em G isomorfo a H (o que significa que H pode resultar de S, remarcando os nós). Desta forma, H pode ser identificado exclusivamente em G. O segundo requisito de H para si mesmo. Com H, o adversário pode ligar um nó de destino w a um subconjunto de nós N em H de modo que, uma vez que H seja localizado, w também estará localizado. A partir dos resultados da teoria de grafos aleatórios [34], a possibilidade de atingir os dois requisitos acima é muito alta, recorrendo ao método de geração de grafos aleatórios.

Felizmente, os ataques ativos não são práticos para a privacidade em grande escala, os ataques estão divididos a três limitações sugeridas por Narayan e Shmatikov [27]. Redes sociais online - Os ataques ativos são restritos às redes sociais online, apenas porque o adversário tem de criar um grande número de contas de utilizador falsas na rede social, antes de seu lançamento. Muitos provedores de redes sociais, como o Facebook, garantem que cada conta tenha pelo menos uma conta de e-mail associada, dificultando a criação de um grande número de contas falsas de utilizadores nessas redes sociais.

Yu et al [28] desenvolvem algumas técnicas, para identificar ataques de sibilas de redes sociais. Muitos provedores de redes sociais consideram que dois utilizadores estão ligados, apenas se ambos os utilizadores concordarem mutuamente em serem amigos, por exemplo, no Facebook, o utilizador A inicia uma amizade com o utilizador B. A amizade só é estabelecida se B confirmar a relação. Restrição de links

mútuos, as vítimas-alvo não concordarão em vincular-se de volta às contas falsas, pelo que os links não são estabelecidos e, portanto, não aparecem na rede libertada.

Uma forma de anonimizar uma rede social é modificar a rede, adicionando ou eliminando nós, adicionando ou eliminando o *edged* e alterando (generalizando) as etiquetas dos nós, em que a rede é libertada, apenas após terem sido feitas alterações suficientes para satisfazer certos critérios de privacidade.

Zhou e pei [29] consideram a 1-vizinhança de um nó, como um possível conhecimento de base do adversário. O objetivo da proteção é a reidentificação do nó. Aqui são permitidas alterações gráficas, por adição de borda e mudança de etiquetas no algoritmo. Há dois passos principais na anonimização, segundo o autor:

1. Encontrar o 1-vizinhança de cada nó da rede.

2. Agrupar os nós por semelhança e modificar as vizinhanças dos nós com o objetivo de ter pelo menos k nós com a mesma 1-vizinhança no gráfico anonimizado.

A segunda etapa do acima exposto, enfrenta o difícil problema do isomorfismo do gráfico, que é um número muito pequeno de problemas conhecidos em NP com incerteza em *NP-completeness*; ou seja, não existe nenhum algoritmo polinomial conhecido e bem como nenhuma prova conhecida de *NP-completeness*.

2.1.3. Datasets

Em quase todos os trabalhos existentes, as ideias ou métodos propostos foram testados nalguns conjuntos de dados. É importante mostrar que os métodos funcionam em alguns dados reais.

Em [30], são testados três conjuntos de dados reais:

1.*hep-th*: dados do arquivo *arXiv* e da base de dados *SPRIES-HEP* do Centro de Aceleradores Lineares de Stanford, que descreve artigos e autores em física teórica de alta energia.

2. *Enron*: introduzimos este conjunto de dados no início do capítulo.

3. *Net-trace*: este conjunto de dados é derivado de uma rede de rastreio de nível IP recolhida numa grande universidade, com 201 endereços internos de um único campus, e mais de 400 endereços externos.

Zhou e Pei [271] adotam um conjunto de dados da *KDD Cup 2003* sobre coautoria para um conjunto de artigos em física de alta energia. Nos gráficos de coautoria, cada nó é um autor e uma margem entre dois autores representa uma coautoria para um ou mais artigos. Há mais de 57.000 nós e 120.000 margens no gráfico, com um grau médio de cerca de 4, não há, de facto, qualquer problema de privacidade com o conjunto de dados acima referido, mas é um verdadeiro conjunto de dados de rede social.

Para além dos conjuntos de dados reais, foram utilizados dados sintéticos do modelo gráfico *R-MAT* que segue a lei do poder [31] sobre a distribuição dos nós e as características do pequeno mundo [32].

Microsoft Azure - A *Microsoft* é o segundo maior fornecedor líder de armazenamento em nuvem, a seguir à *Amazon*. A plataforma de nuvem e infraestrutura denominada *Azure* foi lançada pela primeira vez em fevereiro de 2010. Fornece modelos de entrega *PaaS* (Plataforma como Serviço) e *IaaS* (Serviço de infraestrutura) e oferece suporte a muitas estruturas e a várias linguagens de programação. Ao utilizar o *Azure*, o utilizador pode criar, implementar e gerir aplicações e serviços, por meio de uma rede global de *data centers* geridos pela

Microsoft. O *Azure* fornece ao utilizador a capacidade de utilizar vários tipos de instâncias com taxas por hora, tal como a *Amazon*. No entanto, estes calculam o custo dos recursos utilizados por minuto. O que significa que se um utilizador alocou um recurso por uma hora e meia, então o pagamento é calculado para o período exato de tempo sem estimar o teto para o período de tempo utilizado. No entanto, a *Microsoft* lida com os modelos de preços com sigilo, o que afeta a sua transparência para com os utilizadores. Consequentemente, afeta o número de clientes com credibilidade no serviço. [33]

Amazon Cloud - A *Amazon* é a maior líder no mercado de computação em nuvem com a sua *AWS* (*Amazon Web Services*). A AWS foi lançada em 2006, o que significa que está prestes a completar 15 anos de experiência neste campo. Além disso, a *Amazon* foi o primeiro provedor de serviços distribuídos em nuvem, a oferecer uma Infraestrutura como serviço (IaaS), permitindo que indivíduos e organizações aluguem máquinas virtuais. A *Amazon Elastic Compute Cloud* (EC2) é uma solução da *AWS* que permite aos utilizadores alocar os computadores necessários para executar os aplicativos de que necessitam. Ela permite a criação de máquinas virtuais, "instâncias" como a Amazon as designa, para obter uma implementação escalonável de aplicações que o utilizador requer. Essas instâncias podem ser criadas, iniciadas e encerradas por um utilizador, de acordo com a procura. O pagamento por instância ativa é calculado por hora, o que representa o significado do termo de elasticidade. [33]

O Fairlearn[6] é um pacote *Python* que pode ser utilizado para analisar modelos, e avaliar a disparidade entre previsões e desempenho de previsão, para uma ou mais características selecionadas. Funciona com o cálculo de um grupo de métricas para as características sensíveis ou especificas. As próprias métricas baseiam-se em métricas padrão de avaliação do modelo *scikit-learn*, tais como precisão ou *recall*. A API *Fairlearn* é extensa, oferece múltiplas formas de explorar a disparidade em métricas, através de agrupamentos de recursos específicos. Para um modelo de classificação binária, é comparado com uma taxa de seleção (número de previsões positivas para cada grupo) e utiliza-se a função *selection_rate*. Esta função devolve a taxa de seleção global, para o conjunto de dados de teste. Também pode utilizar funções padrão *sklearn.metrics* (como *accuracy_score*, *precision_score*, ou *recall_score*) para ter uma visão geral do desempenho do modelo. Em seguida, pode definir uma ou mais funcionalidades no seu conjunto de dados, com os quais pretende agrupar subconjuntos e comparar a taxa de seleção e o desempenho preditivo. A *Fairlearn* inclui uma função *MetricFrame,* que lhe permite criar um *dataframe* de várias métricas pelo grupo. Por exemplo, num modelo binário de classificação para previsão do reembolso do empréstimo, em que a característica sensível "*Age"* consiste em dois valores categóricos possíveis (25 e acima de 25). [34]

Quando se testa um modelo de mineração de dados, utilizando uma técnica supervisionada, como regressão ou classificação, utilizam-se métricas conseguidas contra dados de validação de retenção, para avaliar o desempenho preditivo geral do modelo. Por exemplo, podemos avaliar um modelo de classificação baseado na precisão ou *recall*. Para avaliar a equidade de um modelo, podemos aplicar a mesma métrica de desempenho preditivo aos subconjuntos dos dados, com base nas características sensíveis em que a amostra está agrupada, e medir a disparidade

[6] https://fairlearn.org/ consultado a 08 de dezembro de 2021

nessas métricas através dos subgrupos. Por exemplo, suponhamos que o modelo de aprovação de empréstimos apresenta uma métrica global de recolha de 0,67 - ou seja, identifica, corretamente, 67% dos casos em que o requerente reembolsou o empréstimo. A questão é se o modelo fornece, ou não, uma taxa semelhante de previsões corretas para diferentes faixas etárias. Para descobrir, agrupar os dados com base na característica sensível (Idade) e mede-se a métrica de desempenho preditivo para esses grupos. Então pode-se comparar as pontuações métricas para determinar a disparidade entre eles. [35]

A deteção de anomalias é um tipo proeminente de aplicativos de aprendizagem de máquina. Vários algoritmos de aprendizagem de máquina foram modificados, para poder distinguir entre instâncias de dados normais e anómalos. Embora a deteção de anomalias seja estudada há décadas, não existe um modelo de processo comum, que facilite o desenvolvimento desses sistemas. Muita da pesquisa é dedicada ao desenvolvimento de novos algoritmos ou ao aperfeiçoamento de algoritmos existentes. Além disso, existem muitas publicações que tratam da aplicação de um algoritmo de deteção de anomalias, num problema específico. Contudo, permanece uma lacuna entre esses dois extremos, a qual abordamos agora.

Embora existam obstáculos e tarefas individuais relacionados com projetos, existem, também, muitas semelhanças para cada problema de deteção de anomalias que se originam do domínio do aplicativo. Os requisitos para um sistema de deteção de anomalias são muito semelhantes, embora a definição de anomalia dependa da aplicação. Consequentemente, as mesmas perguntas devem ser respondidas para a maioria dos projetos de deteção de anomalias. Isso inclui a definição de anómalo e de normal em relação aos dados, a compensação necessária entre a sensibilidade e a

especificidade, a quantidade de dados que deve ser processada pelo sistema, em que momento e outros.

Normalmente, os algoritmos utilizados não dependem do domínio do aplicativo, mas da textura dos dados, bem como de outros requisitos, como o nível de interpretação do modelo que é necessário. Para a seleção dos algoritmos adequados, importa identificar o tipo de anomalia (anomalia pontual, contextual ou coletiva). Os casos de teste, geralmente podem ser derivados dos algoritmos aplicados. Como todos os tipos de instâncias de dados, são mapeados como anómalos ou normais, pode testar-se com essas duas classes em vez de examinar os próprios dados do aplicativo. Estando claro que, a origem de certos resultados de teste deve ser interpretada com recurso aos dados originais.

Copyright, não podemos utilizar fotos, ou vídeos que encontramos online porque podem ter direitos de autor, muitos conteúdos digitais têm licenças que podem ser livres ou pagas, não podemos assumir que, porque está na Internet, pode ser utilizado gratuitamente, temos de consultar a fonte, ou consultar a licença para sabermos se o que estamos a fazer é permitido.

Hoje é indiscutível que, a cada um, assiste o direito à privacidade sendo esta protegida a diferentes níveis. A declaração universal dos direitos humanos reza:

Artigo 8 [36] - Qualquer pessoa tem direito ao respeito da sua vida privada e familiar, do seu domicílio e da sua correspondência.

Artigo 12 [37] - Ninguém sofrerá intromissões arbitrárias na sua vida privada, na sua família, no seu domicílio ou na sua correspondência, nem ataques à sua honra e reputação. Contra tais intromissões ou ataques, toda a pessoa tem direito à proteção da lei.

Ataques *em Data Mining* - Existem vários casos mediáticos, de violação da privacidade: alguns deles são devidos a uma má compreensão dos fundamentos da privacidade de dados, e à falta do uso de tecnologias adequadas; outros, a maioria, devem-se a violações de segurança, por exemplo, dados encriptados esquecidos num transporte publico.

Sobejamente conhecido, foi o caso de violação da privacidade de dados cometida pelo Facebook. Em 2013, pesquisadores [38] do Centro de Psicometria da Universidade de Cambridge analisaram resultados, que foram obtidos através de um teste de personalidade, realizado por voluntários, disponibilizado no Facebook, para avaliar o seu perfil psicológico. Este teste, denominado “OCEAN”, recolhia indicadores comportamentais de um determinado individuo e correlacionava-os com as suas preferências no Facebook (gostos e partilhas).

Esta pesquisa atraiu 350.000 participantes norte-americanos e, por meio dela, conseguiu-se ter uma perceção clara, da relação entre as atividades do Facebook e outros indicadores online. Este trabalho permitiu criar o perfil *“OCEAN”,* útil para avaliar qualquer indivíduo, já que pode ser reproduzido com uma precisão razoável usando essas métricas e sem recurso a um instrumento psicográfico formal. Todavia, não há confirmação de que essa pesquisa tenha exposto os utilizadores do Facebook e a sua rede de amigos a um abuso intencional da sua de privacidade. Existem, sim, indicações, que a universidade recusou compartilhar dados individuais ou critérios resultantes, com a *Cambridge Analytica*.

Após este teste foi desenvolvido [39] um segundo projeto de pesquisa, executado agora pela *Global Science Research (GSR),* em cooperação com a *Cambridge Analytica,* com o intuito de identificar os parâmetros necessários para desenvolver os perfis *“OCEAN”*, criando para esse efeito um questionário de personalidade na *Amazon's Plataforma Mechanical Turk* e *Qual-trics*, uma plataforma de pesquisa.

O questionário exigia que os utilizadores autorizassem o acesso ao seu perfil no Facebook, partilhando, em simultâneo, os dados dos seus amigos, através da API aberta do Facebook, até maio de 2015.

Foi assim que a *Cambridge Analytica* acedeu aos dados do Facebook, não sendo necessário extrair os dados individuais específicos, para concluir o objetivo principal da pesquisa - que era o de estabelecer uma metodologia, para obter um perfil psicográfico dos indivíduos, com base nas redes sociais e noutros indicadores.

A Cambridge Analytica percebeu que poderia acumular essas informações com as das outras redes sociais, plataformas de *média*, navegadores, compras online, resultados de votação, etc., para ter na sua posse “mais de 5.000 dados individuais de 230 milhões de adultos nos EUA.”

A *Cambridge Analytica* desenvolveu, assim, a capacidade de ter um “micro alvo” de consumidores ou eleitores, com o perfil OCEAN, utilizando mensagens personalizadas, para poder influenciar comportamentos aquando da tomada de decisões.

O *know-how* adquirido pelo teste *OCEAN* foi combinado com um grande número de mensagens do “*Projeto Alamo*”, utilizado para a campanha eleitoral do presidente Trump. Algumas dessas mensagens foram criadas para a campanha de Trump e, outras, simplesmente alavancaram as “notícias” disponíveis na Internet (talvez incluindo conteúdo financiado pela Rússia para condicionar as eleições nos Estados Unidos). Conforme descrito pelo CEO da *Cambridge Analytica*: a chave era identificar aqueles que poderiam ser atraídos a votar no seu cliente, ou desencorajá-los de votar no seu oponente.

Não ter uma conta no Facebook não é garantia de proteção de dados, pois a aquisição destes não se limita ao Facebook e a análise pode ser facilmente aplicada a outras manifestações pessoais de preferências. Além de que todos os sites com o

logotipo do Facebook estão vinculados ao Facebook conduzindo ao rastreamento de não registados. Existem muitas fontes semelhantes de rastreamento online - por exemplo, *web beacons* - a maioria dos quais está ligada a *cookies* que podem ser utilizados em sites, cujo acesso pode ser vendido a compradores interessados. Além disso, ao combinar notícias reais com desinformação ou conteúdo irrestrito da Internet, os eleitores-alvo encontrarão mensagens de reforço em muitos sites sem perceber que são algumas das poucas pessoas no mundo a receber estas mensagens.

A ideia de que a análise psicográfica pode ter um impacto significativo no comportamento tem sido questionada. Todavia, Stanford Michal Kosinski e os seus colegas, os quais fizeram parte da equipa de pesquisa da Cambridge University em 2013, acreditam num impacto significativo, com prova numa base de amostra de 3,5 milhões de utilizadores.

Anonimato e não vinculação o anonimato de um sujeito, significa que o mesmo não é identificável dentro de um conjunto de dados, isto é, não será distinguível dos outros. A não vinculação, na perspetiva de um atacante de sistema, garante que não se conseguirá distinguir o suficiente a pertença, ou não, a um conjunto de dados. Outro conceito importante de privacidade é o de partilha e é utilizado, principalmente, no controle de divulgação de dados, dentro da comunidade de mineração de dados, para perseverar a privacidade. A divulgação de dados pode acontecer [40] sem identificação, caso sejam inferidos a partir dos dados anónimos, de todas as unidades populacionais que possuem um conjunto de atributos -X 1, ..., X n – ou o atributo X n +1 (divulgação de atributo positivo) ou, se toda a população que possui um conjunto de atributos (X 1, ..., X n) não possuir o atributo X n +1 (revelação de atributo negativa). Ora, se um invasor possuir um registo de identificação incluindo os atributos (X 1, ..., X n), então X n +1 pode ser atribuído ou distribuído ao registo de identificação.

Por outro lado, se um intruso já conhece todos os atributos de um registro (ou célula de uma tabela), a identificação pode ocorrer sem atribuição. Essas duas situações encapsulam o problema do risco de divulgação, para dados tabulares agregados, onde há mais preocupação com a atribuição. Esses dados são frequentemente divulgados como dados populacionais completos, tornando fácil a verificação das atribuições.

Pseudônimos e identidade - A pseudonímia é definida como a substituição do nome [40] ou de outros identificadores por um número, a fim de tornar a identificação do titular dos dados impossível, ou substancialmente mais difícil. O titular de um pseudónimo corresponde ao assunto a que o pseudónimo se refere. Vários pseudónimos podem corresponder ao mesmo assunto.

Pseudónimos numa rede de comunicação - Nalgumas circunstâncias, os pseudónimos permitem cobrir o intervalo entre anonimato (sem rastreabilidade) e a responsabilidade. Um individuo pode ter um conjunto de pseudónimos para evitar a rastreabilidade *[40]*. No entanto, em alguns sistemas é fácil estabelecer ligações entre um pseudónimo e o seu titular.

Transparência - Essa abordagem é semelhante ao *Princípio de Kerckhoffs* [40] em criptografia, que afirma que um cripto-sistema deve ser seguro, ainda que tudo sobre o sistema seja do conhecimento publico, exceto a chave. Quando os dados são partilhados, é boa-prática fornecer detalhes sobre a recolha desses dados. Isso, naturalmente, inclui uma descrição de qualquer processo de proteção de dados aplicado, bem como os parâmetros e os métodos A. Kerckhoffs,1883.

A transparência tem um efeito positivo na partilha de dados, se os pesquisadores souberem como os dados foram protegidos, eles podem levar essas informações em consideração na análise dos dados.

O público deve ser capaz de identificar os tipos de dados que estão a ser recolhidos, por meio de qualquer site ou outro meio eletrônico, quais os dados que são retidos, como serão utilizados, bem como o que é compartilhado com terceiros (direta ou indiretamente). Essas mesmas informações devem ser disponibilizadas por esses terceiros.

Todos os mecanismos de recolha de dados devem ser divulgados aos utilizadores, incluindo *web beacons* ou outros mecanismos, usados para rastrear a atividade ou dados do utilizador. Essas informações devem ser suficientes e claras para que os utilizadores possam identificar e procurar a divulgação e os controlos que esses coletores de dados fazem.

Cada site ou aplicativo deve divulgar qualquer conteúdo colocado no dispositivo dos utilizadores, bem como a utilização desse conteúdo.

3. METODOLOGIA

Neste capítulo faz-se uma descrição do caso em estudo, e define-se o planeamento, tendo em consideração as ferramentas de mineração de dados, de forma a dar resposta à problemática levantada. Este trabalho foi realizado para garantir a escolha de uma técnica ideal e coerente de proteção da privacidade dos dados.

Existem diferentes denominações para a conceção metodológica, alguns autores consideram tratar-se de um cânone explicativo ou de um esquema esclarecedor. Segundo Demo [6] é necessário que o pesquisador explicite qual a metodologia a utilizar.

Martins [41] argumentou que existem três abordagens de pesquisa: quantitativa, qualitativa e a combinação de ambas, denominada de método misto, que possibilita uma melhor compressão. Porém, esta abordagem não é muito precisa, a tendência para destacar uma única abordagem é maior, tirar conclusões entre duas metodologias é extremamente desafiante, Bryman [42]

Numa abordagem quantitativa as variáveis de pesquisa são previamente definidas. É classificada como uma abordagem mais objetiva, a analise das variáveis, é frequentemente feita com recurso a métodos estatísticos. Martins [43] também definiu que, os métodos para uma pesquisa quantitativa são: *survey*, modelagem ou simulação, experimentação. O estado da arte é necessário para ambas as abordagens, algumas variáveis não são possíveis de quantificar, o que não causa distinção, pois existem alguns pesquisadores que quantificam variáveis, uma das características que permite diferenciá-las, é que a qualitativa considera a perspetiva das pessoas, e, portanto, trabalha com a realidade subjetiva da amostra.

Este trabalho tem uma abordagem quantitativa, pois para atingir o objetivo desta dissertação, decidiu-se que a plataforma escolhida para a implementação da solução de mineração de dados, num sistema distribuído, fosse a *Microsoft Azure* (ver a secção 2.1) [44], com o suporte de um processamento paralelo. Foi utilizada uma base de dados de um censo populacional, para efeitos de teste deste trabalho. Este capítulo descreve o processo de mineração com a metodologia escolhida, e que foi dividida nas seguintes fases:

- Análise de dados;
- Processo de mineração de dados;
- Implementação de um processo de mineração.

3.1. Análise dos dados

No sector da estatística, graças à iniciativa de recolha de dados de diversas fontes, é possível obter mais conhecimentos acerca das condições sociais dos habitantes de uma determinada zona geográfica [40]. O conjunto de dados utilizados foi fornecido pela plataforma UCI[7], que contém informações relativas ao censo populacional. A extração foi feita por Barry Becker do conjunto de dados do Censo de 1994. Neste, estão contidos dados de 48.842 registos de adultos, provenientes de habitantes dos EUA, com diferentes nacionalidades. Podemos encontrar informações demográficas, da situação social, das habilitações literárias, bem como informações acerca da composição do agregado familiar.

Neste capítulo dá-se uma descrição do repositório dos dados e dos conjuntos que irão ser sujeitos a análise. O conjunto de registos encontra-se razoavelmente

[7] Informação relativa à data de dezembro de 2021

limpo e foi extraído utilizando as seguintes condições: ((AGE> 16) && (AGE> 100) && (AFNLWGT> 1) && (HRSWK> 0)).

No que concerne ao conceito da privacidade, em relação a extração de dados a partir de modelos de *machine learning*, é necessário estabelecer o que realmente é "privado".

Suponhamos um modelo de classificação para a aprovação de um processo de candidaturas a uma vaga de emprego, que permitirá, portanto, prever a probabilidade de uma contratação. O modelo deverá utilizar características, tais como as seguintes:

1. **Idade**
2. **Habilitações literárias**
3. **Situação de empregabilidade**
4. **Número de dependentes**
5. **Estado civil**

Interrogações - Com estas características, é possível criar um modelo de classificação, e determinar as probabilidades de uma possível integração nos quadros da empresa. Portanto, este trabalho procura apurar as competências para uma determinada vaga, por via de um modelo de classificação.

Suponhamos que o modelo prevê que, tendo em conta a situação económica mundial, que 30% dos candidatos não tem competências para a vaga. No entanto, começa-se a suspeitar que, maioritariamente, são aprovadas para entrevista pessoas do sexo masculino, com mais de 30 anos que trabalham e que não têm experiência. Como ter a certeza que este modelo preserva a privacidade de todos os candidatos?

Uma das formas de avaliar a privacidade de um conjunto de dados, consiste em comparar os resultados para cada grupo, através dos dados sensíveis. Para o

modelo de aprovação de candidaturas, o sexo e a raça são característicos sensíveis, para que seja possível dividir os dados em subconjuntos por cada faixa etária e comparar a taxa de seleção por cada grupo [45].

***Tabela 2** - Caraterísticas dos dados demográficos sensíveis*

	Característica	**Tipo**	**Relevante**
1	*General*	*Binário*	Sim
2	*White*	*Binário*	Sim
3	*Black*	*Binário*	Sim
4	*Asian-Pac-Island*	*Binário*	Sim
5	*Amer-Indian-Eski*	*Binário*	Sim
6	*Other*	*Other*	Sim
7	*Sex*	*String*	Sim

Cada uma das colunas correspondentes à raça possui um valor binário e é exclusiva, ou seja, não é possível ter um requerente com duas raças em simultâneo, este conjunto contém 14.653 instâncias e 2 características sensíveis.

Digamos que o modelo prevê que 56% dos candidatos, com idade igual ou superior a 30 anos irão ser aprovados, mas prevê enquadramentos bem-sucedidos para 36% dos candidatos com mais de 30 anos, logo haverá uma disparidade de 18% nas previsões.

3.2. Processo de mineração de dados

No processo de mineração de dados, seguiu-se a especificação *CRISP-DM* (ver secção 1.2), O objetivo desta dissertação consiste em testar um modelo, com suporte de técnicas de privacidade, e não existe qualquer intenção de fazer o *deployment* da solução num ambiente real de produção. Portanto, considera-se que esta documentação corresponde a esta fase.

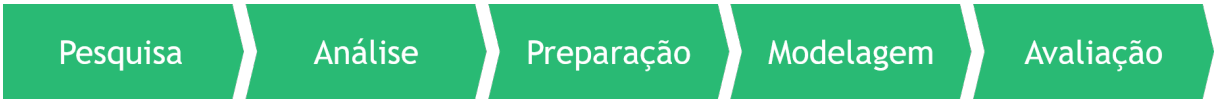

Figura 4 *- Esquema geral do processo de mineração de dados*

Na figura 4 observa-se o diagrama do processo de mineração adotado para a realização do trabalho, que foi totalmente baseado na metodologia de *CRISP-DM* (ver secção 1.2). Este foi dividido em 5 etapas principais (Pesquisa e Estudo, Análise exploratória, Preparação dos dados, Modelagem de Dados e Avaliação).

Pesquisa e Estudo - Esta etapa, que corresponde ao *Business Understanding* representa o ponto de partida da análise dos dados, é o estudo do problema em questão. Foram seguidos os seguintes passos:

1. Entender o conjunto de dados e as características técnicas da área de estudo da estatística;

2. Consultar a documentação disponível no repositório da UCI e descrição da mesma;
3. Pesquisa de artigos sobre estudos relacionados já realizados com o mesmo repositório de dados;

Análise exploratória - Foram seguidas as seguintes fases**:**

1. Entender as características sensíveis disponíveis no conjunto de dados do repositório, e a forma como se relacionam com os outros componentes,
2. Questionar com o objetivo de entender se as características podem ser agrupadas.

Tabela 3 *- Contagem de instâncias por raça*

Raças	**Contagem**
White	84%
Black	91.6%
Asian-Pac-Island	81.1%
Amer-Indian-Eski	91%
Other	85.6%

Durante esta fase teceram-se as seguintes considerações: no conjunto de dados do censo considerou-se que as variáveis de entrada são uma mistura de tipos de dados numéricos, categóricos e ordinais, onde as colunas não numéricas são representadas por *strings*. No mínimo, as variáveis categóricas precisarão ser codificadas ordinal ou *one-hot*.

Também podemos constatar que a variável principal é representada por *strings*. Essa coluna precisará de ser codificada por rótulo, com 0 para a classe maioritária e 1 para a classe minoritária, como é habitual para tarefas de classificação binária desequilibrada.

Os valores ausentes são marcados com um '?'. Esses valores precisarão de ser imputados ou, dado o pequeno número no conjunto, essas linhas podem ser excluídas do conjunto de dados.

Preparação dos dados - No pré-processamento de dados foi feita uma seleção das características sensíveis. O algoritmo escolhido foi SVM porque é o mais adequado quando é necessário segregar duas classes principais, Após a análise preliminar do repositório optou-se pela criação de uma coluna, que agrupa as raças: *white, black, asian-pac-island, amer-indian, other*. Nesta, o conjunto de dados, antes de ser modelado é repartido num conjunto de Treino de (70%) e Testes (30%).Com base nos testes efetuados verificou-se em primeiro lugar que ao fazer o *split* com esta percentagem, trouxe melhorias significativas aos modelos de teste. Também se verificou com utilização desta técnica observou-se a normalização dos

dados.Empregou-se outras taxas de treino é de teste diferentes(5%,10%,15%,20% e 25%)

Modelação dos Dados - Os dados foram preparados para serem inseridos nos algoritmos de mineração escolhidos. As características selecionadas foram raças e sexo, o *Fairlearn* integra-se com o *Azure Machine Learning*; permitiu-se executar uma experiência na qual as métricas do *dashboard* são carregadas para o ambiente de trabalho *Azure Machine Learning*. Isto permite a partilha do *dashboard*. Neste processo utilizaram-se os seguintes:

- Classificador *Support vector machine* (SVM) – O classificador permite resolver problemas de regressão, permite que apenas um subconjunto dos dados de preparação. Este ignora as outras amostras dentro do conjunto.

Avaliação - Realização de testes de desempenho e avaliação das métricas de qualidade dos resultados obtidos (ver Capítulo 4).

3.3. Implementação de processo de mineração de dados

A implementação do processo de mineração foi desenhada para tirar partido do paralelismo durante as fases de preparação e modelação dos dados. Entretanto, isso não significa que todos os componentes da arquitetura podem ser paralelizáveis, pois existem determinadas tarefas que foram desempenhadas de forma distinta. O processo foi implementado de uma maneira genérica, para ambos os ambientes.

Para o conjunto de dados, usou-se o conhecido conjunto de dados do censo de adultos, que foi carregado no *OpenML*.

Neste projeto, escolheu-se um dos principais fornecedores de serviços em nuvem; o *Microsoft Azure*. Foram extraídos dois conjuntos de dados. O conjunto de dados é destinado ao *Microsoft Azure*.

3.3.1. Módulo de mineração em *Azure* com *Fairlearn*

O modulo foi implementado no ambiente de mineração de dados do *Azure* com a biblioteca do *Fairlearn* (um pacote de avaliação de equidade e mitigação de injustiças *open source*) para um problema de classificação binária. O pressuposto é o de que as previsões modelo são utilizadas, para decidir se um indivíduo deve, ou não, ser contratado. A sua finalidade é puramente ilustrativa de um fluxo de trabalho que inclui um painel de controlo da equidade - O algoritmo de *SearchGrid* do pacote *Fairlearn* utiliza uma função específica de equidade chamada Paridade Demográfica. Isso produz um conjunto de modelos, possível de ser visualizado num *dashboard* no ambiente do *Azure*. Através da extração deste conhecimento, é possível a criação de várias estratégias de recrutamento e seleção como:

- Seleção de um perfil adequado para a vaga,
- Criação de uma proposta salarial adequada.

Pré-processamento - Esta fase é comum a todos os modelos. Trata-se do pré-processamento inicial dos dados. (ver listagem 3.1). Efetuaram-se as seguintes etapas:

1. Extração dos dados através do *openml* do *Azure,* que contém toda a informação relevante para a formação das amostras;
2. Seleção e substituição das características relevantes para a pesquisa;
3. Dividir o conjunto de dados por raças e género;
4. Separar o conjunto em treino e teste.

```python
# Carregar o conjunto de dados do Censo
dados = fetch_openml(data_id=1590, as_frame=True)
X_bruto = dados.data
y_transformados = (dados.target == ">50K") * 1

#Separaçao dos dados sensiveis como o "sexo" e "raça" dos dados principal
X_bruto = dados.data
y_transformados = (dados.target == ">50K") * 1
A = X_brutos[["raca", "sexo"]]
X = X_brutos.drop(labels=['sexo', 'raca'],axis = 1)

# Separar o conjunto de dados em dois "treino" e "teste"
(X_treino, X_teste, y_treino, y_teste, A_treino, A_teste) = conjunto_treino(
    X_bruto, y_transformados, A, test_size=0.3, random_state=12345, stratify=y
)
```

Figura 5 *– Pré-processamento em Python :Carregar,adicionar,juntar e particionar em paralelo*

Processamento - Foram aplicados os classificadores SVM, para cada ambiente distinto, nesta fase efetuou-se o tratamento dos valores em vazio.

```python
# Definir o processamento do  pipeline Isto acontece depois de separar os dados para evitar fuga de dados
numeric_transformer = Pipeline(
    steps=[
        ("impute", SimpleImputer()),
        ("scaler", StandardScaler()),
    ]
)
categorical_transformer = Pipeline(
    [
        ("impute", SimpleImputer(strategy="most_frequent")),
        ("ohe", OneHotEncoder(handle_unknown="ignore")),
    ]
)
preprocessor = ColumnTransformer(
    transformers=[
        ("num", numeric_transformer, selector(dtype_exclude="category")),
        ("cat", categorical_transformer, selector(dtype_include="category")),
    ]
)
```

Figura 6 *- Implementação do algoritmo SVM*

4. RESULTADOS E DISCUSSÃO

Para a validação da presente dissertação foi necessário proceder a diversas etapas de caracter prático e, para tal, foram utilizadas ferramentas com um nível tecnológico capaz de cumprir todas as tarefas.

Considerando que o pretendido era a aplicação de técnicas de proteção de dados de *Data Mining,* consideraram-se várias hipóteses: desde aplicações *opensource*, aplicações já conhecidas e a exploração de novas ferramentas.

Neste trabalho apresenta-se o conjunto das métricas obtidas através da realização de testes de desempenho, neste caso duas classes de teste *CPU time* e tempo decorrido de processamento, onde é possível balancear as vantagens de utilizar uma técnica em detrimento da outra.

O capítulo está dividido em duas secções que passo a descrever:

1. Teste dos modelos de mineração - Testes para avaliar a qualidade dos modelos de mineração.
2. Teste de desempenho – Pretende conceder uma visão sobre qual o ponto de saturação da plataforma na qual foi executada.

4.1. Testes de modelo de mineração

Os resultados obtidos da mineração dos dados foram avaliados, com base nos seguintes parâmetros [46]:

1. *Precision* – Positivos verdadeiros a dividir pelo número de falsos positivos e positivos verdadeiros;

2. *Recall* – Total de positivos verdadeiros a dividir pelo total de positivos verdadeiros, somados ao total de falsos negativos;

3. *F1-measure* – Média harmónica entre Precisão e *Recall;*

Estes parâmetros foram escolhidos com base na credibilidade dos mesmos mediante a comunidade. Os testes foram realizados utilizando a *API Fairlearn* do *phyton.*

Diferença da taxa de falsos negativos, intervalos entre 28.3% e 42.5%. O modelo com o melhor desempenho consegue uma precisão de 85.3% e uma diferença na taxa de falsos negativos de 28.3%. O modelo com o maior valor de métrica da integridade alcança a precisão.

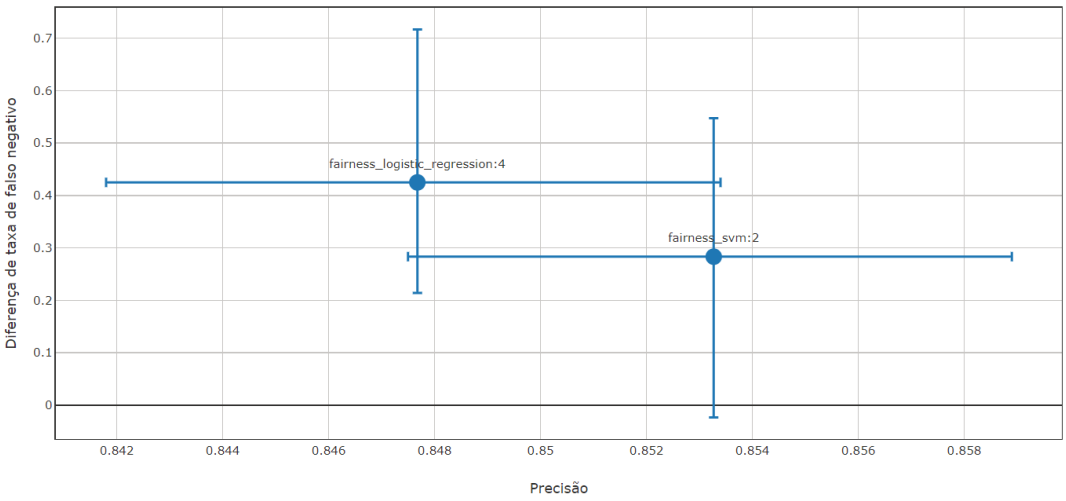

***Figura 7**- Diferença de falsos negativos*

A precisão varia de 84.8% para 85.3%. A diferença da taxa de falsos positivos em intervalos entre 7.22% e 7.94%. O modelo com o melhor desempenho consegue uma precisão de 85.3% e uma diferença na taxa de falsos positivos de 7.22%. O modelo com o maior valor de métrica da integridade alcança uma precisão de 85.3% e diferença na taxa de falsos positivos de 7.22%.

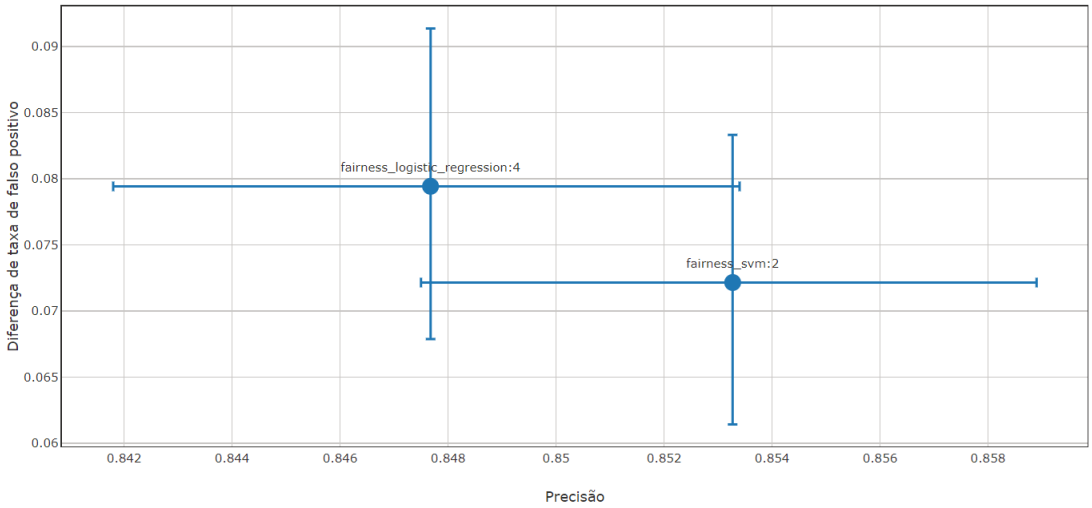

***Figura 8** - Diferença de falsos positivos*

	Precisão	Taxa de seleção	Diferença de taxa de falso positivo	Taxa de falso positivo	Taxa de falso...
Em geral	85.3%	18.9%	7.22%	6.32%	41.2%
Male	81.6%	24.9%		9.13%	39.3%
Female	92.8%	6.77%		1.91%	52.1%

Figura 9 *- Taxa de desempenho com a métrica de precisão*

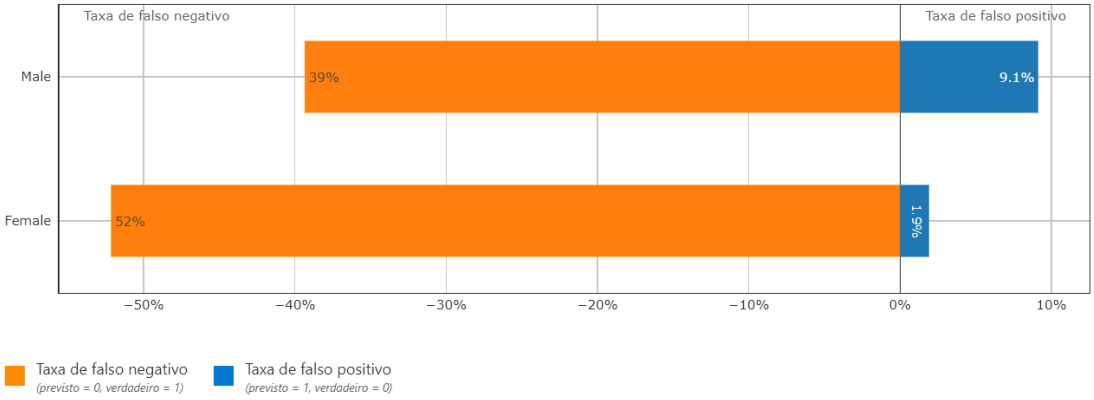

Figura 10 *-Taxa de desempenho com Recall*

	Precisão	Taxa de seleção	Proporção de taxa de...	Taxa de fals...	Taxa de f...
Em geral	85.3%	18.9%	54.1%	6.32%	41.2%
White	84.6%	20.1%		6.73%	40.6%
Black	91.9%	7.73%		2.44%	51.8%
Asian-Pac-Island	80.5%	25.7%		12.8%	38.3%
Amer-Indian-Eski	93.8%	7.59%		0%	45%
Other	87.4%	8.11%		4.17%	66.7%

Figura 11 *-Taxa de desempenho com F1-meseaure*

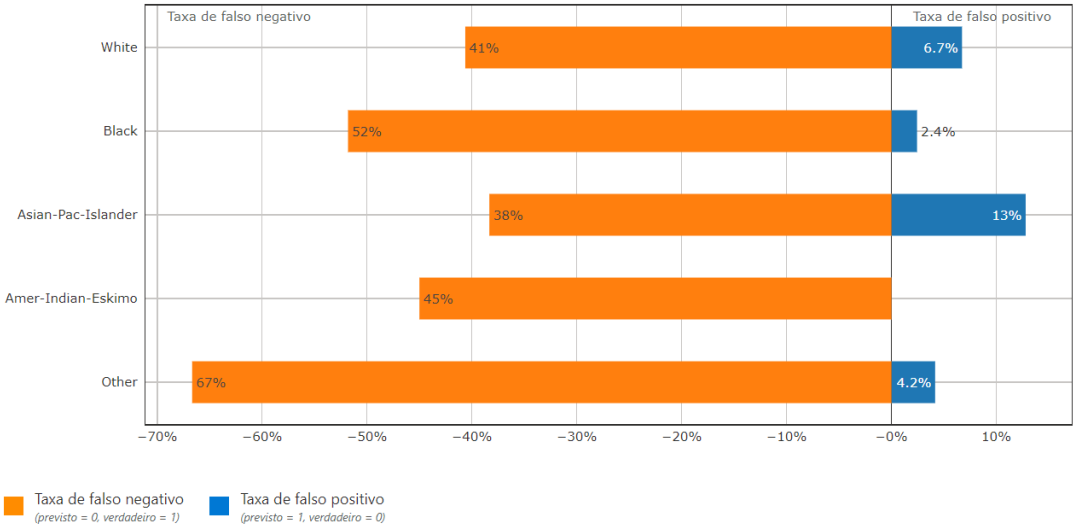

Figura 12 *- Taxa de desempenho do atributo raça*

	Pontuação F1	Taxa de seleção	Proporção de taxa...	Taxa de falso positivo	Taxa de falso negativo
Em geral	0.657	18.9%	0.789	6.32%	41.2%
Male	0.669	24.9%		9.13%	39.3%
Female	0.583	6.77%		1.91%	52.1%

Figura 13 *- Taxa de desempenho do atributo sexo com os algoritmos de métrica de desempenho*

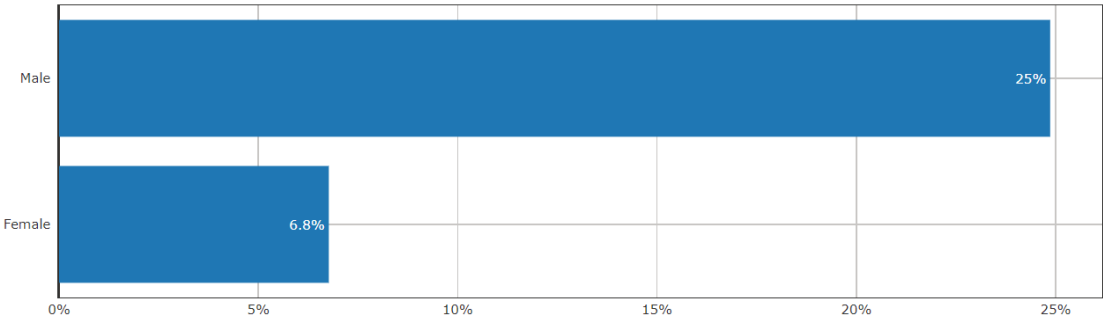

Figura 14 *- Taxa de desempenho do atributo sexo*

4.2. Testes de desempenho

Executámos os testes num PC com 6 GB de RAM, 4 núcleos Intel, i7 (2,0 GHz cada). Para os nossos testes, utilizámos o *Azure Machine Learning*, para implementar a metodologia proposta. Ambos os conjuntos de dados utilizados consistem em 14.568 registos que representam os registos dos questionários realizados no censo. Depois de aplicar os dois modelos de regressão explicados na seção anterior, os tempos de *CPU* foram obtidos a partir da média aritmética das várias interações feitas. Os parâmetros utilizados foram os seguintes:

- Tempo decorrido – Tempo decorrido durante todo o processamento.
- CPU time – É o tempo gasto pelo *Core*, o mesmo é medido em segundos, e excluindo assim os tempos de espera em operações de I/O. Este tempo é crucial para estimar gastos na plataforma de *cloud*, a cobrança é efetuada com base neste parâmetro.

Tendo em vista a realização de um teste de escalabilidade, que verifique qual o algoritmo que apresenta maior desempenho, tal como a figura 0.6 apresenta. Os dados foram replicados, mas a sua dimensionalidade foi mantida. Com esta técnica é possível ter uma análise da escalabilidade do algoritmo.

Implementação dos Testes - Os testes foram implementados em *Azure Machine Learning* e em Phyton seguindo os seguintes passos:

1. Criação de um fluxo de processamento para cada algoritmo,
2. Gravação dos tempos obtidos no teste.

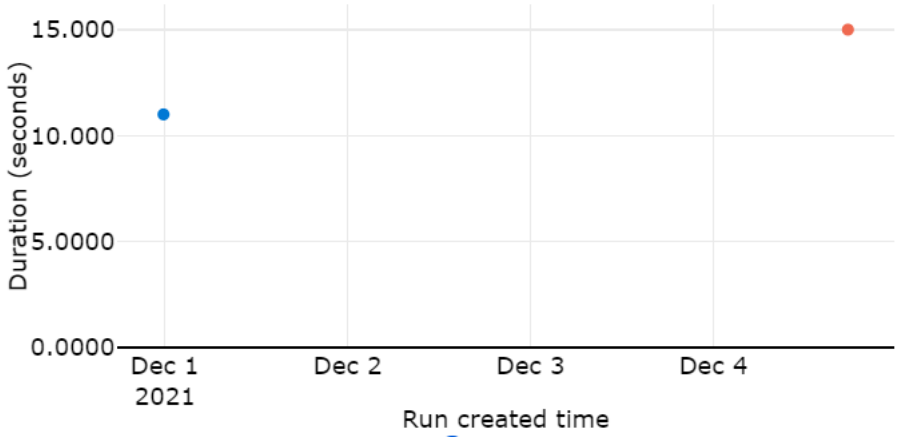

Figura 15 *– Performance da máquina virtual*

5. CONCLUSÕES

Este capítulo faz uma síntese de como o trabalho foi exposto, e é como uma reflexão sucinta sobre os resultados obtidos e as futuras contribuições desta dissertação. É proposta uma descrição de trabalhos futuros.

5.1. Síntese de discussão de resultados

A análise de dados é um dos campos de estudo e implementação mais atrativos para vários tipos de organizações e prestadores de serviços. Após a revisão de muitas aplicações, conforme mencionado na seção de trabalho relacionado, tivemos a ideia de relacionar esse campo interessante e de criar um teste aos fornecedores de computação em nuvem. E isso ocorre porque o ramo da pesquisa tem focado toda a atenção para a computação em nuvem, nos últimos anos. Além disso, este estudo foi realizado para aplicar a técnica de preservação da privacidade num dos principais fornecedores de computação em nuvem, para identificar se existe equidade da informação nesta plataforma.

Como os resultados demonstraram, a polaridade positiva no Azure era maior com a aplicação do *Fairlearn*, enquanto a polaridade negativa sem a aplicação era superior. Tal significa que, por exemplo, uma pessoa que se candidata a um anúncio de emprego, com uma faixa etária acima dos 30 anos de idade, tem a classificação de 83%. Quanto à classificação sem a aplicação do *Fairlearn*, a categoria “desconhecido” era exatamente a mesma que a da *API* com 79%. Enquanto a categoria “*white*” foi de 8% com a aplicação da técnica sem a biblioteca e 10% para a *Azure*, o que representa uma pequena diferença. Outras categorias diferem em percentuais ligeiros, que podem ser considerados insignificantes.

A informação analisada encontra-se disponível num repositório aberto, criado no âmbito de um censo populacional. Os dados foram previamente analisados e criou-se um processo de mineração, que permite apurar a equidade em processos de recrutamento, para responder a interrogação atual foram criados modelos com a utilização do algoritmo SVM *Support Vector Machine*. Os testes de desempenho procuraram avaliar quando é que se torna mais vantajoso utilizar uma *API* externa. Para efetuar essa avaliação, foram utilizados dois tipos de métricas: o tempo decorrido e o CPU *Time.* Foi preservada a dimensionalidade dos dados.

Os testes demostram que a *API* é mais vantajosa, em relação à utilização da regressão simples. A análise foi desafiante, pois o desempenho de um determinado algoritmo varia mediante o conjunto de dados. Contudo, será justo afirmar que para um conjunto de dados similar, os resultados podem ser idênticos, independentemente das técnicas utilizadas.

5.2. Principais contribuições

O trabalho contribui para uma melhor compreensão da temática da mineração de dados em plataformas *cloud*, transmitindo uma análise clara do momento, a partir do qual, se torna mais vantajoso utilizar o *Fairlearn.* O estado da arte das plataformas *cloud*, a escolha de uma plataforma em detrimento de outra, este trabalho contribui para um melhor entendimento deste processo.

5.3. Trabalho futuro

A dissertação procurou descobrir diversos tópicos, contudo, esta área é diversificada e encontra-se numa dinâmica de evolução.

Há trabalho que poderia ser desenvolvido futuramente, como:

- Repetir o teste noutras plataformas de serviço de *cloud computing,* tais como *Google Cloud Plataform, Amazon AWS;*
- Realizar os testes com linguagens de programação diferentes do *Python;*
- Realizar o teste com um conjunto de dados diferentes.

6. REFERÊNCIAS

[1] L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty,” *Fuzziness and Knowledge-Based Systems,* pp. 571-588, 2002.

[2] T. Dalenius, “Finding a needle in a haystack or identifying anonymous census records,” *Journal of official statistics,* vol. 2, p. 329, 1986.

[3] K. A. S. S. D, “Architecture for preserving privacy during data mining by hybridization of partitioning on medical data,” *Architecture for preserving privacy during data mining by hybridization of partitioning on medical data,* pp. 93-97, 2010.

[4] J. C. R. K. T. K. T. Pete Chapman, “CRISP-DM 1.0. CRISP-DM,” vol. p.3, p. 76, 2000.

[5] K. LeFevre, D. J. DeWitt e R. Ramakrishnan, “Incognito: Efficient Full-Domain K-Anonymity,” pp. 49-60, 2005.

[6] Demo, “Metodologia do conhecimento cientifico,” *Atlas,* 2000.

[7] D. Tranfield, D. Denyer e Smart, “Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review,” *British Journal of Management,* vol. v.14, pp. 207-222, 2003.

[8] R. A. U. J. Leskovec J, “Mining of massive data sets.,” *Cambridge university press,* p. 9, 2020.

[9] S. Parsons, D. J. Hand, H. Mannila e P. Smyth, “The Knowledge Engineering Review 19,” em *Principles of Data Mining* , vol. nº2, MIT Press, 2004, p. 183.

[10] J. Dean, “. Big Data, Data Mining, and Machine Learning Value Creation for Business Leaders and Practitioners,” p. 265, 2014.

[11] M. &. R. I. Santos, “Business Intelligence: Tecnologias da Informação na Gestão de Conhecimento,” [Online]. Available: http://repositorium.sdum.uminho.pt/bitstream/1822/6198/1/Resumo_Livro_BI_MYS_IR.pdf. [Acedido em 01 10 2021].

[12] M. Ester, H.-p. Kriegel e J. S. e. X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” em *Knowledge Discovery and Data Mining*, Portland, USA,, 1996.

[13] B. Bengfort e J. Kim, “Data analytics with Hadoop: an introduction for data scientists,” O'Reilly Media, Inc, 2016.

[14] Hand, D. Mannila e H. Smyth, Principles of Data Mining Cambridge, MIT Press, 2001.

[15] L. Sweeney, “k-anonymity: a model for protecting privacy,” *International Journal on Uncertainty,* Vols. %1 de %2Fuzziness and Knowledge-based Systems, pp. 557-570, 2002.

[16] P. Samarati, “Protecting respondents identities in microdata release,” *IEEE Transactions on Knowledge and Data Engineering,* pp. 1010-1027, 2001.

[17] R. J. Bayardo, “Data Privacy Through Optimal k-Anonymization,” *IEEE international conference on Data Engineering,* pp. 217-228, 2008.

[18] "A proteção de dados na UE," [Online]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_pt. [Acedido em 04 november 2021].

[19] "Summary of the HIPAA Security Rule," [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html. [Acedido em 4 November 2021].

[20] "California Consumer Privacy Act (CCPA)," [Online]. Available: https://oag.ca.gov/privacy/ccpa. [Acedido em 06 December 2021].

[21] "Family Educational Rights and Privacy Act (FERPA)," [Online]. Available: https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html. [Acedido em 4 November 2021].

[22] L. Backstrom, C. Dwork e J. Kleinberg, "Wherefore Art Thou R3579X? Anonymized Social," 2007.

[23] D. L. C. a. H. T. Wang, "Privacy protection in social network data disclosure based on granular computing," em *IEEE International Conference on Fuzzy Systems*, 2006.

[24] M. &. M. G. &. J. D. &. T. D. &. W. P. Hay, " Resisting structural identification in anonymized social networks," vol. PVLDB. 1. 10.14778/1453856.1453873, pp. 102-114, 2008.

[25] Hay, Michael, G. Miklau, D. Jensen, D. Towsley e P. Weis, "Resisting structural re-identification in anonymized social networks," *Proceedings of the VLDB Endowment 1,* pp. 102-114, 2008.

[26] K. A., Motwani, N. R., S.U. e Y. Xu, “ Link privacy in social networks,” em *conference on Information and knowledge management* , 2008.

[27] N. A. e S. V., “De-anonymizing social networks,” *IEEE symposium on security and privacy,* pp. 173-187, 2009.

[28] H. Yu, P. Gibbons, Kaminsky e F. M. Xiao, “Sybillimit: A near-optimal social network defense against sybil attacks,” *IEEE Symposium on Security and Privacy,* pp. 3-17, 2008.

[29] J. Zhou B. Pei, “Preserving privacy in social networks against neighborhood attacks,” em *IEEE 24th International Conference on Data Engineering*, 2008.

[30] H. M., Miklau, J. G., D. Weis e S. S., “Anonymizing social networks,” *Computer science department faculty publication series,* p. 180, 2007.

[31] M. Faloutsos, Faloutsos e C. P. Faloutsos, “On power-law relationships of the internet topology,” em *In The Structure and Dynamics of Networks*, 2011.

[32] J. Travers e S. Milgram, “In The structure and dynamics of networks,” *An experimental study of the small world proble ,* pp. 130-148.

[33] L. M. Qaisi e I. Aljarah, “A twitter sentiment analysis for cloud providers: A case study of Azure vs. AWS,” *I Conference on Computer Science and Information Technology,* 2016.

[34] “Fairlerarn documentation,” [Online]. Available: https://fairlearn.org/. [Acedido em 06 June 2021].

[35] "Fairlearn metrics," [Online]. Available: https://fairlearn.org/v0.7.0/auto_examples/plot_quickstart_selection_rate.html#sphx-glr-auto-examples-plot-quickstart-selection-rate-py. [Acedido em 04 June 2021].

[36] "Guide on Article 8," [Online]. Available: https://www.echr.coe.int/documents/guide_art_8_eng.pdf.

[37] "Universal Declaration of Human Rights," [Online]. Available: https://www.un.org/en/about-us/universal-declaration-of-human-rights. [Acedido em 04 November 2021].

[38] J. Isaak e M. J. Hanna, "User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection," *User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection,* pp. 56-59, 2018.

[39] J. Isaak e M. J. Hanna, "User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection," *User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection,* pp. 56-59, 2018.

[40] V. Torra, " Data privacy: Foundations, new developments and the big data challenge," em *Data privacy: Foundations, new developments and the big data challenge*, Berlin, Springer International Publishing, 2017, pp. 20-27.

[41] R. Martins, "Abordagens quantitativa e qualitativa," em *Metodologia de pesquisa em engenharia de produção e gestão de operações*, Rio de janeiro, Elsevier, 2012, pp. 7-31.

[42] A. BRYMAN, "Barriers to integrating quantitative and qualitative research," *Journal of Mixed Methods Research,* vol. 1, pp. 24-35, 2011.

[43] R. MARTINS, “Principios da pesquisa cientifica,” em *Metodologia de pesquiisa em engenharia de produçao e gestao de operaçoes*, Rio de Janeiro, Elsevier, 2000, pp. 7-31.

[44] “Atenuar a injustiça com Fairlearn,” [Online]. Available: https://docs.microsoft.com/pt-pt/learn/modules/detect-mitigate-unfairness-models-with-azure-machine-learning/4-mitigate-with-fairlearn. [Acedido em 04 November 2021].

[45] “Atenuar a injustiça com Fairlearn,” [Online]. Available: https://docs.microsoft.com/pt-pt/learn/modules/detect-mitigate-unfairness-models-with-azure-machine-learning/4-mitigate-with-fairlearn. [Acedido em 01 November 2021].

[46] D. M. Powers, “Evaluation: from Precision, Recall and F-measure to,” *journal of Machine Learning,* p. 42, 2011.

[62] IBM. [Online]. Available: ftp://ftp:software:ibm:com/software/data/swlibrary/.

63] M. Ester, H.-p. Kriegel e J. S. e. X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” em *Knowledge Discovery and Data Mining*, Portland, USA, Portland, USA.

ANEXOS

Extração do dataset num open ml para o processamento distribuido, pré-procesamento dos dados.

```
# Pre processamento do data set Census
data = fetch_openml(data_id=1590, as_frame=True)
X_raw = data.data
y = (data.target == ">50K") * 1

# Separar as caracteristicas sensiveis sex e race
X_raw = data.data
y = (data.target == ">50K") * 1
A = X_raw[["race", "sex"]]
X = X_raw.drop(labels=['sex', 'race'],axis = 1)

# Separar o conjunto de dados em treino e teste
(X_train, X_test, y_train, y_test, A_train, A_test) = train_test_split(
    X_raw, y, A, test_size=0.3, random_state=12345, stratify=y
)

# Separar os indices em  X, y and A,
# Separar os dados em DataFrames
X_train = X_train.reset_index(drop=True)
X_test = X_test.reset_index(drop=True)
y_train = y_train.reset_index(drop=True)
y_test = y_test.reset_index(drop=True)
A_train = A_train.reset_index(drop=True)
A_test = A_test.reset_index(drop=True)
```

Código utilizado para a amostra estratificada, aplicando o algoritmo da regressão linear em *Azure ml*

```
# Definir o processamento em pipeline
numeric_transformer = Pipeline(
    steps=[
        ("impute", SimpleImputer()),
        ("scaler", StandardScaler()),
    ]
)
categorical_transformer = Pipeline(
    [
        ("impute", SimpleImputer(strategy="most_frequent")),
        ("ohe", OneHotEncoder(handle_unknown="ignore")),
    ]
)
preprocessor = ColumnTransformer(
    transformers=[
        ("num", numeric_transformer, selector(dtype_exclude="category")),
        ("cat", categorical_transformer, selector(dtype_include="category")),
    ]
)

# Put an estimator onto the end of the pipeline
lr_predictor = Pipeline(
    steps=[
        ("preprocessor", copy.deepcopy(preprocessor)),
        (
            "classifier",
            LogisticRegression(solver="liblinear", fit_intercept=True),
        ),
    ]
)
```

Código utilizado para criar um teste e carregar o *dashborad* em *Azure Machine Learning.*

```
 exp = Experiment(ws, "Fairness_Censo")
print(exp)

run = exp.start_logging()

# Upload the dashboard to Azure Machine Learning
try:
    dashboard_title = "Fairness insights of Logistic Regression Classifier"
    # Set validate_model_ids parameter of upload_dashboard_dictionary to False if you have not registered your model(s)
    upload_id = upload_dashboard_dictionary(run,
                                            dash_dict,
                                            dashboard_name=dashboard_title)
    print("\nUploaded to id: {0}\n".format(upload_id))

    # To test the dashboard, you can download it back and ensure it contains the right information
    downloaded_dict = download_dashboard_by_upload_id(run, upload_id)
finally:
    run.complete()
```