UNIVERSIDADE FEEVALE

JUNIOR MAURICIO STAUDT

MACHINE LEARNING PARA ANÁLISE DO DESGASTE DA FORÇA DE TRABALHO

Novo Hamburgo

2017

JUNIOR MAURICIO STAUDT

MACHINE LEARNING PARA ANÁLISE DO DESGASTE DA FORÇA DE TRABALHO

Trabalho de Conclusão de Curso
apresentado como requisito parcial
à obtenção do grau de Bacharel em
Sistemas de Informação pela
Universidade Feevale

Orientador: Rodrigo Rafael V. Goulart

Novo Hamburgo
2017

AGRADECIMENTOS

Quero agradecer minha esposa Micheli pelo amor, carinho, compreensão e apoio durante todos os anos que passamos juntos principalmente nos momentos difíceis.

Agradeço aos meus familiares pelo incentivo que sempre me deram, especialmente meu pai por me ensinar o valor do trabalho duro e da dedicação, minha mãe por sempre estar cuidando de mim e por ter me ensinado que o estudo pode me tornar uma pessoa melhor.

Um agradecimento especial para meu avô Valmor por suas histórias, conselhos, por acreditar na minha capacidade, o que me dá mais responsabilidade ainda para honrar toda a confiança depositada em mim.

RESUMO

As técnicas de aprendizado de máquina estão revolucionando o mundo da tecnologia dando mais poder de decisão às máquinas. O aprendizado de máquina faz com que o software possa aprender de acordo com as variáveis que são analisadas e tomar uma decisão. O presente trabalho consiste no estudo sobre *machine learning* ou aprendizado de máquina e como essa tecnologia pode ajudar as empresas a reter sua força de trabalho. O estudo consiste na execução de modelos de aprendizado de máquina para avaliação do desgaste ou atrito dos colaboradores em seus empregos. O resultado proveniente do estudo de algoritmos de aprendizado de máquina e de técnicas de gestão de pessoas, gerará uma base de conhecimento e modelos de aprendizado de máquina para avaliação de um conjunto de dados de treinamento e aplicação dos modelos em uma base de dados que foi coletada via questionário. Os resultados obtidos com os modelos de aprendizado de máquina foram analisados e comparados com as respostas dos participantes do questionário para verificar se existe uma correlação entre sua situação no emprego atual com o resultado da predição dos modelos elaborados. Por fim, apresento a conclusão da análise e a sugestão de trabalhos futuros que podem ser desenvolvidos a partir do estudo realizado nesse trabalho.

Palavras-chave: Inteligência artificial. Aprendizado de máquina. Recursos humanos. Análise preditiva. Gestão de pessoas.

ABSTRACT

Machine learning techniques are revolutionizing the technology world by giving more power to machines. Machine learning enables the software to learn according to the variables that are analyzed and make a decision. The present work consists of the study on machine learning and how this technology can help companies to retain their workforce. The study consists in to test machine learning models to evaluate the employees attrition in their jobs. The result of the study of machine learning algorithms and people management information will generate a knowledge base and machine learning models to evaluate a training dataset and the application of the models in a dataset collected through questionnaire. The results obtained with the machine learning models were analyzed and compared with the participants' answers to verify if there is a correlation between their situation in current job and the result of the predictions by the machine learning models. Finally, I present the conclusion of the analysis and the suggestion of future work that can be developed from the study presented in this work.

Key words: Artificial intelligence. Machine learning. Human resources. Predictive analysis. People management.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

TI	Tecnologia da informação
IBM	International Business Machine
SRB	Solution Representative Brand Specialist
TIC	Tecnologia da informação e comunicação
IBGE	Instituto Brasileiro de Geografia e Estatística
CV	Currículo Vitae
MLP	Multilayer Perceptron
LMT	Logistic Model Tree
CSV	Comma-separated values
ARFF	Attribute-Relation File Format
RNA	Rede neural artificial
RH	Recursos Humanos

SUMÁRIO

INTRODUÇÃO

Manter as pessoas engajadas e motivadas em um ambiente profissional tão competitivo como o atual é um desafio enorme para grandes, médias e pequenas empresas. As técnicas de *machine learning* (aprendizado de máquina) estão revolucionando o mundo da tecnologia dando mais poder de decisão para as máquinas. O aprendizado de máquina faz com que o software possa aprender de acordo com as variáveis que são analisadas e tomar a melhor decisão.

"*Can machines think?*" foi como Alan M. Turing questionou em seu *paper* para o periódico Mind da Universidade de Oxford, em 1950, no que foi chamado de o jogo da imitação (*Imitation Game*), onde ele descreve um jogo em que uma máquina responde perguntas assim como um outro participante humano e um juiz, também humano, deveria identificar quais das respostas pertenciam a máquina e quais pertenciam ao jogador humano. Turing é creditado como um dos pais da inteligência artificial (Turing, 1950).

De acordo com Bell (2015), pode-se desenhar sistemas que podem aprender e serem treinados a partir de suas experiências de modo a melhorar o seu modelo com o tempo e com a experiência adquirida para poder prever resultados baseados em um aprendizado já ensinado aos sistemas.

Machine learning é uma técnica utilizada para auxiliar os programas a aprenderem a partir de informações existentes em bases de dados cujo o principal objetivo é a previsão de resultados futuros, por exemplo, indicando um produto do agrado de um consumidor de acordo com o comportamento de compras dele (Gronlund, 2016).

Machine learning pode ser utilizado para vários objetivos, para ofertar programação de acordo com sua utilização na Netflix, saber o que estão falando sobre sua marca no Twitter e detecção de fraudes em compras com cartão de crédito, por exemplo (SAS, 2016). Todas essas análises de dados são realizadas sem que uma pessoa faça isso manualmente. A volumetria de dados existente atualmente é gigantesca, somente o Facebook gera 500 TB de informação por dia (McGlaun, 2012) e para conseguir realizar uma análise sobre esse volume imenso de dados, o *machine learning* é fundamental possibilitando o reconhecimento de padrões para sugerir resultados futuros.

Machine learning é a mesma coisa que *data mining* (mineração de dados)? A resposta para essa pergunta é não. *Machine learning* utiliza abordagens estatísticas e matemáticas para que possa aprender com os dados que são analisados e tomar decisões sem a intervenção

humana. A mineração de dados também utiliza abordagens utilizadas pelo *machine learning*, porém, a mineração de dados descobre informações e padrões ainda desconhecidos já o *machine learning* é utilizado para reconhecer padrões conhecidos e aplica-los em ações sem intervenção humana (SAS, 2016).

Esse trabalho consiste no estudo sobre *machine learning* e como essa tecnologia pode ajudar as empresas a reter seus talentos. Muitas organizações possuem dificuldades em identificar, capacitar e reter seus colaboradores, por outro lado, os colaboradores das empresas muitas vezes não conseguem se identificar com a empresa devido à falta de oportunidades de crescimento, valorização profissional através de promoções ou compensações financeiras ou até mesmo sentem falta de ter um ambiente desafiador no qual possam contribuir com suas habilidades e adquirir outras novas através da expertise nas funções desempenhadas dentro da organização (Martin, 2010).

O estudo consiste em construir um ativo intelectual sobre as técnicas e ferramentas de aprendizado de máquina e gestão de pessoas além de estudar trabalhos similares para propor uma aplicação das técnicas de *machine learning* em dados coletados ou públicos de funcionários de empresas para avaliar seu desgaste ou atrito.

O capítulo 1 apesenta uma abordagem sobre gestão de pessoas, quais fatores contribuem para a retenção de profissionais e porque essas pessoas são o motor da nova economia. O capítulo 2 aborda uma revisão bibliográfica sobre o que é *machine learning* e como essa tecnologia pode ajudar a resolver problemas complexos. O capítulo 3 descreve trabalhos e artigos publicados que tentaram resolver problemas similares e seus resultados. O capítulo 4 apresenta uma proposta de solução ao problema que este trabalho visa resolver. O capítulo 5 apresenta toda a execução do projeto e os resultados obtidos. Por fim, a conclusão é apresentada com alguns *insights* obtidos através das execuções de modelos de aprendizado de máquina além de apresentar sugestões para trabalhos futuros que podem ser desenvolvidos baseando-se nos resultados obtidos.

1 GESTÃO DE PESSOAS

Este capítulo aborda os conceitos básicos sobre gestão de pessoas como qual o seu contexto e quais os objetivos da sua aplicação. Será apresentado o que empresas bem-sucedidas estão fazendo para gerir e reter seus funcionários.

1.1 Conhecendo gestão de pessoas

Para entender melhor o que significa a gestão das pessoas, este capítulo abordará os seus conceitos básicos. Gestão de recursos humanos, também conhecido como RH, ou gestão de pessoas é um termo muito conhecido dentro do ambiente das organizações. Segundo Marques (2016):

Gestão de pessoas é o departamento dentro da empresa responsável por administrar e gerir o capital humano, também conhecido como Departamento Pessoal. Pode-se dizer que é o coração da organização, pois todos os processos pessoais de todos os colaboradores passam por essa área.

Seguindo a ideia de Marques (2016), é possível identificar que a gestão das pessoas é um departamento de extrema importância dentro das empresas. Também de acordo com o autor, esse departamento é responsável pela disseminação da cultura organizacional, da contratação das pessoas e dentre as competências do departamento de gestão de pessoas, está descrição dos cargos, treinamentos, planejamento de carreira, avaliação de desempenho e *feedback* aos colaboradores.

Para Chiavenato (2008), o RH pode ser visto de três formas, sendo a primeira delas funcionando como um departamento de operação prestador de serviços nas áreas de recrutamento e seleção, treinamento, remuneração, benefícios e comunicação. A segunda forma vista pelo autor é de que o RH é um conjunto de práticas que as empresas utilizam para dar suporte ao trabalho operacional mencionado no primeiro conceito. E por fim o último significado de RH refere-se a todos os colaboradores que trabalham no setor de RH e que são responsáveis pela aplicação das práticas aos serviços prestados pelo setor operacional. Para Gil (2007), a gestão de pessoas pode ser definida da seguinte maneira: “a gestão de pessoas abrange amplo leque de atividades, como recrutamento de pessoal, descrição de cargos, treinamento e desenvolvimento, avaliação de desempenho entre outras atividades”.

Analisando a descrição dos autores é perceptível que todos possuem o mesmo entendimento sobre o significado do que é gestão de pessoas ou RH. Mas qual seria o

significado das pessoas que trabalham para as empresas? Para Chiavenato (2008) a forma como as organizações determinam ou classificam as pessoas que compõem a sua força de trabalho é extremamente importante pois deixa claro qual é o papel e qual é a valorização que essas pessoas possuem para a sua empresa. Dentre as classificações citadas pelo autor estão funcionários, colaboradores, associados, operários, capital humano, capital intelectual, mão-de-obra direta e indireta e de acordo com o autor algumas empresas cometem a aberração de classificar suas pessoas como pessoal produtivo e improdutivo. Algumas empresas se referenciam a suas pessoas ou parte delas como talentos.

Mas o que diferencia as poessoas comuns dos talentos? De acordo com Chowdhury (2003), talentos são pessoas que dão grandes contribuições para as empresas, gerando grandes retornos sobre seus trabalhos e que precisam ser reconhecidas, estimuladas e alavancadas para que melhorem os resultados positivos que elas podem atingir. São pessoas diferenciadas, que podem ser consideradas as estrelas do time pois contribuem mais e devem ser melhor remuneradas do que os profissionais de nível mediano

Talentos geram inovação e a inovação leva à riqueza. Vivemos um momento único nos dias de hoje, onde quem possui uma boa ideia e talentos, consegue colocá-la no mercado rapidamente antes dos concorrentes podendo gerar ganhos exponenciais. O que os líderes buscam em um talento é a capacidade dele de gerar valor agregado mas, esse objetivo não é tão fácil assim de ser alcançado. No livro A Era do Talento, o autor Chowdhury (2003) indica que os talentos possuem uma cadeia de valor. Os ambientes empresariais onde a inovação prevalece são os lugares onde pessoas talentosas desejam trabalhar e essas empresas precisam saber reter seus talentos e seus colaboradores de bons desempenhos. A gestão de pessoas entra nesse contexto como um conjunto de práticas que podem auxiliar as organizações a superarem esses desafios.

1.2 O motor da nova economia

Talentos são pessoas criadoras, agentes de mudança e é quando as coisas estão mudando que eles se destacam mais pois são responsáveis por essas mudanças. De acordo com Chowdhury (2003) um talento inova, tem o tempo certo do negócio, otimiza os recursos e a produtividade, energiza as pessoas ao seu redor e assume mais responsabilidades tanto no sucesso quanto no fracasso. O desafio das empresas é cercar-se de pessoas talentosas.

Existem diferenças entre o que Chowdhury (2003) classifica como talento e profissional do conhecimento. Os profissionais do conhecimento são indispensáveis para as empresas, mas

são os talentos que fazem a diferença na hora de criar um produto ou serviço inovador e com alto valor agregado. Talentos costumam quebrar regras pois são inventores, pensam diferente e os profissionais do conhecimento costumam obedecer às regras e isso não significa que não sejam inteligentes, mas sim que são melhores na implementação do que na concepção de um produto ou serviço inovador que possa mudar o rumo da empresa. Talentos são geradores de mudanças enquanto profissionais do conhecimento dão sustentação a essas mudanças. As pessoas de talento lideram os profissionais do conhecimento. Talentos são agentes inspiradores e motivadores enquanto os profissionais do conhecimento são responsáveis por receber essas inspirações e a motivação. Os talentos geram grande contribuição e riqueza enquanto os profissionais do conhecimento compartilham dessa riqueza e das experiências que são geradas pelos talentos. A qualidade das pessoas é um fator fundamental para se obter sucesso.

A questão de se gerenciar as pessoas com maior qualidade vai ser um fator vital para as organizações da economia moderna. Pensando nisso Chowdhury (2003) sugere algumas formas de se lidar com a questão do gerenciamento das pessoas, sendo elas profissionais do conhecimento ou talentos.

1.3 Gestão de talentos e pessoas

Para reter talentos uma empresa precisa de um sistema de gerenciamento organizacional, esse sistema consiste em nove elementos e cada elemento possui uma série de boas práticas (Chowdhurry, 2003).

1. Sistema de foco no cliente voltado para os talentos;
2. Sistema de medição de satisfação e desempenho;
3. Sistema de gerenciamento participativo;
4. Sistema de gerenciamento da mudança;
5. Sistema de inovação constante;
6. Processo de formação de equipe de projetos;
7. Sistema de desenvolvimento de funcionários;
8. Sistema de gerenciamento de recursos humanos;
9. Sistema de suporte financeiro.

Chowdhury (2003) reforça muito que os talentos devem ter um tratamento diferenciado nas organizações, eles devem ser tratados como os principais clientes, fornecedores e acionistas da empresa. Criar um sistema diferenciado para atrair, reter e gerenciar os talentos não é algo fácil e se não for implantado de forma suave pode gerar uma ruptura entre os profissionais do

conhecimento e os talentos devido a diferenciação que é criada entre eles. A gerencia de linha deve ser responsável pela gestão dos talentos tendo o setor de RH da empresa apenas como uma área de suporte. Você precisa criar uma atmosfera que faça os melhores profissionais quererem ficar na sua empresa. Outra lição importante a ser aprendida na gestão de talentos é que não basta sua empresa ter vários talentos e não os colocar no local adequado isso pode gerar frustração e os talentos podem partir para outros locais. Se os talentos forem gerenciados estrategicamente eles irão produzir o máximo retorno.

Bichuetti (2015) critica alguns executivos por ainda não sabem quem são responsáveis pela gestão das pessoas, das quais ele chama de capital humano e pelo fato de considerar esses como custo e não como ativo. O autor prega que os responsáveis pela gestão de pessoas são os gestores e não o setor de RH. Pessoas são os ativos mais importantes dentro de uma empresa e para o autor, existem quatro fatores que atrapalham o criação e manutenção de times de alta performance, são eles:

1) Líderes que não enxergam as pessoas como um capital humano ou como um ativo da empresa, o que influencia a cultura organizacional e as ações dos gestores;
2) Falta de preparo dos executivos para gerenciar seus times e se tornando mau exemplo;
3) Falta de valorização do setor de RH e de alinhamento estratégico nas empresas; e
4) Não tratar esse tema como relevante no ensino superior.

Gerenciar pessoas de uma maneira eficaz é saber identificar atrair, contratar, reter, avaliar, remunerar, demitir e identificar as necessidades diferentes para cada perfil específico. Para o autor as empresas falam em falta de talentos e algumas vezes as pessoas até estavam na organização, porém "escondidas atrás da incompetência de seus gestores". Segundo o autor é preciso ter uma cultura que valorize as pessoas para que elas desejem continuar trabalhando nas empresas em que estão. Políticas de retenção de pessoas e principalmente gestores qualificados para gerir, são os fatores principais na retenção de colaboradores. Somente políticas de retenção não serão o suficiente para reter as pessoas, gestores qualificados são fundamentais para que pessoas desejem permanecer em suas organizações (Bichuetti, 2015).

1.4 Rotatividade e retenção da força de trabalho

Segundo Travis Bradberry (2015), em artigo publicado no LinkedIn, não é raro ver gestores reclamando que seus funcionários pediram demissão, além de tentarem esconder o sol com a peneira, ignorando o real motivo de seus empregados deixarem as empresas. Para ele,

pessoas não deixam os empregos, eles deixam seus gestores. Em sua opinião, existem nove fatores gerados pelos gestores que fazem os seus funcionários saírem.

1. Sobrecarga de trabalho: uma pesquisa publicada por John Pencavel (2014) pela Universidade de Stanford revela que a produtividade dos funcionários cai drasticamente após uma semana de trabalho com mais de 50 horas e cai mais drasticamente ainda após 55 horas semanais. Longas jornadas de trabalho estão diretamente ligadas ao absenteísmo e rotatividade dos colaboradores. Se o excesso de trabalho não for devidamente compensado os colaboradores podem pensar que estão sendo penalizados por terem bons desempenhos. Recompensar o excesso de trabalho com aumentos, promoções e maior status são formas de contornar o desgaste dos colaboradores.
2. Não recompensar nem reconhecer os méritos: gestores precisam compreender seus colaboradores para identificar suas necessidades e dessa forma prover o reconhecimento adequado para cada colaborador. Uns podem gostar de reconhecimento em público, outros podem preferir um aumento de salário e assim por diante.
3. Não se importar com seus funcionários: para Bradberry (2015) a maioria dos profissionais deixa seu emprego por causa de seu chefe. Gestores que não se importam com suas pessoas terão alto índice de rotatividade, pois, será difícil para que essas pessoas trabalhem quando não existe importância por parte dos gestores.
4. Não honrar compromissos: não cumprir com o que foi prometido pode passar a impressão de falta de comprometimento ou mesmo falta de honestidade do gestor. Esse comportamento pode causar a falta de confiança dos funcionários e seu desligamento.
5. Contratar e promover pessoas erradas: perder uma promoção por um profissional mal contratado ou de pior desempenho pode ser fatal para a permanência dos profissionais soa como um grande insulto.
6. Não deixar as pessoas perseguirem suas paixões: não deixar as pessoas correrem atrás de suas paixões com medo de que a produtividade diminua pode ser um fator desanimador para os bons profissionais. Pessoas que perseguem suas paixões e mantém um bom fluxo de trabalho podem ser até cinco vezes mais produtivas.
7. Não desenvolver as habilidades das pessoas: encontrar áreas onde os colaboradores podem evoluir é fundamental para que eles não fiquem entediados e complacentes. Gerenciar e dar *feedback* são tarefas indispensáveis para manter bons profissionais.

8. Não exercitar a criatividade: profissionais talentosos precisam de desafios e gostam de melhorar tudo que fazem. Conter ou evitar que esses profissionais exercitem seu poder de criatividade pode ser um erro por parte dos gestores e podem fazer as pessoas detestarem o seu trabalho.
9. Falta de desafios intelectuais: fazer com que seus profissionais busquem sair da sua zona de conforto através de metas desafiadoras podem estimular seu desejo e vontade de seguir trabalhando. Se a maior parte do tempo for de trabalho fácil ou chato, profissionais talentosos podem querer buscar uma oportunidade que desafiem sua inteligência.

Trabalhar todos esses fatores é fundamental se as empresas desejarem que as melhores pessoas trabalhem para elas.

2 CONHECENDO MACHINE LEARNING

2.1 O que é machine learning

Machine learning ou aprendizado de máquina é uma ramificação da inteligência artificial que resumidamente tem o objetivo de aprender com informações históricas através do método de treinamento e processamento dos dados. Os sistemas que usam aprendizado de máquina podem aprender através de experiências e com o tempo podem ser refinados para prever informações baseadas em questionamentos que são baseados no aprendizado já obtido (Bell, 2015).

Tom M. Mitchell é considerado *Chair of Machine Learning* pela universidade de Carnegie Mellon e a sua definição para *Machine Learning* é:

Um programa de computador é dito para aprender com a experiência *E* com relação a alguma classe de tarefas *T* e medida de desempenho *P*, se o seu desempenho em tarefas em *T*, medida por *P*, melhora com a experiência *E* (MITCHELL, 2016, tradução nossa).

Arthur Samuel definiu machine learning como um campo de estudo que dá aos computadores a habilidade para aprender sem ser programado explicitamente para isso. Aprendizado de máquina possui uma grande gama de algoritmos que podem ser utilizados para prever informações relevantes e são essas informações que definem qual algoritmo deve ser usado. Os algoritmos de aprendizado de máquina geralmente são enquadrados em duas ramificações, o aprendizado supervisionado e aprendizado não supervisionado. O aprendizado supervisionado os dados de treinamento são classificados com objetivo de aprender uma regra geral que vai mapear as entradas e saídas geradas pelo modelo aplicado. Já o aprendizado não supervisionado não etiqueta os dados, o algoritmo é responsável por identificar de forma automática os padrões existentes no conjunto de dados analisado. Os dados podem mudar, os requisitos podem mudar e os resultados podem mudar, porém para Bell (2015) não se deve pensar que os algoritmos irão sempre servir para a solução desejada ou seja, isso significa que a manutenção dos modelos de aprendizado de máquina exige a intervenção humana para manter os algoritmos sempre atualizados atendendo os requisitos e gerando os resultados desejados. Não espere criar um modelo de aprendizado e pensar que ele irá resolver seu problema para sempre.

2.2 Aplicações de machine learning

Aprendizado de máquina pode ser aplicado em várias situações para inúmeras funcionalidades, é utilizada especialmente no campo de desenvolvimento de software pois pode aprender o comportamento do usuário e após um determinado tempo prever quais serão as ações desse usuário. Além disso, aprendizado de máquina é utilizado na detecção de *spam*. *Spam* são e-mails não solicitados que por vezes lotam nossa caixa de entrada. Esse é um exemplo básico da utilização de aprendizado de máquina. Reconhecimento feito por assistentes de voz como a Siri da Apple, são exemplos do uso de *machine learning* pois elas vão aprendendo o comportamento do usuário e usam uma busca muito avançada baseada em computação na nuvem para analisar as perguntas dos usuários.

Aprendizado de máquina é amplamente utilizado no mercado de ações também, os algoritmos são utilizados para auxiliar investidores nas suas tomadas de decisão, ajudando na predição de resultados baseados em dados históricos. Segundo Bell (2015) outra área que utiliza amplamente os algoritmos de *machine learning* é a robótica. Não é raro ver nas redes sociais ou em sites de tecnologia novos robôs sendo apresentados ao mundo, robôs que se adaptam ao ambiente, robôs que imitam os movimentos dos animais ou mesmo que servem como uma espécie de *pet* ou bichinho de estimação. Todos esses robôs utilizam aprendizado de máquina para aprender com o ambiente ao seu redor sempre se baseando em dados estatísticos que são coletados e usados como fonte para predição e tomada de decisão.

Computadores com grandes capacidades de processamento estão ajudando no campo da medicina. Um exemplo desse cenário é o famoso Watsom, um supercomputador construído pela IBM para ajudar médicos ao redor do mundo a diagnosticar doenças. Não se assuste com essa informação, o objetivo é que o médico ainda seja o tomador da decisão e caberá a ele indicar a doença e o tratamento, porém muitas vezes existem inúmeras causas para sintomas em pacientes e o objetivo do Watsom é justamente ajudar os médicos a identificar de forma muito mais rápida a causa dos sintomas. Atualmente as pessoas geram muitos dados online e com esses dados gerados por celulares, dispositivos vestíveis e equipamentos de IoT, cada vez mais será possível acessar informações e utilizá-las para reconhecer padrões e sugerir ações (Bell, 2015).

Outro exemplo dado por Bell (2015) para aplicações de aprendizado de máquina são o e-commerce, lojas e supermercados que podem utilizar de ferramentas como cartões de fidelidade para obter dados sobre o comportamento de seus clientes e usar essas informações para criar campanhas de marketing e promoções específicas para seus clientes. Além disso o campo de

jogos digitais também utiliza dados para aprender com seus jogadores, até mesmo aprender quando um jogador possui dificuldades em concluir uma fase e adequar esse nível de dificuldade para mantê-lo no jogo evitando a sua desistência.

2.3 Linguagens e ferramentas de *machine learning*

Existem várias linguagens de programação que podem ser utilizadas no aprendizado de máquina. Dentre elas destacam-se Python, R, Matlab, Scala, Cloujure e Ruby são algumas e mais conhecidas linguagens de programação utilizadas em *machine learning*. Além das linguagens de programação, também existem ferramentas que podem auxiliar na criação de modelos de aprendizado além de possibilitarem a realização de experimentos em *datasets*. Uma dessas ferramentas é o software Weka que foi utilizado neste trabalho. Esse software permite fazer experimentos de mineração de dados e aprendizado de máquina. Grandes empresas como Microsoft, SAS, IBM e Google também possuem ferramentas de aprendizado de máquina.

Além disso, existem várias bibliotecas disponíveis na internet para integrar com o código Java entre outras linguagens, essas bibliotecas fornecem funções de aprendizado de máquina que podem ser utilizadas no desenvolvimento de software. Além disso outras ferramentas relacionadas são o Mahout, SpringXD e Hadoop.

Mas nada acontece no aprendizado de máquina se você não possuir um *dataset* ou conjunto de dados para usar, e esses dados podem ser adquiridos gratuitamente em alguns sites pela internet, entre eles se destacam o Kaggle que é um site que disponibiliza *datasets* para competição, além desse a Universidade de Irvine também disponibiliza um amplo repositório de dados, além de cidades como a de Nova Iorque nos EUA que disponibiliza dados gerais sobre a cidade para que as pessoas possam usar para o aprendizado de máquina (Bell, 2015). Aqui no Brasil o site do IBGE e os portais de transparências são ótimos lugares para conseguir dados.

2.4 Algoritmos utilizados

Esta seção apresenta os fundamentos dos algoritmos utilizados neste trabalho. A seção 2.4.1 apresenta o algoritmo C4.5 que é uma árvore de decisão. A seção 2.4.2 apresenta do algoritmo LMT (*Logistic Model Tree*). Por fim, a seção 2.4.3 apresenta o algoritmo MLP (*Multilayer Perceptron*). O capítulo 5 apresenta a execução dos algoritmos citados neste capítulo.

2.4.1 C4.5

Uma árvore de decisão é um modelo de classificação usado em inferências indutivas. Esse modelo é treinado para prever classes baseado nos valores dos atributos de um conjunto de dados de treinamento. No Weka o algoritmo C4.5 é representado pelo J48 (Bell, 2015). O C4.5 é um modelo do tipo árvore de decisão (*decision tree*) derivado do algoritmo ID3, também proposto por Ross Quinlan, devido ao fato de trabalhar com valores indisponíveis, com valores contínuos, podar árvores de decisão e derivar regras. A partir de uma árvore de decisão é possível derivar regras. As regras são escritas considerando o caminho do nodo raiz até uma folha. Estes métodos são geralmente utilizados em conjunto. Devido as árvores de decisão possuírem tendência a crescer muito elas são muitas vezes substituídas pelas regras. Isto ocorre em razão das regras poderem ser modularizadas. Uma regra pode ser compreendida sem que haja a necessidade de se referenciar outras regras (Quinlan, 1993).

O objetivo de uma árvore de decisão é criar um modelo viável para prever o valor de uma variável de saída tendo como base um conjunto de dados de entrada. Será preciso explicar onde árvores de decisão são mais utilizadas e quais as suas limitações para que seja possível compreender como ela pode ajudar na solução proposta por este trabalho. Árvores de decisão podem ser utilizadas para várias finalidades, desde prever se um cliente irá comprar determinado tipo de produto até a indústria de jogos e de redes sociais que utilizam árvores de decisão para fazer reconhecimento de movimentos e reconhecimento facial. A Microsoft utilizou três árvores de decisão para treinar sua plataforma Kinect no reconhecimento de movimentos usando um milhão de imagens e um cluster de 1.000 núcleos (Bell, 2015).

Para Bell (2015) árvores de decisão são fáceis de ler e essa é uma de suas grandes vantagens. Ela permite que você utilize informações numéricas ou categorizadas. É possível criar um modelo de trabalho usando dados formalizados em variáveis separadas por vírgula. Mesmo que você tenha um poder computacional razoável é possível ter uma boa performance com árvores de decisão mesmo que seja um conjunto de dados grande. Dependendo dos dados usados no conjunto de treinamento, árvores de decisão podem se tornar um modelo excessivamente complexo. Pode-se concluir que árvores de decisão são bons modelos para um conjunto de dados numérico e categorizado o que é o caso do *dataset* usado nesse trabalho, logo, árvore de decisão é um modelo a ser considerado como possível solução para o problema.

2.4.2 LMT – *Logistic Model Tree*

Logistic Model Tree combina os modelos de regressão logística com a indução de árvore, e, portanto, é um análogo de modelos de árvores para classificação de problemas. Uma LMT consiste basicamente em uma estruturada de árvore padrão com funções de regressão logística nas folhas, muito parecido com um modelo de árvore é uma árvore com funções de regressão regressão nas folhas. Como nas árvores de decisões comuns, um teste em um dos atributos é associado a cada nodo interno. Para um atributo nominal (enumerado) com valores *k*, o nodo tem *k* nodos filhos, e as instâncias são ordenadas abaixo de um dos ramos de *k* em função do valor de seus atributos. Para os atributos numéricos, o nodo tem dois nodos filhos e o teste consiste em comparar o valor do atributo: uma instância é classificada abaixo do ramo esquerdo se o valor para esse atributo é menor que o limite e são ordenados abaixo do ramo direito caso contrário. Formalmente, uma LMT consiste em uma estrutura de árvore feita de um conjunto de nodos internos ou não-terminais *N* e um conjunto de nodos terminais ou folhas *T*. Através do *S* é evidenciado todo o espaço de exemplo, gerado por todos os atributos que estão presentes nos dados. Em seguida, a estrutura da árvore gera uma subdivisão de *S* em regiões disjuntas *St*, e cada região é representada por uma folha na árvore (Landwehr, Hall, & Frank, 2006).

2.4.3 MLP – *Multilayer Perceptron*

MLP é um tipo de rede neural artificial. As RNAs oferecem a habilidade de aprender o desconhecido usando meios convencionais. Usando RNAs pode ser estabelecido um modelo livre de estimação do ambiente, permitindo um sistema que se adapte e que seja robusto. Como no cérebro humano a unidade base de processamento de uma RNA é o neurônio. As RNAs apresentam algumas características como várias unidades de processamento, ligações entre as unidades de processamento com pesos associados, processamento altamente paralelo e distribuído e a aprendizagem é realizada ajustando os pesos das conexões entre os neurônios. O MLP é um modelo de múltiplas camadas que estão ordenadas onde os neurônios de uma camada estimulam os neurônios da camada posterior sem estimular os neurônios da mesma camada ou das camadas anteriores, seu nome em inglês significa MultiLayer Perceptron (Perceptrons de múltiplas camadas). O MLP consiste em uma rede fortemente conectada com conexões *feedfoward* (Scikit-learn, 2017).

A Figura 1 mostra um esquema de como funciona o MLP, ou seja, ele é alimentado com várias entradas que são processadas pela camada de entrada que não possui capacidade para gerar os dados, essa camada gera estímulos para a segunda camada chamada de camada oculta. A camada oculta processa as informações de entrada e além disso podem haver mais camadas

ocultas que são usadas para gerar conexões entre as camadas de entrada e saída. A complexidade do problema é que determina a quantidade de camadas ocultas. A última camada é responsável por processar os dados de saída que é o resultado da predição (Djuris, et al., 2012).

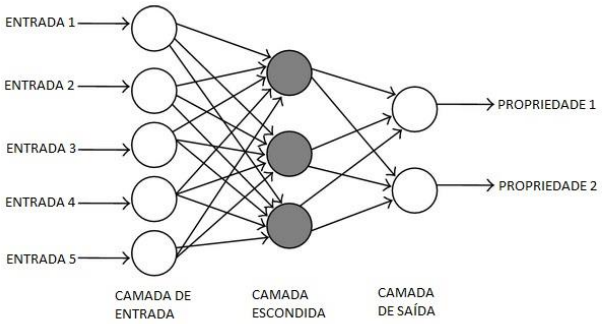


Figura 1 - Esquema do RNA Multilayer Perceptron.
Fonte: Scikit-learn, 2017.

3 APLICAÇÃO DE *MACHINE LEARNING* EM CASOS REAIS

Este capítulo aborda artigos com propostas similares a deste trabalho, como os autores tentaram resolver os problemas e quais foram os resultados obtidos. Essas características ajudarão na elaboração da proposta desse trabalho. Como foco central dos artigos analisados está uma empresa de serviços de TI, a IBM.

3.1 A predição de expertise dos empregados para gestão de talentos

A IBM é uma das maiores empresas do mundo com mais de 400 mil empregados e é essencialmente importante para sua estratégia de negócio e tomada de decisão, saber os dados completos, precisos e atualizados de seus empregados. Saber que time tem capacidade para dar suporte a determinado cliente ou qual área da empresa possui as competências necessárias para dar suporte a um produto ou serviço são dados crucias para a empresa (Varshney, et al., 2014).

3.1.1 Problemática

A IBM possui uma estrutura de 5 níveis para classificação dos seus empregados. Essa estrutura é composta por uma categoria principal, uma categoria secundária, função ou cargo, especialidade do cargo e competência ou habilidade, sendo que cada, empregado possui apenas uma única função e especialidade. A IBM possui um sistema que foi liberado no ano 2000 que possui três funções principais: a primeira é coletar informações das áreas de negócio onde o funcionário trabalha; a segunda é o empregado ter escolhido um cargo e uma especialidade; e a terceira é o sistema apresentar o conjunto de habilidades correspondente ao cargo selecionado e pedir para que o colaborador classifique as competências em uma escala até 5 pontos.

O número de cargos e especialidades cresceu muito ao longo dos anos e as regras de negócios de avaliação do cargo com as habilidades ficaram muito complexas pois as vezes o sistema chega a sugerir mais de 100 habilidades para o colaborador avaliar.

3.1.2 Objetivos

Em primeiro lugar, o trabalho apresenta uma dupla tarefa. A primeira é preencher os dados de cargos e especialidades dos cargos para os empregados que possuem essas informações em branco ou de forma inválida. A segunda é identificar os empregados que

possuem dados válidos, mas que não condizem com o seu cargo e especialidade atual que exercem.

A solução analítica não busca ser totalmente automática, mas sim recomendar predições e correções para os usuários, que podem ou não serem aceitas para cada empregado. Umas das abordagens para a solução do problema é a formulação de recuperação da informação que consiste na indexação de várias bases de dados que fornecem informações sobre a expertise dos funcionários, então através de consultas aos termos relacionados para cada cargo ou especialidade é possível ver no resultado de cada consulta o empregado melhor ranqueado. Porém, construir os termos da consulta de forma apropriada que diferenciem as características similares dos cargos e especialidades é difícil e isso é uma forma indireta de solucionar o problema. A maneira mais direta de abordar o problema é através da classificação pois a tarefa é rotular um indivíduo de um pequeno conjunto de rótulos.

De acordo com Varshney, et al. (2014) para uma abordagem *machine learning* é necessário ter dados de treinamento confiáveis. A primeira tarefa foi pegar uma fração dos empregados sem rótulos em branco e com rótulos válidos. A segunda tarefa é fazer uma validação cruzada que consiste em construir conjuntos de treinamento e de testes de forma que possam ser identificados os empregados que não se encaixam no padrão quando fazem parte do conjunto de testes. Para o problema enfrentado a formulação mais direta e adequada é a classificação de multi categoria supervisionada. (VARSHNEY, et al. apud FAN, CHANG, HSIEH, WANG, & LIN, 2014).

A *feature* consiste em pegar todas as oportunidades que um vendedor possui e a partir disso criar conjunto de palavras dessas descrições. Para criar esse conjunto de palavras os termos foram coletados do sistema de RH e da base de dados de negócios e do perfil pessoal do empregado. Os termos foram convertidos em caixa baixa e receberam um identificador (*token*). Foi computado também um conjunto de palavras para o título do cargo e para as *tags* (uma espécie de palavra-chave) da rede social da IBM.

Muitos modelos de classificação diferentes podem ser aplicados ao problema descrito. A métrica de performance desejada nesse problema é a precisão da classificação. Não é a métrica mais sensata em muitas aplicações, mas é a mais sensata para a predição de cargos porque as classes não são tão desequilibradas e diferentes tipos de erros de classificação não tem custos diferentes. O foco de estudo foram os empregados do setor de vendas da IBM no mundo inteiro. Para a linha de negócio estudada existem onze cargos válidos. Os nomes dos cargos são muitos similares apesar de serem diferentes e por isso não é fácil para os funcionários rotular-se com rapidez e precisão.

Como descrito anteriormente, todos os empregados com rótulos válidos foram utilizados no conjunto de testes totalizando 36.709 empregados que representam 89% da população de vendedores da IBM. A maior classe, SRB (*Solution Representative Brand Specialist*) representa 0.2633, ou seja, pouco mais de um quarto de todos os vendedores e essa é a *baseline* para a precisão da classificação. Foram comparados quatro algoritmos um contra o resto: regressão logística linear com regularização L2 e L1, *linear support vector machine* e Naive Bayes. Os parâmetros de regularização para os três primeiros modelos são encontrados por *cross-validation*. Foram realizados cinco testes de *cross-validation* para classificadores diferentes. Também foi comparado os quatro conjuntos individualmente e combinados: título do cargo, informação do RH, oportunidades de vendas e *tags* sociais.

3.1.3 Resultados obtidos

Foi percebido que o classificador Naive Bayes teve o pior desempenho em quase todos os conjuntos de atributos e *support vector machine* é consistentemente um pouco pior que a regressão logística. A precisão das *tags* sociais (C) foi pior que a precisão da *baseline*. A performance das oportunidades de vendas (D) foi bem pobre isso devido à falta de clareza das palavras na descrição das oportunidades. A melhor performance foi alcançada usando as informações do RH (B) e o conjunto de palavras do título do cargo (A) tanto individualmente como em conjunto (Varshney, et al., 2014).

Feature Set	ℓ_2-Reg. Logistic Regress.	ℓ_1-Reg. Logistic Regress.	Support Vec. Mach.	Naïve Bayes
Job Title (A)	0.6746	0.6749 ▲	0.6695	0.6410
HR Info (B)	0.7661 ▲	0.7641	0.7604	0.6807
Social Tags (C)	0.2320	0.2396	0.2380	0.2573 ▲
Sales Opp (D)	0.3374	0.3404	0.3473 ▲	0.2306
(A) + (B)	0.8016	0.8031 ▲	0.7899	0.7330
(A) + (B) + (C)	0.7671	0.7703 ▲	0.7504	0.6118
(A) + (B) + (C) + (D)	0.7720	0.7733 ▲	0.7655	0.3952

Figura 2 - Precisão do teste de validação cruzada de cinco dobras.
Fonte: Varshney, et al., 2014.

O estudo avança utilizando essas duas últimas funcionalidades para atingir uma melhor precisão na classificação. O modelo escolhido para a continuação do trabalho foi o de regressão logística linear com ℓ2-norm. Foi feita uma matriz de confusão que indicou que os problemas estavam concentrados em dois *clusters* (grupos) de cargos similares que são os vendedores técnicos e os representantes de soluções de marca. Foram aplicados métodos de pós processamento considerando a estrutura organizacional da empresa o que trouxe um decréscimo de precisão pois as informações obtidas do RH já continham dados relativos a estrutura organizacional do empregado.

A liberação do modelo preditivo foi feita em 2014 para todo o setor de vendas da IBM com a objetivo de reduzir o esforço manual de aproximadamente 2500 horas para a atualização dos cargos, funções e habilidades. Desde que as predições fossem de aproximadamente 80%, poucos gerentes teriam que fazer uma alteração manualmente. Estima-se que o retorno alcançado com uma pessoa de vendas seja de 1 milhão de dólares ao ano, sendo assim, o esforço reduzido no processo foi de aproximadamente 1 pessoa ao ano, ou seja, é como se uma pessoa a mais estivesse trabalhando pois não dispenderá tempo fazendo um processo demorado e manual e assim poderá gerar um retorno financeiro do seu setor para a IBM. As pessoas de vendas representam 10% da força de trabalho da IBM, a expectativa é de que quando o modelo seja aplicado na empresa toda a economia de tempo chegue em 20 pessoas ano.

3.1.4 Conclusões

O modelo proposto por Varshney, et al. (2014) foi concebido com o objetivo de diminuir um trabalho manual das pessoas do setor de vendas da IBM. A redução obtida chegou a atingir o esforço inteiro de uma pessoa por ano. O modelo preditivo também gerou benefício no processo de atualização das competências e habilidades dos empregados que podem ser facilmente atualizados mais do que uma vez ao ano. Foram usadas quatro diferentes fontes de dados para desenvolver a abordagem.

O modelo escolhido para fazer as análises foi o de regressão logística Liblinear L2-regularizado (normatizado). Para o futuro o objetivo é continuar a implementação e evangelização para as demais áreas da IBM e melhorias no modelo preditivo. É perceptível que a solução proposta por Varshney, et al., atende a necessidade da IBM para o setor de vendas da empresa. Porém a aplicação proposta sugere apenas a troca de função do empregado por uma oportunidade na qual ele tenha maior compatibilidade porém não indica quais habilidades e competências o colaborador deveria melhorar ou adiquirir para que possa continuar na mesma posição caso seja a vontade dele.

3.2 Avaliando as competências dos colaboradores para troca de área de atuação

Complementando o trabalho relatado na seção 1.1, um outro trabalho proposto por Wei, Varshney, & Wagman (2015) propôs solucionar um problema de demanda por profissionais em novas tecnologias como computação na nuvem, mobilidade, análise de dados entre outras, através da análise de profissionais internos da IBM, que atualmente trabalham com tecnologias legadas, para novas áreas de desenvolvimento. Fazer essa transferência sem gerar custos que

onerem o preço dos seus serviços e gerenciando o declínio da demanda do mercado pelas tecnologias legadas ou obsoletas, é um fator crucial de sucesso para a IBM.

3.2.1 Problemática

A demanda por profissionais qualificados em novas tecnologias é grande, conforme a consultoria IDC Brasil (2016) só no Brasil e no ano de 2016 o crescimento da demanda de profissionais de tecnologias deve ser de 2,6% o que pode parecer pouco mas, se comparado a taxa de desemprego que vem aumentando ao longo do tempo (IBGE, 2016) e a crise econômica pela qual o país está passando, pode-se dizer que a demanda por esses profissionais é alta porém a oferta desses profissionais é escassa visto que só no Brasil 50 mil vagas no setor de TIC não estão preenchidas (Dino, 2016).

Levando em consideração o cenário descrito acima, demitir funcionários que trabalham em tecnologias legadas para contratar funcionários que trabalham com novas tecnologias seria uma das possibilidades a serem adotadas, porém, os custos gerados por esse processo podem comprometer o preço do serviço prestado pela IBM e gerar uma perda de competitividade (Wei, Varshney, & Wagman, 2015). Além desses fatores o tempo e os custos de recrutamento, integração e a perda de produtividade de um novo empregado são problemas que prejudicam a escolha pela demissão. Segundo Wei, Varshney, & Wagman (2015) a melhor abordagem para solucionar o problema é encontrar e transferir colaboradores que trabalham com tecnologias legadas, que já possuam habilidades pré-requisitadas e estejam dispostos a se submeter a uma pequena carga de treinamento para adaptar-se a uma nova função na área de desenvolvimento.

3.2.2 Objetivos

De acordo com Wei, Varshney, e Wagman (2015) a principal contribuição do trabalho é gerar um algoritmo que permita a análise de dados das habilidades dos empregados para realizar suas transferências internas para as áreas de desenvolvimento.

Um dos objetivos do trabalho é identificar profissionais que possuam o perfil exatamente compatível para ocupar as novas vagas das áreas de desenvolvimento, podendo ser realizada uma transferência quase que imediata. Outro objetivo da solução que foi proposta por Wei, Varshney, e Wagman (2015) é analisar empregados que possuam os pré-requisitos que os habilitem a obter as competências exigidas para as novas vagas através de alguns poucos treinamentos. O objetivo do trabalho é aplicar a solução proposta para o setor de serviços de TI da IBM.

No trabalho proposto por Wei, Varshney, e Wagman (2015), foram utilizadas quatro bases de dados as quais são: avaliação da expertise (base principal), currículo vitae do empregado, dados históricos de projetos e informações básicas do RH. Um apontamento relevante feito pelos autores é de que o currículo vitae é uma ferramenta de dados não estruturados e utilizada pelos empregados tanto interna como externamente por isso não devem ser a base principal de consulta da solução e sim apenas ser usado como uma base de informação suplementar.

Foram utilizadas várias informações para a coleta de dados usadas como base para solução. A Figura 3 é um diagrama de blocos que demonstra o início do processo que passa primeiro pelas informações do RH. O processo pode ser continuado de duas formas através das definições para a filtragem dos dados sendo uma delas definindo a população fonte onde serão computados os perfis dos candidatos ou pela definição de uma população alvo onde serão computados os perfis dos indivíduos alvo. As duas filtragens utilizaram a base de avaliação da expertise (EA) e CVs como fonte de dados. A filtragem da população alvo resulta em mais um passo no processo que consiste em computar o perfil da população alvo. Após esse processamento ambas as filtragens geram a listagem dos candidatos pontuados versus o perfil alvo. Essas informações juntamente com suas informações do RH (local de trabalho, departamento, unidade de negócio, grade de pagamentos, entre outras) são ranqueadas.

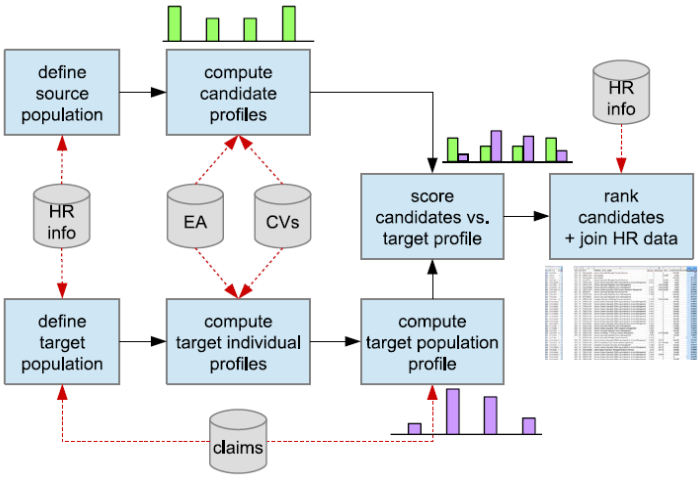


Figura 3 - Diagrama de bloco proposto para o algoritmo de análise.
Fonte: Wei, Varshney, e Wagman (2015).

O algoritmo proposto por Wei, Varshney, e Wagman (2015) deveria ser aplicado sobre uma população alvo bem definida versus um perfil de cargo ou expertise que representa essa população alvo. Uma população alvo pode ser definida como um ou mais serviços ofertados pela IBM e é composta pelos empregados que realizaram essa função ou ela pode ser definida pelo gerente de contratação do cargo que a empresa deseja suprir. Uma população alvo também

pode ser definida como um departamento inteiro para aumentar o público fonte. Público fonte pode ser definido como os empregados que empresa deseja considerar para o cargo a ser preenchido e, devido ao tamanho da empresa, pode abranger um público bem extenso.

O processo foi simplificado para que o público alvo seja somente da unidade de negócio onde estão os empregados candidatos. Também houve a exclusão dos funcionários que já atuam em uma área de desenvolvimento. Outras questões limitadoras no processo são o estado ou cidade para o qual a oportunidade é ofertada e também o país devido a questões de idioma ou legalidade para o trabalho. A grade de pagamento é um típico fator de limitação visto que geralmente devem ser as mesmas ou menor que o cargo disponível (Wei, Varshney, & Wagman, 2015).

O nível de experiência do candidato é dado por um perfil de quatro vetores positivos sendo um para cada medida (cargo, especialidade do cargo, habilidades e currículo vitae) sendo que o total de entradas (dimensão do vetor) dar-se-á pela quantidade de cargos e especialidades ocupados por uma pessoa na população alvo. Em relação ao CV a análise é diferente pois se trata de um texto não estruturado. Em relação as habilidades, sua quantidade pode ser muito expressiva gerando uma alta complexidade computacional para a solução, bem como, um número elevado de habilidades por candidato pode dificultar a seleção de um candidato com potencial. Para solucionar essas questões os autores definiram um conjunto de palavras relevantes ao perfil que se deseja buscar limitando a busca dos CVs e habilidades que contém ao menos uma das palavras relevantes.

Um perfil de especialização é criado para os indivíduos da população alvo e também define um perfil médio para a população alvo. Os candidatos são pontuados ao quão compatíveis com o perfil médio eles são, através da seguinte fórmula matemática:

$$s = \alpha^{\mathrm{JR}} S\left(\mathbf{x}^{\mathrm{JR}}, \mathbf{y}^{\mathrm{JR}}\right) + \alpha^{\mathrm{JRS}} S\left(\mathbf{x}^{\mathrm{JRS}}, \mathbf{y}^{\mathrm{JRS}}\right) + \alpha^{\mathrm{S}} S\left(\mathbf{x}^{\mathrm{S}}, \mathbf{y}^{\mathrm{S}}\right) + \alpha^{\mathrm{CV}} S\left(\mathbf{x}^{\mathrm{CV}}, \mathbf{y}^{\mathrm{CV}}\right)$$

Figura 4 - Fórmula de compatibilidade dos perfis candidatos.
Fonte: Wei, Varshney, e Wagman (2015).

Cada elemento da fórmula acima recebeu um peso sendo α^{JR} peso 12,50 α^{S} peso 37,50, α^{JRS} peso 25,00 e α^{CV} 25,00 totalizando peso 100. O elemento α^{JR} possui peso menor pois o cargo é muito genérico já as habilidades (α^{JRS}) possuem maior peso porque estão mais disponíveis do que os CVs. A função apresentada a seguir é uma forma de medir os vetores de expertise x e y entre si.

$$D(\mathbf{x},\mathbf{y}) = \sum_j w_j \frac{(y_j - x_j)_+}{y_j}$$

Figura 5 - Função D(x, y) para pontuar a diferença entre os vetores x w y.
Fonte: Wei, Varshney, e Wagman (2015).

O objetivo da função apresentada é dar uma avaliação de qualitativa do custo D(x,y) necessário para que um empregado candidato aprenda os conhecimentos necessários para conseguir a vaga ofertada. A expressão $(y_j - x_j)+$ é a quantidade de conhecimentos do tipo j que o candidato deve ganhar em relação ao alvo. A aquisição dessa experiência geralmente requer recursos, sejam treinamentos, cursos ou em tempo de trabalho. Os pesos w_j representam a importância relativa e o custo de aquisição de cada tipo de especialização. O custo é zero se o candidato já tiver a experiência desejada, isto é, se $x_j \geq y_j$.

Para cargos e especialidades a fórmulada é dada pela Figura 6. Para as habilidades, a variável y_j representa o número médio de habilidades sobre a população-alvo associada à palavra-chave ou especialidade de função de trabalho j, enquanto para a os CVs, a variável y_j é a incidência média da palavra-chave j. Estes podem ser menos indicativos de importância do que fracções de tempo de trabalho. Por exemplo, a taxonomia da IBM pode incluir mais habilidades contendo uma palavra-chave ou uma especialidade de função de trabalho para outra. Por esta razão, geralmente não definimos $w_j = y_j$ para habilidades e currículos, usando pesos ou pesos uniformes escolhidos pelo gerente de contratação ou pelo HR.

$$\begin{aligned} D(\mathbf{x},\mathbf{y}) &= \sum_j (y_j - x_j)_+ = \sum_j (x_j - y_j)_+ \\ &= \frac{1}{2}\sum_j |y_j - x_j| \,. \end{aligned}$$

Figura 6 - Fórmula dada para pontuação dos cargos e especialidades.
Fonte: Wei, Varshney, e Wagman (2015).

A forma funcional implica que D (x, y) ∈ [0,1] independentemente dos vetores de entrada x e y. Assim, seria razoável renunciar à normalização e aplicar diretamente uma média ponderada ao papel desempenhado, à especialidade do papel desempenhado, à habilidade e às distâncias CV. No entanto, verificou-se na prática que algumas das medidas de especialização não podem utilizar o intervalo completo da unidade. Isso acontece porque o vetor de especialidade de destino y é uma média sobre muitos indivíduos e nenhum funcionário único pode aproximar-se da amplitude em y, isto é, o menor valor observado de D (x, y) pode ser significativamente maior do que zero. Para corrigir esse viés, calculam-se as distâncias D (x, y) não apenas para todos os candidatos da população fonte, mas também para todos os membros da população alvo para fornecer uma comparação, tratando os funcionários alvo como

exemplos de trabalhadores candidatos bem qualifiados. Ao calcular D (x, y) para um empregado alvo, a média y pode ser modificada para deixar de fora o empregado que está sendo avaliado. Dado valores de distância para todos os empregados alvo, primeiro tomamos o complemento para obter uma pontuação 1-D (x, y) que é maior para as melhores correspondências. Em seguida, calculamos o p-quantil para obter um único ponto de referência.

$$S_p(\mathbf{y}) = Q\left(\{1 - D(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in \text{target}\}; p\right)$$

Figura 7 - Fórmula para calcular um ponto de referência.
Fonte: Wei, Varshney, e Wagman (2015).

Foi escolhido o percentual 95, isto é, p = 0,95, o que corresponde aos empregados alvo com melhor pontuação (p = 1). O quantil da população alvo é usado para normalizar as pontuações dos candidatos da seguinte forma:

$$S(\mathbf{x}, \mathbf{y}) = \frac{\min\{1 - D(\mathbf{x}, \mathbf{y}), S_p(\mathbf{y})\}}{S_p(\mathbf{y})}$$

Figura 8 - Fórmula de normalização das pontuações dos candidatos.
Fonte: Wei, Varshney, e Wagman (2015).

O passo final é produzir uma lista de candidatos classificados em ordem decrescente e seus escores globais. As pontuações dos componentes para diferentes medidas de especialização também podem ser mostradas para uma melhor compreensão da pontuação global do candidato. Como ponto de referência adicional, pode ser calculada uma estatística das pontuações globais para indivíduos alvo (novamente tratados como se fossem candidatos), por exemplo a mediana. Além de pontuações numéricas e identificadores de funcionários, a lista pode incluir informações úteis de RH, como gerente, departamento e outros detalhes organizacionais, grau de remuneração atual e localização geográfica. Todas essas informações são apresentadas a equipe de recursos humanos, equipes de gerenciamento de recursos humanos e aos gerentes de contratação para revisão.

3.2.3 Resultados obtidos

Nesta seção, serão resumidos os resultados obtidos até a data na IBM usando a abordagem analítica proposta para transferências de tarefas internas. Primeiro, discutimos o caso de uma grande equipe em um país europeu prestando serviços em um subcampo de uma das áreas de crescimento. Esta equipe foi uma das primeiras onde o algoritmo foi aplicado. A população fonte correspondente era composta por todos os funcionários daquele país que trabalhavam na divisão de serviços de TI, mas não na área de desenvolvimento. Foram avaliadas quatro medidas

de especialização: funções profissionais, especialidades de funções do trabalho, competências e certificações profissionais, sendo as últimas tratadas de forma semelhante às competências e substituindo os CVs. As pontuações foram computadas tanto para os candidatos como para os membros existentes da equipe, sendo que os últimos resultados ajudaram a normalizar o primeiro.

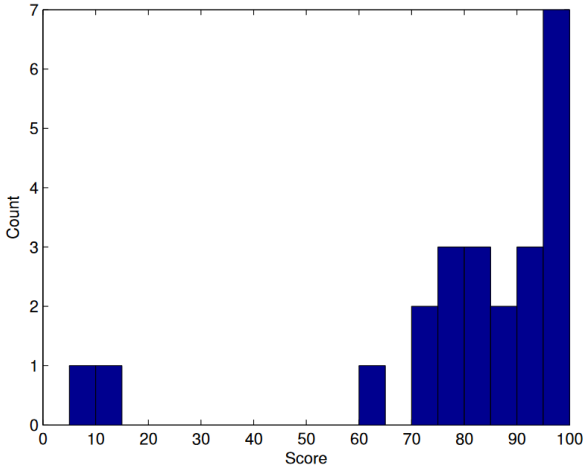


Figura 9 - Histograma de pontuações globais.
Fonte: Wei, Varshney, e Wagman (2015).

A calibração fornecida pela população-alvo é ilustrada na Figura 9, que mostra um histograma dos escores globais para a equipe, obtido como uma média das pontuações para as quatro medidas de especialidade com pesos $\alpha^{JR} = 100/7$, $\alpha^{JRS} = 200/7$, $\alpha^{S} = 200/7$ e Acert = 200/7. A maioria dos membros da equipe tem alta pontuação, com dois terços acima de uma pontuação de 80 e quase um terço acima de 95. Essa concentração de pontuação aumenta a confiança na capacidade do método de pontuação para identificar candidatos com qualificações semelhantes. Por outro lado, o histograma também mostra dois indivíduos com pontuações baixas, um resultado de ter papéis de trabalho, especialidades e habilidades muito diferente da maioria da equipe. Como a maioria das equipes não é completamente homogênea, a presença de tais *outliers* é talvez inevitável. Além disso, se vê o problema de transferência de trabalho como um de classificação, então os resultados na Figura 9 podem ser vistos como uma forma de validação cruzada, mas apenas para a classe positiva de funcionários qualificados. A validação similar para a classe negativa é dificultada pela falta de uma amostra pura - se a maior população de serviços de TI contém alguns candidatos qualificados como esperado, é por definição impura.

A partir dos escores calculados para a população fonte, uma lista dos 125 melhores candidatos foi compilada e analisada pelos gerentes de contratação e recursos humanos no país. As conclusões da revisão são apresentadas na Figura 10. Talvez a validação mais forte de nosso

algoritmo seja representada pelos 10 candidatos que, desconhecidos na época, já haviam sido previamente abordados sobre juntar-se à equipe, mas recusaram, foram contratados recentemente para a equipe ou foram ex-membros da equipe. Trinta e quatro dos candidatos foram considerados promissores o suficiente para ter seus CVs recuperados manualmente. Destes, 13 foram determinados como candidatos adequados para entrevistas. No entanto, nenhuma outra medida foi tomada, uma vez que esta avaliação se destinava principalmente a testar o algoritmo proposto. No lado negativo, apenas 14 candidatos foram excluídos totalmente como tendo experiência inadequada, enquanto o restante receberam avaliações neutras, nem promissoras nem inadequadas.

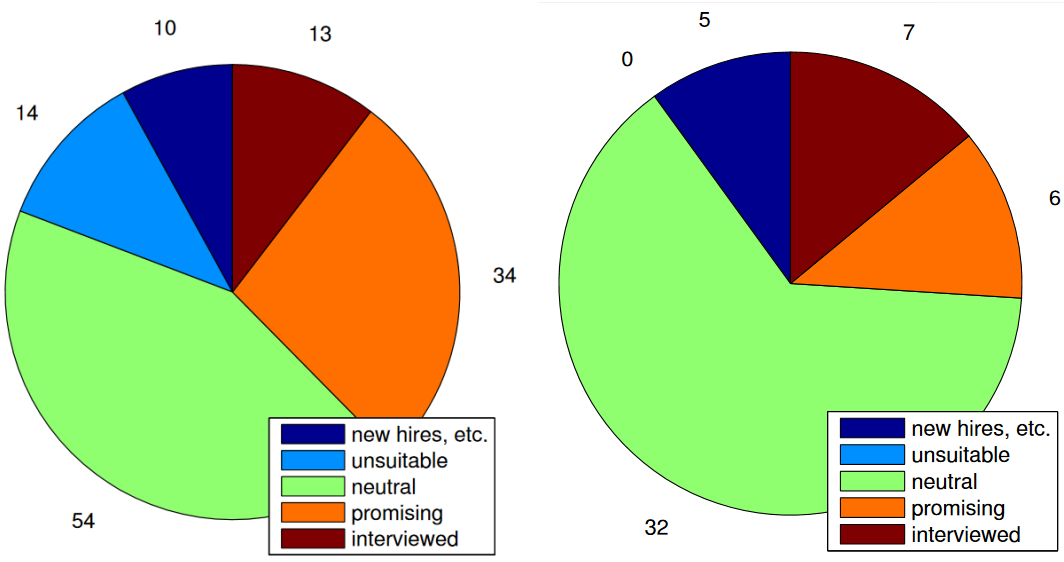


Figura 10 - Conclusões da revisão.
Fonte: Wei, Varshney, e Wagman (2015).

A Figura 10 também mostra uma lista de candidatos para uma equipe similar em um país latino-americano. Neste caso, os 50 candidatos principais foram escolhidos de uma população de origem que consiste em todos os funcionários de serviços de TI no país de fora da área de crescimento. As mesmas quatro medidas de especialização foram consideradas. Os resultados são amplamente semelhantes ao do país europeu. Além disso, o gerente de contratação na América Latina, motivado por uma posição aberta em sua equipe, deu o passo adicional de entrevistar 7 dos candidatos cujos currículos foram revisados, e selecionou um para preencher a abertura. Entretanto, o gerente do candidato escolhido era relutante em liberar o candidato na hora porque não havia nenhuma política ou procedimentos para encontrar um substituto.

Desde as avaliações nos dois países, o algoritmo foi refinado e aplicado para cerca de meia dúzia de lotes de posições abertas, no total fornecendo mais de 2.300 candidatos classificados para pelo menos 90 posições. O feedback que foi recebido geralmente era positivo. No entanto, não houve conhecimento de transferências reais que

tenham sido concluídas como resultado dessa abordagem. O algoritmo também forneceu valiosos resultados negativos: em um caso, a aparente falta de candidatos adequados deu aos gerentes de recursos humanos a confiança de que proceder com contratações externas foi a decisão certa.

Especificamente, o benefício de uma transferência interna resulta da evitação de custos associados a uma demissão e da evitação de custos associados à contratação externa, nomeadamente custos de recrutamento e de integração, bem como perdas iniciais de produtividade. Para estimar esses custos, de vários departamentos da IBM, foram coletados dados de custos de demissão em cada país, além de contratar dados de custo e salário em um nível mais refinado: país, função e remuneração. Em seguida, foram coletados dados sobre a distribuição de funcionários na divisão de serviços de TI por país, função e remuneração. Reunindo esses dados em uma média ponderada, foi obtido uma estimativa de poupança de mais de 50.000 dólares norte-americanos por transferência interna não incluindo perdas de produtividade. A maior parte desse montante foi devido a custos de demissão. A conclusão é que mesmo um pequeno número de transferências internas representaria um grande benefício para uma empresa.

Embora existam incentivos muito fortes, principalmente por meio de medições de utilização, para que os gerentes de serviços de TI "emprestem" os membros de suas equipes a projetos ou outras áreas do negócio, há um viés muito forte contra o deslocamento de indivíduos de forma permanente para um novo emprego. Os fatores que contribuem para este viés incluem:

1) Compromissos com contratos atuais ou futuros para indivíduos com a habilidade e treinamento: embora um indivíduo possa ter pausas em seu trabalho, permitindo-lhe trabalhar em outros projetos por um período definido de tempo, muitas vezes os gerentes têm trabalho atual ou estão antecipando um novo trabalho que exigiriam esses indivíduos. Manter esses indivíduos, com suas habilidades conhecidas e capacidades, na equipe aumenta a capacidade do gerente para executar.

2) A contratação de cargos abertos, mesmo de substituição, é cuidadosamente analisada: rigorosas metas e processos financeiros exigem a capacidade de substituir ou contratar um novo indivíduo em uma equipe desafiadora para um gerente. Se uma pessoa deixa a equipe, então trabalho adicional é criado para o gerente no desenvolvimento do caso para substituí-lo.

3) Potenciais candidatos são empregados na demanda: em virtude de ter as habilidades e experiência que os identificou como fortes candidatos ao algoritmo de análise, os funcionários identificados têm habilidades muito desejáveis para sua área atual de atribuição, bem como para a área de crescimento. Algumas discussões iniciais sobre a limitação do pool de potenciais

candidatos àqueles cuja utilização não é alta ou aqueles que não atendem a outras metas foram rapidamente descartadas como inadequadas para a atribuição de áreas estratégicas do negócio, a menos que fossem adequadas ao perfil alvo. Ter um plano de sucessão para cada abertura antecipada na cadeia é conceitualmente viável, mas tem limitações práticas.

4) Contratação externa é mais fácil: no momento em que o gerente de contratação fez o caso para construir uma equipe e tem os acordos financeiros e de gestão necessários para contratar para uma área de crescimento, o processo de recrutamento externo promete resultados mais rápidos e menos negociações internas. Gerentes de contratação que olhavam para potenciais candidatos identificados pelo processo, ficaram satisfeitos com os resultados, mas rapidamente encontraram resistência do atual gerente do funcionário e, na maioria das vezes, nem sequer concedeu permissão para ter uma discussão de qualificação com o empregado. Uma lição aprendida é que as equipes menos propensas a encontrar potenciais candidatos internos já tinham pesquisas externas ou equipes de recrutamento no lugar.

5) Não foram implementados programas de transformação amplos: o que é necessário é um programa de transformação amplamente apoiado, mas sem a comprovação da contratação e benefícios empresariais realizados, um investimento num programa de transformação é difícil de suportar.

Os autores relatam que foram aprendidas várias lições na implantação da abordagem de transferência de trabalho além da principal lição dos desafios organizacionais para adoção. Foi aprendido com o exame de listas de trabalho internas escritas pelos gerentes de contratação que é difícil para o pessoal que não é do RH menos familiarizado com a taxonomia de especialização da IBM aplicá-lo, especialmente para habilidades em áreas de crescimento emergentes. Por outro lado, os títulos de trabalho livres são altamente não padronizados e variáveis.

Um desafio inesperado no processo acabou por excluir todos os funcionários já trabalhando em uma área de crescimento. Muitas vezes áreas de crescimento são áreas novas ou emergentes dentro de uma empresa que pode não ter uma estrutura organizacional, como departamento ou códigos de faturamento, que pode ser usado para identificá-los. Portanto, muitas vezes, inadvertidamente, incluir alguns funcionários da área de crescimento em nossa população fonte. No entanto, encontrar esses indivíduos no processo ajudou a validar a eficácia do algoritmo, especialmente para aqueles que podem ser inicialmente céticos da abordagem. Outra lição é que os dados de perícia disponíveis para nós não capturam fatores pessoais relevantes para a adequação dos candidatos, por exemplo vontade de mudar de horário de trabalho ou relocalização. Esses fatores só podem ser obtidos dos próprios funcionários depois de terem sido altamente classificados.

3.2.4 Conclusões

No trabalho exposto neste capítulo, foi apresentada uma solução baseada em dados, para permitir a transferência de funcionários de uma grande empresa de serviços de TI de áreas legadas para áreas de crescimento. A principal fonte de dados é a informação de avaliação de experiência a partir da qual é possível entender as habilidades e competências dos funcionários analisados. O algoritmo de análise de dados cria um perfil estatístico para uma equipe de área de crescimento direcionada e classifica os funcionários de uma ampla fonte de população em toda a empresa para adequação contra esse perfil. Foram estimadas grandes transferências internas habilitadas por dados podem proporcionar benefícios financeiros muito significativos para as empresas. O algoritmo foi testado com equipes de serviços de TI do mundo real dentro da IBM Corporation e as saídas são mais do que satisfatórias através da validação cruzada empírica e através das experiências dos gestores de recrutamento reais. Contudo, não foram facilitadas transferências internas reais devido a barreiras organizacionais independentes da análise. Há esperança de que as mudanças institucionais virão e delinearão algumas etapas na direção certa.

1) Fornecer incentivos para a equipe de gestão para participar: para os gerentes de contratação, calibrar o processo de contratação para agilizar a contratação de candidatos atualmente empregados pela empresa, por exemplo, priorizando buscas internas antes de dar permissão para contratar externamente. Para os gerentes dos candidatos, fornecer um processo simplificado para candidatos selecionados que remove a dor da negociação de recursos de substituição, e também compreender e respeitar onde os gestores têm pressões de custo a curto prazo significativo.

2) Compromisso forte dos stakeholders: no ambiente de gerenciamento de uma grande empresa, os executivos que estão fortemente comprometidos com um programa de transformação podem não ser os executivos a quem o gerente de contratação ou o gerente atual do candidato relatam. Assim, o compromisso e as medidas dos gestores de ambos os lados da transação pode não ser suficiente, ou eles podem não ter a visibilidade incentivando-os a agir. Portanto, é necessária uma ampla liderança executiva e apoio, com o ponto de tomada de decisão favorecendo o movimento para o crescimento da área estratégica para ajudar a remover a inércia organizacional.

3) Aumentar a conscientização dos funcionários: os funcionários estão animados em trabalhar em áreas de desenvolvimento estratégico, mas muitas vezes não percebem como suas habilidades poderiam ser aplicadas ou como elas poderiam ser identificadas. Sabendo que há

um veículo neutro, como o modelo apresentado, sugerindo oportunidades isso irá aumentar a sua confiança nas oportunidades de carreira da organização.

Além das mudanças organizacionais recomendadas, também recomendamos a seguinte pesquisa técnica futura. O algoritmo atual não modela a similaridade entre papéis de trabalho diferentes, especialidades e habilidades ou como é fácil adquirir uma nova habilidade dada as existentes. Se fizermos tais análises, quer através do exame de dados históricos sobre trajetórias de aquisição de habilidades ou sobre a ocorrência de habilidades entre os empregados, pode-se generalizar a métrica de distância da variação total da distância.

4 PROPOSTA DE SOLUÇÃO

A proposta inicial deste trabalho era desenvolver um sistema informatizado que faria a avaliação das competências dos talentos de uma empresa e informe se os profissionais analisados estão adequados a função que exercem dentro de suas organizações. Além disso, outro objetivo buscado por esse projeto é obter resultados relevantes que possam dar aos gestores e RH, um indicativo de troca de função ou de habilidades que precisariam ser adquiridas para que sua compatibilidade com a função exercida aumente. Esses objetivos foram alterados ao longo do trabalho dada a dificuldade de se conseguir um conjunto de dados que atendesse as necessidades para gerar um *dataset* de treinamento além de deixar uma quantidade de dados para teste de acordo com o objetivo inicial que era prever as habilidades e sugerir mudanças de cargos.

Contudo, ao descrever a sessão 1 (Gestão de pessoas) percebe-se de acordo com Chowdhury (2003) que existe uma distinção entre talentos e profissionais do conhecimento. Devido a essa diferenciação não seria correto construir um sistema de gestão de talentos, considerando todos os avaliados como talentos. Seria necessário fazer uma distinção entre os profissionais avaliados e fazer a gestão somente dos que forem classificados como talentos. Como o objetivo do trabalho era inicialmente avaliar toda uma força de trabalho de uma ou mais empresas, sendo estes talentos ou não, e ajuda-los a serem enquadrados nas funções das quais podem aumentar a sua contribuição, esse trabalho não contemplará a identificação de talentos. Portanto, o trabalho será referido a partir de agora a gestão dos profissionais do conhecimento ou simplesmente força de trabalho e não talentos. A identificação dos talentos e uma gestão específica para este tipo de profissional poderá ser sugerida para trabalhos futuros.

As empresas bem-sucedidas e inovadoras tendem a ter um trabalho muito forte na gestão e retenção da sua força de trabalho. É notável que ao longo dos anos, as evoluções tecnológicas e industriais estão modificando a forma como as pessoas atuam em suas empresas, diminuindo o trabalho "braçal" e aumentando o trabalho mental (BRYNJOLFSSON & MCAFEE, 2016). Com o auxílio da inteligência artificial, sua subárea de aprendizado de máquina, o trabalho objetiva identificar as competências técnicas, pessoais e comportamentais dos profissionais. Por meio de uma análise dessas características, que serão cruzadas com as características da função que o profissional exerce dentro da organização, haverá um resultado de compatibilidade entre profissional e função exercida.

Com base nos trabalhos relacionados e na análise de gestão de talentos, a abordagem a ser usada é a da criação de modelos de aprendizado de máquina que possam prever se um funcionário está desgastado ou em atrito com seu atual trabalho. A contribuição do trabalho é disponibilizar um método de análise que ajude na identificação do desgaste da força de trabalho e possa dar subsídio para os gestores de equipes e times de RH na tomada de decisão para manter ou não os colaboradores.

4.1 Metodologia científica a ser aplicada

O projeto foi desenvolvido por meio de pesquisas bibliográficas em trabalhos acadêmicos, livros, publicações e periódicos abrangendo os assuntos de inteligência artificial e sua subcategoria *machine learning* (aprendizado de máquina), análise preditiva de dados, recursos humanos, gestão de talentos, planejamento de talentos e análise de força de trabalho para que deem embasamento teórico sobre os assuntos abordados no projeto de pesquisa.

O trabalho é composto por alguns modelos de pesquisa com base na metodologia de pesquisa proposta por Prodanov e Freitas (2013). O método de pesquisa foi fenomenológico buscando na experimentação, a aplicação dos conceitos de *machine learning* no ambiente coorporativo do mundo real. Em relação a natureza da pesquisa, ela foi uma pesquisa aplicada que visa gerar um produto ao final do trabalho. Quanto aos objetivos da pesquisa, foi uma pesquisa exploratória na sua fase de concepção visto que é necessário realizar levantamento bibliográfico. Quanto a abordagem de pesquisa, ela foi qualitativa visto que seu principal objetivo é identificar colaboradores que podem ter desgaste ou atrito no emprego atual e dar aos gestores ou ao setor de RH da empresa uma ampla visão para a tomada de decisão (VIANNA, 2016). Quanto aos procedimentos a serem adotados na pesquisa eles serão vários, por meio de pesquisa bibliográfica, como já citado, questionário para levantamento de dados, estudo de caso e pesquisa-ação (GIL, 2010).

O universo de pesquisa deste projeto são os diversos profissionais que estão em atividade até o momento em que responderem o questionário. A validação dos resultados se dará por meio da comparação do questionário, aplicado para coleta de dados dos profissionais, comparados aos resultados do algoritmo de predição deste projeto.

5 IMPLEMENTAÇÃO

Este capítulo apresenta como o *dataset* foi obtido, quais foram os processamentos executados nesse conjunto de dados para que eles ficassem prontos para serem usados no Weka. Também será apresentado todos os atributos existentes e qual o seu significa, visando dar uma visibilidade ao leitor de quais informações estão sendo utilizadas para a criação do modelo de aprendizado de máquina. Será apresentado os resultados relevantes das execuções dos algoritmos aplicados sobre os *datasets*, concluindo com uma análise sobre os resultados obtidos e comparando os mesmo com as respostas dos participantes que responderam o questionário para participar desse trabalho.

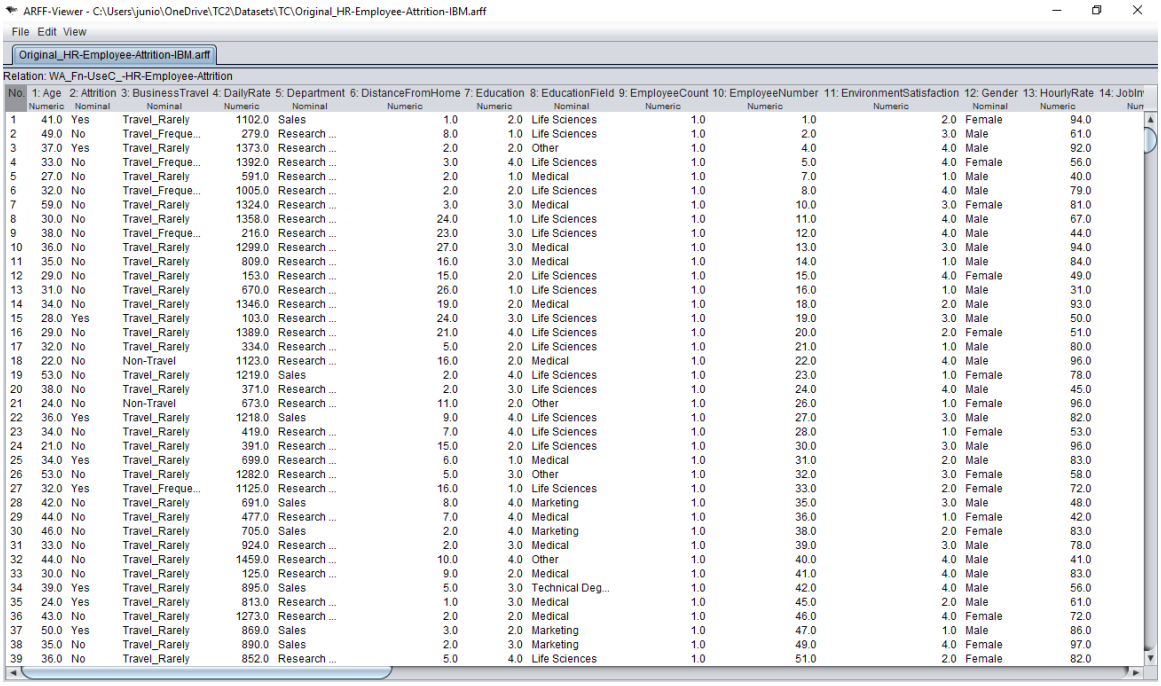
5.1 Base de dados

Este subcapítulo apresenta a escolha do *dataset* de treinamento juntamente com a elaboração da base de dados para aplicação dos modelos de aprendizado de máquina que foram utilizados na implementação. Para este trabalho foi utilizada uma base pré-existente, com disponibilidade pública e com uma quantidade significativa de registros ou instâncias para a realização da construção do modelo de treinamento que foi aplicado em um outro *dataset*, nomeado de produção, onde existe um conjunto de dados diferente dos dados de treinamento e que é usado para aplicação do modelo de melhor desempenho e análise dos resultados obtidos.

O *dataset* possui dados fictícios, ou seja, é um conjunto de dados sintéticos criado por cientistas da IBM e disponibilizado em seu site através de uma planilha eletrônica com 1.470 instâncias e 35 atributos (IBM Watson Analytics, 2015). Apesar de ser uma base sintética, não foram encontradas informações adicionais sobre como a base de dados foi criada ou se os dados foram coletados de pessoas reais.

Feito o *download* do arquivo CSV que contém os dados usados para treinamento do modelo de aprendizado de máquina, o mesmo foi renomeado para Original_HR-Employee-Attrition-IBM.csv, esse procedimento é importante para o restante do trabalho visto houveram alterações nos arquivos e foram geradas novas versões a partir da versão original. Para que o arquivo original não se misture ele foi renomeado para uma fácil identificação. Após isso, antes das execuções dos modelos de treinamento, foi necessário verificar se o arquivo CSV está pronto para ser utilizado no Weka, que é um software de *data mining* e *machine learning* e que foi utilizado nesse projeto. Para validar o arquivo é preciso usar o recurso ArffViewer do Weka, abrindo o arquivo CSV e conferindo todas as colunas existentes no conjunto de dados. A Figura

11 mostra o arquivo original no formato CSV aberto no ArffViewer. Não houve erros ao executar esse procedimento. Para que o arquivo fique padronizado com a extensão padrão executada pela Weka, após abrir o arquivo CSV no ArrfViewer, é recomendado salvar o arquivo na extensão padrão ARFF.



No.	1: Age	2: Attrition	3: BusinessTravel	4: DailyRate	5: Department	6: DistanceFromHome	7: Education	8: EducationField	9: EmployeeCount	10: EmployeeNumber	11: EnvironmentSatisfaction	12: Gender	13: HourlyRate	14: JobIn
	Numeric	Nominal	Nominal	Numeric	Nominal	Numeric	Numeric	Nominal	Numeric	Numeric	Numeric	Nominal	Numeric	Nun
1	41.0	Yes	Travel_Rarely	1102.0	Sales	1.0	2.0	Life Sciences	1.0	1.0	2.0	Female	94.0	
2	49.0	No	Travel_Freque...	279.0	Research ...	8.0	1.0	Life Sciences	1.0	2.0	3.0	Male	61.0	
3	37.0	Yes	Travel_Rarely	1373.0	Research ...	2.0	2.0	Other	1.0	4.0	4.0	Male	92.0	
4	33.0	No	Travel_Freque...	1392.0	Research ...	3.0	4.0	Life Sciences	1.0	5.0	4.0	Female	56.0	
5	27.0	No	Travel_Rarely	591.0	Research ...	2.0	1.0	Medical	1.0	7.0	1.0	Male	40.0	
6	32.0	No	Travel_Freque...	1005.0	Research ...	2.0	2.0	Life Sciences	1.0	8.0	4.0	Male	79.0	
7	59.0	No	Travel_Rarely	1324.0	Research ...	3.0	3.0	Medical	1.0	10.0	3.0	Female	81.0	
8	30.0	No	Travel_Rarely	1358.0	Research ...	24.0	1.0	Life Sciences	1.0	11.0	4.0	Male	67.0	
9	38.0	No	Travel_Freque...	216.0	Research ...	23.0	3.0	Life Sciences	1.0	12.0	4.0	Male	44.0	
10	36.0	No	Travel_Rarely	1299.0	Research ...	27.0	3.0	Medical	1.0	13.0	3.0	Male	94.0	
11	35.0	No	Travel_Rarely	809.0	Research ...	16.0	3.0	Medical	1.0	14.0	1.0	Male	84.0	
12	29.0	No	Travel_Rarely	153.0	Research ...	15.0	2.0	Life Sciences	1.0	15.0	4.0	Female	49.0	
13	31.0	No	Travel_Rarely	670.0	Research ...	26.0	1.0	Life Sciences	1.0	16.0	1.0	Male	31.0	
14	34.0	No	Travel_Rarely	1346.0	Research ...	19.0	2.0	Medical	1.0	18.0	2.0	Male	93.0	
15	28.0	Yes	Travel_Rarely	103.0	Research ...	24.0	3.0	Life Sciences	1.0	19.0	3.0	Male	50.0	
16	29.0	No	Travel_Rarely	1389.0	Research ...	21.0	4.0	Life Sciences	1.0	20.0	2.0	Female	51.0	
17	32.0	No	Travel_Rarely	334.0	Research ...	5.0	2.0	Life Sciences	1.0	21.0	1.0	Male	80.0	
18	22.0	No	Non-Travel	1123.0	Research ...	16.0	2.0	Medical	1.0	22.0	4.0	Male	96.0	
19	53.0	No	Travel_Rarely	1219.0	Sales	2.0	4.0	Life Sciences	1.0	23.0	1.0	Female	78.0	
20	38.0	No	Travel_Rarely	371.0	Research ...	2.0	3.0	Life Sciences	1.0	24.0	4.0	Male	45.0	
21	24.0	No	Non-Travel	673.0	Research ...	11.0	2.0	Other	1.0	26.0	1.0	Female	96.0	
22	36.0	Yes	Travel_Rarely	1218.0	Sales	9.0	4.0	Life Sciences	1.0	27.0	3.0	Male	82.0	
23	34.0	No	Travel_Rarely	419.0	Research ...	7.0	4.0	Life Sciences	1.0	28.0	1.0	Female	53.0	
24	21.0	No	Travel_Rarely	391.0	Research ...	15.0	2.0	Life Sciences	1.0	30.0	3.0	Male	96.0	
25	34.0	Yes	Travel_Rarely	699.0	Research ...	6.0	1.0	Medical	1.0	31.0	2.0	Male	83.0	
26	53.0	No	Travel_Rarely	1282.0	Research ...	5.0	3.0	Other	1.0	32.0	3.0	Female	58.0	
27	32.0	Yes	Travel_Freque...	1125.0	Research ...	16.0	1.0	Life Sciences	1.0	33.0	2.0	Female	72.0	
28	42.0	No	Travel_Rarely	691.0	Sales	8.0	4.0	Marketing	1.0	35.0	3.0	Male	48.0	
29	44.0	No	Travel_Rarely	477.0	Research ...	7.0	4.0	Medical	1.0	36.0	1.0	Female	42.0	
30	46.0	No	Travel_Rarely	705.0	Sales	2.0	4.0	Marketing	1.0	38.0	2.0	Female	83.0	
31	33.0	No	Travel_Rarely	924.0	Research ...	2.0	3.0	Medical	1.0	39.0	3.0	Male	78.0	
32	44.0	No	Travel_Rarely	1459.0	Research ...	10.0	4.0	Other	1.0	40.0	4.0	Male	41.0	
33	30.0	No	Travel_Rarely	125.0	Research ...	9.0	2.0	Medical	1.0	41.0	4.0	Male	83.0	
34	39.0	Yes	Travel_Rarely	895.0	Sales	5.0	3.0	Technical Deg...	1.0	42.0	4.0	Male	56.0	
35	24.0	Yes	Travel_Rarely	813.0	Research ...	1.0	3.0	Medical	1.0	45.0	2.0	Male	61.0	
36	43.0	No	Travel_Rarely	1273.0	Research ...	2.0	2.0	Medical	1.0	46.0	4.0	Female	72.0	
37	50.0	Yes	Travel_Rarely	869.0	Sales	3.0	2.0	Marketing	1.0	47.0	1.0	Male	86.0	
38	35.0	No	Travel_Rarely	890.0	Sales	2.0	3.0	Marketing	1.0	49.0	4.0	Female	97.0	
39	36.0	No	Travel_Rarely	852.0	Research ...	5.0	4.0	Life Sciences	1.0	51.0	2.0	Female	82.0	

Figura 11 - ARFF Viewer com lista de dados de treinamento.
Fonte: Elaborado pelo autor.

O *dataset* possui 35 atributos que serão descritos a seguir e que formaram a base para coletar os dados do conjunto de produção por meio de questionário. A seguir é listado os atributos do *dataset*:

1. *Age*: atributo numérico que indica a idade da pessoa relativa a instância analisada;
2. *Attrition*: atributo nominal alvo (atributo classe), é a informação que desejamos descobrir ao aplicar um modelo de aprendizado de máquina. Indica se a instância ou registro possui ou não atrito, ou seja, indica se a pessoa que tem seus dados analisados, possui atrito com seu atual emprego;
3. *BusinessTravel*: atributo nominal, possui três categorias sendo uma delas que indica se a pessoa relativa a instância analisada viaja frequentemente, se viaja raramente ou se não viaja;
4. *DailyRate*: atributo numérico, indica um valor diário que não foi identificado para qual propósito ou ao que se refere;
5. *Department*: atributo nominal que indica o departamento onde a pessoa relativa a instância analisada trabalha;

6. *DistanceFromHome*: atributo numérico que indica a distância que a pessoa relativa a instância analisada mora do seu local de trabalho. Não foi identificada se a distância foi informada em milhas ou em quilômetros;
7. *Education*: atributo numérico que é usado para identificar qual o nível de educação da pessoa relativa a instância analisada;
    a. 1 – *Below College*;
    b. 2 – *College*;
    c. 3 – *Bachelor*;
    d. 4 – *Master*; e
    e. 5 – *Doctor*.
8. *EducationField*: atributo nominal usado para identificar qual o campo de estudo da pessoa relativa a instância analisada;
9. *EmployeeCount*: atributo numérico que indica apenas a contagem da pessoa relativa a instância analisada. Terá o valor 1 para todos os registros;
10. *EmployeeNumber*: atributo numérico que não foi possível identificar para qual finalidade foi usado;
11. *EnvironmentSatisfaction*: atributo numérico que é usado para identificar qual o nível de satisfação com o ambiente de trabalho da pessoa relativa a instância analisada;
    a. 1 – *Low*;
    b. 2 – *Medium*;
    c. 3 – *High*; e
    d. 4 – *Very High*.
12. *Gender*: atributo nominal, indica o gênero da pessoa relativa a instância analisada;
13. *HourlyRate*: atributo numérico, indica uma taxa por hora que não foi identificada para qual finalidade está no *dataset*;
14. *JobInvolvment*: atributo numérico que indica o envolvimento ou quão motivada a pessoa relativa a instância analisada está com seu trabalho;
    a. 1 – *Low*;
    b. 2 – *Medium*;
    c. 3 – *High*; e
    d. 4 – *Very High*.
15. *JobLevel*: atributo numérico que não foi identificado para qual finalidade foi usado na base de dados;

16. *JobRole*: atributo nominal, que indica o cargo ocupado pela pessoa relativa a instância analisada;
17. *JobSatisfaction*: atributo numérico que indica a satisfação com o trabalho atual da pessoa relativa a instância analisada;
18. *MartialStatus*: atributo nominal que indica o estado civil da pessoa relativa a instância analisada;
19. *MonthlyIncome*: atributo numérico que indica o rendimento mensal da pessoa relativa a instância analisada;
20. *MonsthyRate*: atributo numérico que indica a taxa mensal que não foi identificada para qual propósito foi utilizada;
21. *NumCompaniesWorked*: atributo numérico que indica a quantidade de empresas que a pessoa relativa a instância analisada já trabalhou;
22. *Over18*: atributo nominal que indica se a pessoa relativa a instância analisada é maior de 18 anos;
23. *OverTime*: atributo nominal que indica se a pessoa relativa a instância analisada possui sobrecarga;
24. *PercentSalaryHike*: atributo numérico que indica o aumento de salário percentual. Não se tem mais detalhes sobre as condições de coleta desse atributo;
25. *PerformanceRating*: atributo numérico que indica a performance da pessoa relativa a instância analisada;
    a. 1 – *Low*;
    b. 2 – *Good*;
    c. 3 – *Excellent*; e
    d. 4 – *Outstanding*.
26. *RelationshipSatisfaction*: atributo numérico que indica a satisfação com o relacionamento da pessoa relativa a instância analisada;
    a. 1 – *Low*;
    b. 2 – *Medium*;
    c. 3 – *High*; e
    d. 4 – *Very High*.
27. *StandardHours*: atributo numérico que indica a quantidade padrão de horas. Não há informação se esse padrão é semanal ou quinzenal;
28. *StockOptionLevel*: atributo numérico que indica o nível de opção de compra de ações. Não foi identificado as opções que foram usadas para coletar as informações;

29. *TotalWorkingYears*: atributo numérico que indica o total de anos trabalhados da pessoa relativa a instância analisada;
30. *TrainingTimesLastYear*: atributo numérico que indica a quantidade de treinamentos realizados no último ano pela pessoa relativa a instância analisada;
31. *WorkLifeBalance*: atributo numérico que indica o nível de balanceamento entre vida pessoal e trabalho da pessoa relativa a instância analisada;
32. *YearsAtCompany*: atributo numérico que indica a quantidade de anos que a pessoa relativa a instância analisada está trabalhando na empresa atual;
33. *YearsInCurrentRole*: atributo numérico que indica a quantidade de anos da pessoa referente a instância analisada no cargo atual;
34. *YearsSinceLastPromotion*: atributo numérico que indica a quantidade de anos da pessoa relativa a instância analisada desde a última promoção recebida;
35. *YearsWithCurrManager*: atributo numérico que indica a quantidade de anos que a pessoa relativa a instância analisada está trabalhando com o gerente atual.

Após tomar conhecimento sobre todas as informações disponibilizadas no *dataset* fica mais claro entender as alterações realizadas para geração dos modelos nos treinamentos. O objetivo do conjunto de dados disponibilizado é identificar através do atributo alvo, chamado "*Attrition*", se uma instância possui ou não atrito ou desgaste no trabalho atual (Kaggle, 2017).

A base de dados chamada de produção contém as informações coletadas de pessoas, via questionário, conforme determina a metodologia científica para o tipo de pesquisa de levantamento ou *survey* (Prodanov & Freitas, 2013). A pesquisa foi elaborada no idioma nativo onde o questionário foi aplicado, ou seja, em português brasileiro. Contudo, toda pesquisa foi baseada no conjunto de dados obtidos da IBM elaborado por cientistas de dados da IBM (IBM Watson Analytics, 2015). As respostas obtidas através do questionário eletrônico aplicado, correspondem as mesmas informações que constam no *dataset* da IBM. Nesse contexto se apresenta o primeiro problema encontrado no projeto.

Nem todas as informações que constam no *dataset* da IBM fazem sentido no contexto onde o questionário foi aplicado. Um exemplo que retrata esse problema é o atributo *StandardHours* que corresponde ao padrão de horas de trabalho da instância. Porém no contexto do *dataset* da IBM não foi possível identificar se o padrão era correspondente a uma única semana, duas semanas ou mensal. De qualquer forma, esse atributo foi mantido na base para forma de comparação. Um atributo que demonstra outro problema é o *HourlyRate* que significa uma taxa diária, porém não foi possível reconhecer com qual atributo ele se assemelha no contexto onde o questionário foi aplicado e por esse motivo uma das soluções para que ambas

as bases ficassem equivalentes foi eliminar os atributos que não foram possíveis coletar via questionário. Os atributos excluídos do *dataset* da IBM foram *DailyRate, EmployeeNumber, HourlyRate, MonthlyRate, PercentSalaryHike, StockOptionLevel, JobLevel* e *JobRole*. Assim, ambos os *datasets* serão compatíveis tanto na execução dos modelos de treinamento quanto no de produção.

Adicionalmente, outra questão verificada foi padronizar as respostas coletadas pelo questionário elaborado no mesmo formato existente na base da IBM, isso significa que as informações textuais dos atributos nominais, tiveram que ser traduzidas para o inglês, garantindo compatibilidade com o *dataset* de treinamento.

5.2 Criação dos modelos de aprendizado de máquina

Existem vários algoritmos que podem ser utilizados para diversas funções na área de aprendizado de máquina. Nesta seção serão abordados os algoritmos apresentados no capítulo 2.4 e suas subseções. Neste trabalho foi utilizado o software Weka, um aplicativo desenvolvido pela Universidade de Waikato na Nova Zelândia e que possui uma coleção de algoritmos de *machine learning* para tarefas de mineração de dados e aprendizado de máquina. As funções do Weka podem ser utilizadas em um *dataset* existente ou chamando os algoritmos através do próprio código em Java. O Weka é um aplicativo de código aberto sob a licença GNU *General Public License*.

Foi necessário transformar o conjunto de dados de produção em formato CSV, abrir o arquivo no Weka e salvar no formato padrão ARFF, essa transformação não interfere na leitura dos dados. A base utilizada possui 35 e foi reduzia a 27 atributos sendo um deles o atributo alvo a ser previsto pelo modelo de aprendizado de máquina que obtiver o melhor desempenho. Nessa primeira execução nenhuma alteração foi feita nos atributos, apenas foi aplicado o algoritmo J48. O atributo escolhido para predição no modelo de treinamento foi o "*Attrition*" que é um atributo nominal já explicado na apresentação dos dados do *dataset*. Dentro das opções de testes do Weka, a opção chamada *Use training set* indica que todo o conjunto de dados será utilizado para fazer o treinamento do modelo, foi a primeira a ser testada.

5.2.1 Execução da árvore de decisão J48 ou C4.5

O resultado obtido na classificação correta foi de 92,38% e 7,62% de instâncias incorretamente classificadas. Esse percentual é bem alto e se candidata a ser um dos modelos usados para aplicar no conjunto de dados de produção. A Figura 12 mostra os parâmetros

default que foram usados para a primeira execução. Essa execução gerou uma árvore com 66 folhas e de tamanho 120. Executando o mesmo algoritmo no modo "*Cross-validation*" a classificação correta foi de 82,86% e a classificação incorreta foi de 17,14%. O tamanho da árvore não foi aletrado em relação a execução realizada no modo *use training set.* O método *Cross-validation* é quando o Weka utiliza parte da base de dados como treinamento e parte como dados de teste (Bell, 2015). Na terceira tentativa, usando "*Cross-validation*", o parâmetro "minNumObj", que é o número mínimo de instâncias por folha da árvore, foi alterado de 2 para 4 e o seu desempenho no tipo de teste "*Cross-validation*" foi maior, atingindo a marca de 83,67% de classificação correta. Para essa tentativa, o modelo da árvore teve um desempenho melhor e também teve o tamanho da árvore reduzida, o número de folhas foi de 32 e o tamanho total da árvore foi de 58.

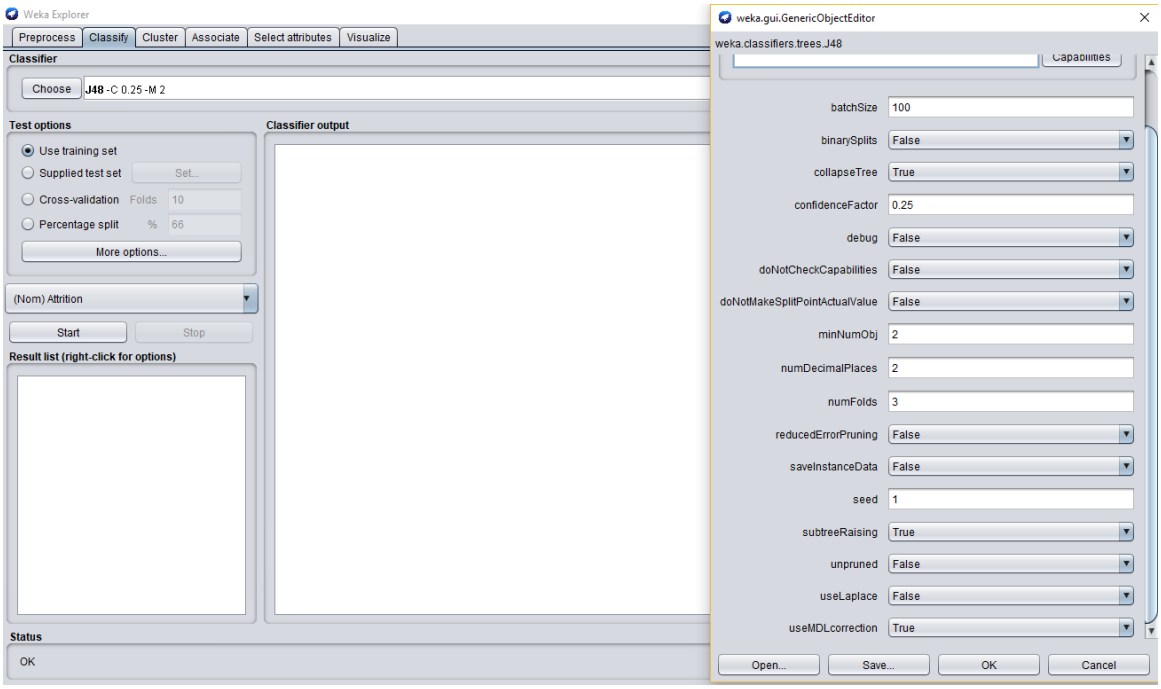


Figura 12 - Configuração utilizada para o J48.
Fonte: Elaborado pelo autor.

Uma matriz de confusão exibe a classificação das instâncias de suas classes reais e previstas mostrando a qualidade do modelo atual e indicando quantas instâncias foram classificadas como verdadeiro positivo, verdadeiro negativo, falso positivo ou falso negativo. Para entender melhor a representação da matriz de confusão no contexto deste trabalho, uma instância classificada como verdadeiro positivo indica que o modelo de *machine learning* previu que essa instância seria classificada como *Yes* no atributo *Attrition* indicando que ela possui atrito, ou seja, o modelo acertou a predição do atributo. Já uma instância classificada como falso positivo, indica que o modelo previu que ela não possui atrito indicando a resposta *No* para o atributo *Attrition*, porém, no *dataset* esse atributo estava registrado como *Yes*, isso

significa que o modelo previu que uma pessoa não teria atrito quando na verdade ela teve atrito. Uma instância classificada como verdadeiro negativo indica que uma pessoa não teve atrito e o modelo classificou ela como *Attrition* igual a *No*, acertando a predição. Já uma instância classificada como falso negativo indica que uma instância foi classificada como se tivesse atrito quando na verdade não possuiu atrito (Microsoft Developer Network, 2017).

Analisando o resultado apresentado na execução *Cross-validation* do algoritmo J48, é possível verificar na matriz de confusão o número de instâncias classificadas corretamente com atrito foi de 58, as outras 179 instâncias que deveriam ter sido classificados com atrito foram classificados como sem atrito, isso se chama de falso positivo. Já o número de pessoas classificadas corretamente sem atrito foi de 1.172 enquanto as pessoas classificadas com atrito e que deveriam ser classificadas sem atrito foi de 61 para a melhor performance. Existe uma falta de balanceamento nos dados desse algoritmo, dado que quase 84% dos resultados para o atributo *Attrition* foram classificados como *No* e o restante como *Yes*. Por fim, foi possível concluir que o melhor desempenho usando J48 no conjunto de dados de teste foi sem alterar qualquer parâmetro do *dataset*, alterando apenas o parâmetro "numMinObj" de 2 para 4 e executando no modo *Cross-validation*.

a	b	<= Classificados como
58	179	a = Yes
61	1172	b = No

Figura 13 - Matriz de confusão J48.
Fonte: Elaborado pelo autor.

5.2.2 Execução do modelo LMT

A segunda execução realizada no *dataset* de treinamento foi usando a LMT. Os índices obtidos com esse modelo sem alterar os parâmetros *default* foram de 88,03% de classificações corretas e 11,97% de classificações incorretas no método *Cross-validation*. O índice de classificação correta teve um melhor desempenho que o modelo J48. A LMT se candidata como um dos modelos mais adequados a serem utilizados na base de produção até esse momento. Analisando a matriz de confusão é possível ver que a quantidade de itens falsos negativos e os itens falsos positivos são os menores entre os modelos executados. Além disso, existe um alto índice de acerto nos verdadeiros positivos e nos verdadeiros negativos. Conclui-se que esse modelo tem uma capacidade muito grande de assertividade para essa base de dados apesar de ser mais lento que o J48.

a	b	<= Classificados como
84	153	a = Yes
23	1210	b = No

Figura 14 - Matriz de confusão LMT.
Fonte: Elaborado pelo autor.

Fazendo alguns ajustes no *dataset* foi possível aumentar o percentual do índice de acerto da LMT. A diferença percentual foi pequena de 88,03% para 88,36%. Já em relação aos verdadeiros positivos, aumentaram de 84 para 86, e os verdadeiros negativos diminuíram de 153 para 151. Esse aumento é significativo já que melhorou a performance de predição. A mudança que gerou essa melhoria foi a troca do atributo *Education* de numérico para nominal, isso transformou o atributo em uma classe e melhorou o desempenho do modelo. A matriz de confusão mostra o aumento da classificação dos itens verdadeiros positivos devido a alteração no atributo de numérico para nominal.

a	b	<= Classificados como
85	151	a = Yes
20	1213	b = No

Figura 15 - Matriz de confusão LMT.
Fonte: Elaborado pelo autor.

Com o objetivo de melhorar a performance do modelo de predição LMT, uma nova tentativa foi executada. A partir do *dataset* em seu estado inicial o seguinte ajuste foi feito, o atributo *JobSatisfaction* foi convertido de numérico para nominal. A nova execução gerou um índice de classificações corretas de 88,84%. O resultado com esse *dataset* ajustado teve um aumento em relação a execução anterior. A matriz de confusão mostra que os itens verdadeiros positivos subiram de 86 para 93. As execuções realizadas nesse conjunto de teste foram executadas no modo *Cross-validation*. Na Figura 16 é possível identificar as mudanças ocorridas na matriz de confusão.

a	b	<= Classificados como
93	144	a = Yes
20	1213	b = No

Figura 16 - Matriz de confusão LMT.
Fonte: Elaborado pelo autor.

Devido ao fato do *dataset* não estar balanceado, isso significa dizer que a proporção entre as predições do atributo *Attrition* como *Yes* são muito menores que as predições do tipo *No*, e com os processamentos que foram feitos, os acertos na predição do resultado *Yes* foram elevados para um total de 93. Esse foi o maior valor atingido entre as execuções realizadas. Por

fim, foram realizadas outras tentativas de ajustes visando aumentar o desempenho, mas o percentual de 88,84% foi o maior obtido para a LMT.

5.2.3 Execução do modelo MLP

Executando o MLP pela primeira vez sem fazer qualquer pré-processamento nos dados do *dataset* o resultado obtido foi um índice de classificação de instâncias corretas de 84,01% que é um valor baixo se comparado ao percentual obtido com a execução da LMT ou *logistic model tree*. Porém, nessa execução nenhuma alteração foi realizada no *dataset.* A Figura 17 mostra a matriz de confusão da primeira execução usando o MLP.

a	b	<= Classificados como
91	146	a = Yes
89	1144	b = No

Figura 17 - Matriz de confusão MLP.
Fonte: Elaborado pelo autor.

O atributo *JobSatisfaction* foi modificado de numérico para nominal para que fosse possível comparar com o resultado do LMT. Mudando o mesmo atributo de numérico para nominal e executando o modelo no modo *Cross-validation*, a primeira percepção é que a execução do MLP em relação a LMT é o tempo de processamento do resultado. O MLP demorou muito mais tempo do que a LMT, sendo que a LMT levou menos de 3 segundos para montar o modelo e o MLP levou 20 segundos para montar o modelo. O resultado obtido foi de 85,71% para classificações corretas, é uma melhora significativa se comparada a primeira execução.

Em uma tentativa de melhorar a performance do modelo, os seguintes atributos foram convertidos de numérico para nominal: *Education*, *EnvironmentSatisfaction*, *JobInvolvment*, *JobSatisfaction*, *PerformanceRating*, *RelationshipSatisfaction*, *WorkLifeBalance* e *YearsWithCurrManager*. Após essas alterações uma nova execução foi realizada e um novo índice foi obtido, melhor que o anterior. O índice de classificações corretas chegou a 86,05%. A matriz de confusão exibida na Figura 18 mostra que os verdadeiros positivos do modelo MLP tiveram melhor performance se comparado a LMT, 116 e 93 respectivamente. Em relação aos falsos positivos a relação foi de 1213 para a LMT e 1149 para o MLP. Essa diferença foi fundamental para que o percentual obtido no MLP tenha sido menor que o percentual obtido na melhor execução da LMT.

a	b	<= Classificados como
116	121	a = Yes
84	1149	b = No

Figura 18 - Matriz de confusão MLP (melhor execução).
Fonte: Elaborado pelo autor.

A partir desta execução conclui-se que o melhor percentual obtido com o MLP estava em seu limite e que não haveriam mais alterações que poderiam melhorar o desempenho desse modelo. Apesar do modelo de árvore de decisão LMT ter obtido um melhor desempenho em relação ao MLP, é notável que a classificação dos itens verdadeiros positivos, na matriz de confusão, pertencentes ao MLP de melhor desempenho são superiores do que as classificações dos mesmos itens se comparado a LMT. Essa diferença deve ser levada em consideração já que o modelo não é balanceado. Concluo que a execução dos dois modelos, o melhor LMT e o melhor MLP serão utilizados na base de produção que contém os dados obtidos através de questionário elaborado pelo autor deste trabalho. Para que isso seja possível, os modelos das melhores performances foram salvos a partir do Weka para que sejam aplicados no *dataset* de produção.

5.3 Dados coletados

Este subcapítulo, apresenta toda a preparação da base de dados ou *dataset* que foi coletado pelo autor através de questionário, além de demonstrar os resultados obtidos na aplicação dos modelos de aprendizado de máquina LMT (*Logistic Model Tree*) e o MLP também conhecido como *Multlayer Perceptron*. Ambos modelos possuem características diferentes, a LMT é um modelo de árvore de decisão e nas execuções apresentadas no subcapítulo 5.2.2 obteve o melhor desempenho em relação ao percentual de assertividade das classificações corretas, obtendo um índice de 88,84%. Já o MLP é um modelo RNA ou rede neural artificial, e nas execuções apresentadas no subcapitulo 5.2.3 obteve o segundo melhor desempenho dos modelos testados, atingindo um percentual de 86,05%. Apesar do percentual ser menor que o LMT a inclusão desse modelo dar-se-á pela sua alta taxa de classificação de verdadeiros positivos na matriz de confusão visto que é um modelo não balanceado.

Com os modelos usados no *dataset* de produção definidos, esta seção apresenta os resultados obtidos com as execuções realizadas. As respostas obtidas via questionário digital, foram tabuladas em uma planilha de eletrônica. Os títulos para as respostas foram incluídos manualmente e seguiram o mesmo padrão utilizado no conjunto de dados da IBM. Todas as respostas foram padronizadas com a inclusão do caractere *underline* para que não possuíssem espaços, além disso os termos foram traduzidos para o inglês. Os atributos que possuem valores com casas decimais tiveram as vírgulas das casas decimais trocadas por pontos. Essa troca foi feita porque o arquivo foi salvo em formato CSV que justamente separa as informações por

vírgula. Na coluna *Atrittion*, diferentemente do *dataset* da IBM, não possui valores no conjunto de dados de produção, isso porque o objetivo do trabalho é justamente aplicar os modelos criados para prever o valor dessa coluna no novo *dataset*. Ao invés de ter uma informação na coluna *Atrittion*, foi colocado um ponto de interrogação "?", essa indicação caracteriza para o Weka que, ao executar o modelo nesse conjunto de dados, essa é a coluna que desejamos prever. Por fim, o arquivo CSV processado com os dados coletados, foi aberto no ARFF Viewer do Weka para verificar se estava em conformidade, e foi salvo uma versão no formato padrão do Weka que é ARFF.

O conjunto de dados de produção possui 28 instâncias que foram coletadas via questionário eletrônico, porém 29 pessoas responderam o questionário. Uma das instâncias não foi incluída no *dataset* porque não estava empregado e um dos critérios escolhidos pelo autor desse trabalho era que as pessoas que respondessem o questionário estivessem empregadas. As pessoas que responderam tinham entre 22 e 51 anos de idade. 23 pessoas eram do sexo masculino e 6 do sexo feminino. 14 pessoas informaram que seu estado civil era solteiro, 11 eram casadas e 4 indicaram ser divorciadas. Em relação a escolaridade, 58,6% dos questionados responderam ter ensino superior, como curiosidade uma das pessoas possuía mestrado e outra doutorado. Em relação a área de estudo, 72,4% das pessoas responderam estudam na área de tecnologia. Essas são as informações gerais sobre os dados coletados. As pessoas que responderam o questionário estão no mercado de trabalho entre 4 e 22 anos e o número total de empresas que trabalharam vai de 1 até 10.

Mais de 60% das pessoas trabalham na área de TI. 12 das 29 pessoas não viajam a trabalho, 13 viajam raramente e 3 viajam frequentemente. A renda mensal variou entre pouco mais de 1.000 reais a mais baixa até 9.500 reais a mais alta. 9 pessoas estão em seus empregos a mais de 5 anos as outras 19 pessoas estão em seus empregos entre 1 e 5 anos. 57,1% das pessoas possui um padrão de 40 horas de trabalho semanais e 32,1% possuem um padrão de 44 horas semanais e o restante das pessoas trabalha menos de 40 horas semanais e apenas uma pessoa possui um padrão de mais de 44 horas semanais. Em relação a carga de trabalho realizada semanalmente, 50% das pessoas informaram que trabalham entre 44 e 55 horas por semana, 8 pessoas informaram que trabalham até 44 horas semanais, 3 pessoas trabalham até 40 horas semanais, 1 pessoa trabalha até 30 horas semanais e 1 pessoa trabalha acima de 55 horas semanais. 35,7% das pessoas informaram que se consideram sobrecarregados.

Mais da metade, 15 pessoas, informaram que sua satisfação com o trabalho atual é média, 6 pessoas informaram que sua satisfação com o trabalho é alta, 3 disseram que é muito alta e 4 informaram que sua satisfação é baixa. 11 pessoas informaram que seu envolvimento ou

engajamento com o trabalho atual é alto, 4 pessoas informaram que seu envolvimento é muito alto a mesma quantidade informou que seu envolvimento é baixo e 9 pessoas informaram que seu nível de envolvimento é mediano. Em relação ao desempenho18 pessoas informaram ter um bom desempenho no seu trabalho atual, 6 indicaram ter um excelente desempenho, 2 indicaram ter um desempenho além das expectativas e outras 2 indicaram ter um baixo desempenho. Em relação a satisfação com o relacionamento no trabalho 42,9% das pessoas indicaram ter um nível alto de satisfação com o relacionamento, 35,7% indicaram ter uma satisfação média, 10,7% indicaram ter uma satisfação muito alta e outros 10,7% indicaram ter uma satisfação baixa.

O balanceamento entre a vida pessoal e a vida profissional também foi avaliada, 18 pessoas indicaram que possuem um bom balanceamento, 6 pessoas indicaram um balanceamento ruim entre vida e trabalho, 3 indicaram que tem um ótimo balanceamento e 1 indicou que tem um balanceamento excelente. Em relação ao ambiente de trabalho, metade das pessoas informaram que possuem satisfação média, 10 pessoas informaram uma alta satisfação com ambiente, 3 pessoas possuem uma satisfação muito alta e 1 possui satisfação baixa. 10 pessoas estão a mais de cinco anos exercendo a mesma função e o restante está até cinco anos exercendo a mesma função, dessas, 6 estão a um ano na mesma função e cinco estão a três anos na mesma função. 25 pessoas indicaram que estão trabalhando até 4 anos com o mesmo gestor, os demais estão trabalhando com o mesmo gestor a mais de 4 anos. Em relação à última promoção, 10 pessoas indicaram que faz um ano desde a última promoção recebida o restante indica que faz mais de um ano que não recebe uma promoção.

Analisando as respostas que foram dadas concluí que nem todas as pessoas estão totalmente satisfeitas com seus trabalhos atuais e isso pode ser um sinal de que exista um atrito ou desgaste desses profissionais, o que poderá gerar predições positivas para o atributo *Attrition*.

5.4 Execução dos modelos

Depois de ter um *overview* sobre os dados coletados, será apresentado os resultados das execuções no *dataset* de produção. O primeiro procedimento a ser realizado é carregar o modelo de melhor desempenho do LMT e do MLP. Após carregar o modelo deve-se abrir o arquivo ARFF que contém as 28 instâncias que serão analisadas. Após esse procedimento, o resultado das predições foi configurado com a opção "*PlainText*" para gerar o resultado em formato texto. Adicionalmente, na seção *Teste options* é preciso selecionar a opção *Supplied test set*. No botão

"*Set...*" é necessário selecionar o conjunto de dados e selecionar o atributo classe. Após os preparativos, é preciso executar o modelo no conjunto de dados, para isso deve-se clicar com o botão direito sobre o modelo carregado e selecionar a opção *Re-evaluate model on current test set*. Ao executar essa opção o modelo será aplicado sobre o conjunto de dados. A Figura 19 mostra uma visão geral do resultado e das opções selecionadas para a execução dos modelos.

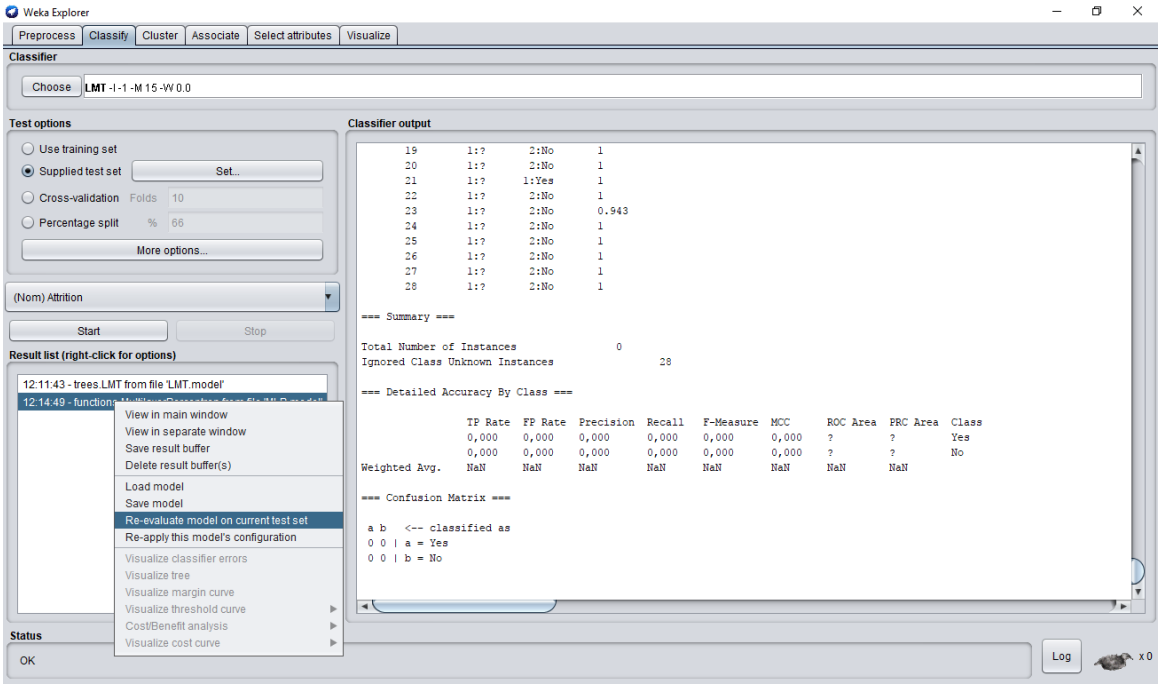


Figura 19 - Configuração para execução do modelo LMT.
Fonte: Elaborado pelo autor.

5.5 Análise dos resultados

A Figura 20 mostra uma compilação dos resultados obtidos na execução do modelo LMT ou *Logistic Model Tree*. O modelo LMT previu que 5 instâncias possuem atrito ou desgaste isso corresponde a 17,86% do conjunto de dados. As outras 23 instâncias foram previstas como sem atrito ou desgaste. Em cinco casos a probabilidade de o modelo estar certo em sua predição foi igual a 1 em uma escala de 0 a 1.

Logistic Model Tree - LMT				
Instância	Atual	Previsto	Erro	Probabilidade
1	1:?	1:Yes		0.863
2	1:?	2:No		0.792
3	1:?	2:No		0.988
4	1:?	2:No		0.996
5	1:?	2:No		0.924
6	1:?	2:No		0.845
7	1:?	2:No		0.528
8	1:?	2:No		0.998
9	1:?	2:No		0.985
10	1:?	2:No		0.915
11	1:?	1:Yes		0.972
12	1:?	2:No		0.999
13	1:?	1:Yes		0.946
14	1:?	2:No		0.999
15	1:?	1:Yes		0.939
16	1:?	2:No		0.982
17	1:?	2:No		0.696
18	1:?	2:No		1
19	1:?	2:No		0.921
20	1:?	2:No		0.565
21	1:?	1:Yes		0.656
22	1:?	2:No		0.997
23	1:?	2:No		0.821
24	1:?	2:No		1
25	1:?	2:No		1
26	1:?	2:No		0.981
27	1:?	2:No		1
28	1:?	2:No		1

Figura 20 - Resultados obtidos para o LMT.
Fonte: Elaborado pelo autor.

A Figura 21 mostra a compilação dos resultados gerados pela execução do modelo MLP ou *Multlayer Perceptron*. Portanto, 23 instâncias foram classificadas como sem atrito ou desgaste e outras cinco instâncias foram classificadas como atrito ou com desgaste (*Attrition = No*) o que significa um percentual 17,86% (5 instâncias) e 82,14% (23 instâncias) como possuindo desgaste ou atrito (*Attrition = Yes*) ou seja, exatamente o mesmo resultado previsto pelo modelo LMT. Em 25 instâncias a probabilidade de o modelo estar certo em sua predição foi igual a 1 em uma escala de 0 a 1, desses 25 casos 5 foram classificados que possuem atrito ou desgaste e as outra s 20 como sem atrito ou desgaste.

Multilayer Perceptron - MLP				
Instância	Atual	Previsto	Erro	Probabilidade
1	1:?	1:Yes		1
2	1:?	2:No		1
3	1:?	2:No		1
4	1:?	2:No		1
5	1:?	2:No		1
6	1:?	2:No		0.999
7	1:?	2:No		1
8	1:?	2:No		1
9	1:?	2:No		1
10	1:?	2:No		1
11	1:?	1:Yes		1
12	1:?	2:No		1
13	1:?	1:Yes		1
14	1:?	2:No		1
15	1:?	1:Yes		1
16	1:?	2:No		1
17	1:?	2:No		0.963
18	1:?	2:No		1
19	1:?	2:No		1
20	1:?	2:No		0.609
21	1:?	1:Yes		1
22	1:?	2:No		1
23	1:?	2:No		1
24	1:?	2:No		1
25	1:?	2:No		1
26	1:?	2:No		1
27	1:?	2:No		1
28	1:?	2:No		1

Figura 21 - Resultados obtidos para MLP.
Fonte: Elaborado pelo autor.

Analisando o resultado de ambos os modelos, houve a mesma correspondência das instâncias classificadas, isso significa dizer que ambos modelos classificaram a mesma quantidade de instâncias como *Yes* e como *No*. Além desse fator, é importante destacar que ambos os modelos previram atrito para as mesmas instâncias, são elas as instâncias de número 1, 11, 13, 15 e 21. O LMT possui uma taxa de confiança na predição que é menor que o MLP visto que em 25 casos o MLP teve uma probabilidade igual a 1 de o modelo estar certo em sua predição contra 5 instâncias no modelo LMT. Essa diferença mostra que o modelo RNA, através do *Multilayer Perceptron* (MLP) teve maior confiança nas suas predições e por esse motivo, se mostrou o modelo mais assertivo nessa execução. A proximidade entre os resultados apresentados por ambos os modelos permite a conclusão de que as predições podem estar

corretas, mas essa correlação será analisada mais adiante quando os resultados previstos forem comparados com o padrão de respostas das pessoas que responderam ao questionário.

É importante fazer uma relação dos resultados obtidos no *dataset* de produção com os dados coletados durante o processo de treinamento e montagem do modelo. Apesar da diferença de instâncias em cada conjunto de dados, sendo 1.470 instâncias para o *dataset* da IBM e apenas 28 para o *dataset* do autor, uma característica ficou evidente em ambos os modelos, a falta de balanceamento. No conjunto de dados da IBM, usado para treinar o modelo, 237 instâncias estavam classificadas como *Yes* para o atributo *Attrition*, isso representa 16,12% de todas as instâncias do conjunto, ou seja, há muito mais instâncias classificadas como sem atrito do que com atrito. Na execução do *dataset* do autor, uma proporção muito parecida com a do conjunto de treinamento foi obtida, ou seja, 17,86% do conjunto foi classificado como *Yes* para o atributo *Attrition*.

Outra análise que deve ser realizada e que pode ser mais esclarecedora para o entendimento das instâncias que foram classificadas com atrito ou desgaste, é comprar as mesmas com suas respostas no questionário realizado. Das cinco instâncias que foram indicadas com atrito ou desgaste, três informaram que não viajam a trabalho, uma viaja raramente e um viaja frequentemente. Apenas uma das cinco instâncias mora a mais de 10 quilômetros do local de trabalho, logo considero que esse fator não me parece ser decisivo para que haja desgaste ou atrito no trabalho. 80% das pessoas (4 instâncias) indicadas com desgaste ou atrito são do sexo masculino e 20% (1 instância) do sexo feminino além disso 80% delas possuem mais de 25 anos de idade sendo que as 5 não são casadas o que pode indicar que não há uma grande preocupação com o fator estabilidade, mas isso é apenas uma hipótese. Dessas seis pessoas classificadas com desgaste ou atrito, 60% possuem rendimentos até 2 mil reais mensais o que pode ser um fator relevante quando se fala de desgaste inclusive relacionado ao fator de que 60% dessas pessoas está a mais de dois anos sem uma promoção no trabalho. Apenas uma pessoa indicou ter uma satisfação alta com seu ambiente de trabalho, as outras quatro indicaram ter uma satisfação média com o ambiente em que trabalham, esse fator pode ter uma forte relação com a possibilidade de haver atrito ou desgaste desses profissionais. 80% (4 instâncias) informaram que seu nível de engajamento com o trabalho atual é médio e 20% (1 instância) informou que seu engajamento é baixo, outro fator relevante para gerar desgaste no profissional. Em relação a satisfação com o trabalho atual, 3 indicaram possuir uma baixa satisfação com seu trabalho atual e 2 indicaram ter um nível médio de satisfação com o trabalho. 80% (4 instâncias) responderam que se sentem sobrecarregados no emprego atual. O fator de sobrecarga de trabalho é considerado por Pencavel (2014), citado no capítulo 1 subcapítulo 1.4,

um fator crítico que leva os funcionários a deixarem seus empregos por muitas vezes acreditarem que estão sendo penalizados por seus bons desempenhos. Em 40% (2 instâncias), as pessoas classificaram como baixo o seu nível de satisfação com o relacionamento na sua empresa, 40% classificou como média a sua satisfação e apenas 20% (1 instância) classificou como alta a sua satisfação. Por fim ao que se refere ao balanceamento entre vida pessoal e trabalho, 3 pessoas classificaram que essa relação é ruim e 2 informaram que essa relação é apenas boa.

Analisando todos os fatores supracitados, é perfeitamente aceitável crer que a predição de ambos os modelos está correta. Um exemplo disso é que outras duas instâncias informaram no questionário que sua satisfação com o trabalho atual é baixa, porém ambos possuem rendimentos mensais que superam a casa dos 6 mil reais, e sabendo que o salário é um fator muito importante para um profissional é possível entender que essas duas instâncias não foram classificadas com atrito apesar de possuírem baixa satisfação. Além do fator salário, essas duas instâncias citadas informaram que não se sentem sobrecarregadas no trabalho atual, contra 80% das pessoas que os modelos previram com desgaste ou atrito. Novamente, indico que em uma opinião do autor, as previsões feitas pelos modelos estão corretas e que essas pessoas indicadas com atrito ou desgaste podem estar em estado crítico na sua relação com o seu trabalho atual. É justamente para prever essas situações que a aplicação de *machine learning* pode ajudar equipes de RH ou outros interessados a manter ou trocar essas pessoas que estão com atrito. Quanto mais exemplos reais forem utilizados para treinar os modelos, mais preciso eles irão se tornar. No próximo capítulo, será apresentada as conclusões tiradas dessa experiência e quais aplicações podem ser feitas com o modelo gerado.

CONCLUSÃO

Com as análises realizadas neste trabalho, várias conclusões foram encontradas sobre todo o processo de aprendizado de máquina aplicado a gestão de pessoas. A primeira diz respeito ao processo de coleta dos dados. Houve grande dificuldade de encontrar voluntários para a pesquisa com o questionário, pois nem todas as pessoas sentem-se à vontade para responder informações pessoais e sobre seus empregos, principalmente quando essas informações podem ser negativas. Contudo, grande parte das informações que foram coletadas pelo questionário, construído nesse trabalho, as empresas já possuem e muitas vezes não fazem boa utilização delas. Outro fator complicado neste trabalho foi pelo fato do *dataset* conter dados fictícios e que correspondem a outro país. No momento de traduzir os atributos para elaborar o questionário foi preciso analisar profundamente o contexto de cada atributo para que ele tivesse o mesmo sentido no questionário em português e para que o modelo mantivesse a mesa consistência sendo executado em ambos os conjuntos de dados.

Outro problema vivenciado é em relação ao conjunto de dados para treinamento dos modelos. Obter dados históricos reais, relacionados ao assunto que se deseja estudar e de domínio público é uma tarefa extremamente difícil mesmo com a existência de sites que disponibilizam dados para estudos. Esse problema ocorreu com este trabalho dado que inicialmente, a ideia era prever as habilidades e a compatibilidade dos funcionários com seus cargos. Porém encontrar um *dataset* com um conjunto de instâncias significativo para treinamento é muito difícil e somado a dificuldade de coletar dados das pessoas via questionário ou entrevista exigiria um tempo de pesquisa muito maior do que o disponível para a execução deste trabalho.

Como já citado, as empresas possuem grande parte das informações usadas nesse trabalho e aquelas que não possuem, podem ser obtidas facilmente através de avaliações de desempenho e de clima organizacional. Através dessas ferramentas uma empresa pode coletar dados valiosos para acompanhar seus colaboradores e construir uma base histórica. A partir dessa base histórica é possível criar modelos de treinamento de aprendizado de máquina para avaliar os fatores que foram analisados nesse trabalho. Além do desgaste é possível prever outras informações como a satisfação com o trabalho ou mesmo se um funcionário pode estar com sobrecarga de trabalho.

Em relação a criar os modelos de aprendizado de máquina, é possível concluir que não é uma tarefa muito difícil de ser fazer, ainda mais contando com a ajuda de ferramentas como

o Weka. É necessário ter um bom *dataset* de treinamento e gastar muito tempo realizando alterações nesse *dataset* para verificar em quais condições os modelos geram as melhores predições. A parte mais complicada desse processo é organizar os resultados que vão sendo obtidos juntamente com as alterações realizadas no *dataset*. Uma planilha onde se registra essas informações além de armazenar o resultado dos modelos em arquivos de texto podem ajudar no controle das execuções. Ainda em relação as ferramentas utilizadas nesse trabalho, especificamente o Weka, é possível concluir que é um software muito fácil de utilizar que abrange muitos algoritmos de *data mining* e *machine learning*, além de possuir muito material disponível na internet e de ser amplamente utilizado em universidades para ensino acadêmico.

Em relação aos resultados obtidos nesse trabalho, é possível concluir que eles foram consistentes com os resultados obtidos pelos modelos de aprendizado de máquina no *dataset* de treinamento. Além disso, quando comparada as predições para pessoas que podem possuir desgaste ou atrito, com as respostas que deram no questionário elaborado para coletar as informações, é possível identificar fatores que de fato podem contribuir para que uma pessoa se sinta desgastada com o trabalho atual. Adicionalmente, vale informar que as predições realizadas pelos modelos MLP e LMT são apenas previsões e que não há como saber de fato se essas pessoas estão desgastadas com o trabalho atual. Por fim posso concluir que no ponto de vista do autor desse trabalho as predições realizadas nos dados coletados estão consistentes e podem indicar de fato um desgaste real para as pessoas que responderam o questionário mostrando que essa pode ser uma ferramenta poderosa para a tomada de decisão dos setores que cuidam da retenção e gestão das pessoas nas organizações.

Para encerrar, os insights obtidos nesse trabalho deixam abertura para que outros estudos e trabalhos possam ser realizados. Entre eles o mais palpável é a construção de um software que armazene, gerencie e preveja o desgaste ou atrito dos colaboradores das empresas, podendo auxiliar os times de RH e gestão de pessoas nas tomadas de decisão. Outra possibilidade é a de melhorar os modelos criados utilizando uma base real de treinamento de uma organização de médio ou grande porte para validar a consistência dos modelos e tirar uma prova real se ele é de fato efetivo. Enfim, muitas ideias podem surgir a partir dos questionamentos e problemas que as equipes de RH e gestão de pessoas enfrentam no dia a dia. Tanto a área de inteligência artificial como sua ramificação de *machine learning* estão se consolidando no auxílio da solução para problemas complexos. No entanto uma afirmação é certa, dados consistentes valem muito e fazem toda a diferença na hora de utilizar aprendizado de máquina e obter resultados consistentes.

REFERÊNCIAS BIBLIOGRÁFICAS

Bell, J. (2015). *Machine Learning: hands-on for developers and technical professionals.* Indianapolis: Wiley.

Bichuetti, J. L. (2015, Maio Não informado). *Gestão de pessoas não é com o RH!* Retrieved 04 29, 2017, from Harvard Business Review: http://hbrbr.uol.com.br/gestao-de-pessoas-nao-e-com-o-rh/

BRYNJOLFSSON, E., & MCAFEE, A. (2016). *The Second Machine Age: work, progress and prosperity in a time of brilliant technologies.* New York: W. W. Norton & Company.

Chiavenato, I. (2008). *Gestão de Pessoas.* São Paulo: Elsevier Editora Ltda.

Chowdhurry, S. (2003). *A era do Talento: Obtendo alto retorno sobre o talento.* São Paulo: Pearson Education.

Dino. (2016, Agosto 16). *Dino*. Retrieved from Terra: https://noticias.terra.com.br/dino/mercado-de-ti-tem-50-mil-vagas-a-serem-preenchidas,78ff21558d76bee1bc0e21b5c69768fcjp7gaatr.html

Djuris, J., Medarevic, D., Krstic, M., Vasiljevic, I., Masic, I., & Ibric, S. (2012, Julho 31). Design Space Approach in Optimization of Fluid Bed Granulation and Tablets Compression Process. *The Cientific World JOURNAL*, pp. 1-10.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008, Agosto 8). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, pp. 1871-1874.

Gil, A. C. (2007). *Gestão de Pessoas: Enfoque nos papéis profissionais.* São Paulo: Editora Atlas S.A.

Gronlund, C. J. (2016, 08 17). *Introdução ao aprendizado de máquina na nuvem*. Retrieved from Microsoft Azure: https://azure.microsoft.com/pt-br/documentation/articles/machine-learning-what-is-machine-learning/

IBGE. (2016, Março 23). *Pesquisa Mensal de Emprego*. Retrieved Novembro 15, 2016, from IBGE - Instituto Brasileiro de Geografia e Estatísticas: ftp://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Mensal_de_Emprego/fasciculo_indicadores_ibge/2016/pme_201602pubCompleta.pdf

IBM Watson Analytics. (2015, Setembro 14). *SAMPLE DATA: HR Employee Attrition and Performance*. Retrieved from IBM:

https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/

IDC Brasil. (2016, Janeiro 28). *IDC Releases*. Retrieved from http://br.idclatin.com/: http://br.idclatin.com/releases/news.aspx?id=1970

Kaggle. (2017, 06 01). *IBM HR Analytics Employee Attrition & Performance*. Retrieved from Kaggle: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Landwehr, N., Hall, M., & Frank, E. (2006, 09 02). Logistic Model Trees. *Kluwer Academic Publishers*, p. 14:21.

Marques, J. R. (2016, 03 03). *Conceito de Gestão de Pessoas*. Retrieved from IBC Coaching: http://www.ibccoaching.com.br/portal/rh-gestao-pessoas/conceito-gestao-de-pessoas/

Martin, J. (2010, Maio). *Leadership Development*. Retrieved from Harvard Business Review: https://hbr.org/2010/05/how-to-keep-your-top-talent

McGlaun, S. (2012, Agosto 23). *Slash Gear*. Retrieved from Slash Gear: https://www.slashgear.com/facebook-data-grows-by-over-500-tb-daily-23243691/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+slashgear+%28SlashGear%29&utm_content=Google+Reader

Microsoft Developer Network. (2017, 06 03). *Matriz de classificação (Analysis Services - Mineração de dados)*. Retrieved from Microsoft Developer Network: https://msdn.microsoft.com/pt-br/library/ms174811(d=printer).aspx

Mitchell, T. M. (2016, setembro 09). *Publications*. Retrieved from Carnegie Mellon University School of Computer Science: http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf

Pencavel, J. (2014, Abril Não informado). The Productivity of Working Hours. *IZA*, pp. 2052-2076.

Prodanov, C. C., & Freitas, E. C. (2013). *Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico.* Novo Hamburgo: Editora Feevale.

Quinlan, R. (1993). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann Publishers.

SAS. (2016, Setembro 19). *Machine Learning: O que é e por que é importante?* Retrieved from SAS: https://www.sas.com/pt_br/insights/analytics/machine-learning.html

Scikit-learn. (2017, Junho 27). *Neural network models (supervised)*. Retrieved from Scikit-learn: http://scikit-learn.org/stable/modules/neural_networks_supervised.html

Turing, A. (1950). Computer machinery and intelligence. *Mind*, pp. 433-460.

Varshney, K. R., Chenthamarakshan, V., Fancher, S. W., Wang, J., Fang, D., & Mojsilović, A. (2014, 08 24). Predicting employee expertise for talent management in the enterprise.

KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 1729-1738.

Wei, D., Varshney, K. R., & Wagman, M. (2015). Optigrow: People Analytics for Job Transfers. *2015 IEEE International Congress on Big Data*, 535-542.