Propósito do sistema de Machine Learning para Gestão de Talentos

O propósito deste sistema foi a criação de um algoritmo de inteligência artificial para averiguar o potencial de um funcionário ser ou se tornar um talento e gerenciar o seu desenvolvimento e retenção. Entendendo-se por talento aquele com desempenho superior ao longo do tempo em relação sob métricas particulares de cada organização.

Adotamos a metodologia de machine learning, a qual objetiva provocar um aprendizado quando exposto a um conjunto de dados. Exige-se que essa base de dados seja grande o suficiente para sustentar o processo de aprendizagem, o qual pode ser aplicado a diversas áreas tais como: saúde, dados financeiros, recursos humanos etc (Baba & Sevil, 2019). O algoritmo foi desenvolvido em linguagem Python 6 na plataforma Jupyter/Anaconda (Disponível no link:https://nbviewer.jupyter.org/github/rdgcdasilva/Jupyternotebook/blob/master/Modelagem_RL%20e%20RF_%20Python_dados%20Melhores.ipynb)

Assim, uma extensa base de dados para o desenvolvimento do sistema foi disponibilizada pela Faculdade FIA de Administração e Negócios. Trate-se da base de respondentes do Guia das Melhores Empresas Para Você Trabalhar, feita em parceria com a revista Você S.A, referente aos dados da edição de 2018. Nesta edição, selecionamos as respostas de 84936 empregados respondentes de 399 empresas que possuíam avaliação de desempenho formal.

Fundamentação teórica

O desenvolvimento do conceito de machine learning, surge por volta da metade da década de 1980, por meio da introdução da teoria de aprendizagem indutiva, a qual objetivava lidar com problemas de aprendizagem ligadas a representação de espaço(Silver, Yang, & Li, 2013).

O uso de métodos de aprendizagem baseados em máquinas vem aumentando recentemente, sendo aplicado em diversos campos, sendo os mais populares os baseados em regressões e classificações para se poder fazer previsões (Baba & Sevil, 2019; Vezza, Muñoz-Mas, Martinez-Capel, & Mouton, 2015).

Esse aumento no uso de aplicações ligadas a aprendizagem baseada em máquinas se deve muito as proposições de aprendizagem de longa duração, surgida durante a metade dos anos 1990, se baseando principalmente na aplicação de redes neurais de aprendizagem, por meio do uso da transferência do conhecimento, por meio de muitas tarefas de aprendizagem

(back-propagation) explorando dessa forma o uso de conhecimento anteriores obtidos em outras tarefas de aprendizagem (Silver et al., 2013).

As modernas técnicas de machine learning tem se mostrado úteis, demonstrando bastante acurácia em diversos setores, tais como predição de risco de crédito, análise de dados geospaciais etc, permitindo dessa forma uma melhoria na tomada de decisões, sendo para isso as técnicas de classificação amplamente utilizadas para esses fins (Arora & Kaur, 2020).

No que tange ao uso de machine learning também está em ascensão o uso de modelos de aprendizagem não supervisionados, tendo como exemplo o modelo ART (Adaptative Resonance Theory) por meio do mapeamento de redes de aprendizagem de baixo para cima na entradas sensoriais dos nós e de cima para baixo com a utilização dos nós de expectativas (ou cluster nodes) (Silver et al., 2013).

Em relação aos métodos utilizados para a aplicação do machine learning, existem diversos algoritmos que podem ser utilizados, citando apenas os métodos de classificação, devido ao foco do artigo, tais como: Nayve Bayes, Redes Neurais, Random Forest, Árvores de Decisão, Regressão Logística, apenas para citar alguns (Arora & Kaur, 2020).

Neste sistema foi escolhida a técnica Random Forest, o qual é base no conceito de árvore de classificação, o qual extrai randomicamente variáveis e dados amostrais, os quais permitem ao algoritmo gerar muitas árvores de classificação, e agregar os resultados dessas classificações (Pan & Zhou, 2019).

O método Random Forest que é um método de criação de classificações foi escolhido por fornecer uma boa acurácia na classificação dos dados, além de possuir algumas vantagens tais como se imune ao efeito de ajuste demasiado *(over adjust*), além de ser rápido, simples e auxiliar na identificação de erros internos, além de possuir uma boa tolerância a *outiliers* e ruídos (Arora & Kaur, 2020).

Uma das características desse método é ser uma ferramenta de aprendizagem que mapeia uma lista de parâmetros de entrada objetivando prever uma resposta, por meio da construção de uma ranqueamento dos parâmetros de entrada mesmo para dados não lineares (Aulia et al., 2019).

O método Random Forest foi proposto por Breiman (2001), sendo um conjunto de métodos que constrói múltiplas árvores de decisão e as junta de forma a se obter uma predição ou classificação mais estável e acurada, por meio da utilização do conceito de re-amostragem *boostrap* (Baba & Sevil, 2019). O processo é demonstrado por meio da Figura 1:

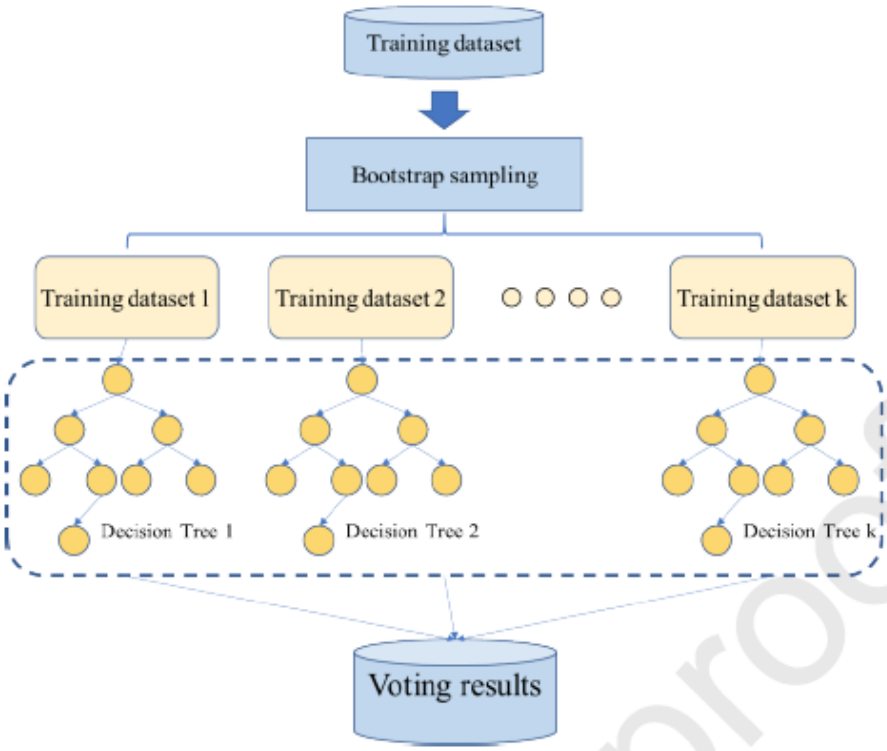


Figura 1 – Processo de Construção de modelos por meio de Random Forest
Fonte: Adptado de Pan & Zhou (2019)

O recurso de seleção do algoritmo random forest pode selecionar diferentes subconjuntos dos recursos quando existem leves variações nos dados de treinamento, o que faz com que resulte numa boa acuracidade das previsões (Arora & Kaur, 2020).

O método Randon Forest é explicado por Baba & Sebil da seguinte forma: assume-se que se tenha um conjunto de dados D. f. $x_1;y_1...X_n;Y_n$, procura-se encontrar a função f: X/Y onde X é a entrada e Y representa as saídas. Além do mais, assume-se que M seja o número de entradas, os seguintes passas serão aplicados:

1. O método Random forest, seleciona randomicamente *N* observações de *D* com substituições para forma a amostra *bootstrap*;
2. Cada árvore cresce utilizando o subjconto de dados *M*, que é um recurso obtido do recurso total. Para regressão, é recomendado que o subconjunto de recursos seja M=3. Então em cada nó, *M* recursos são selecionados de forma randômica e a melhor divisão entre os recursos *M* é selecionada de acordo com a medida de impureza (Gini impurity);

3. As árvores crescem até a máxima profundidade sem poda.

Segundo Breiman (2001), o método Random Forest é a combinação de árvores de preditores dependentes dos valores de um vetor randômico com amostragem independente, utilizando a mesma distribuição para todas as árvores que existem na floresta.

Análise dos Resultados

Primeiramente, foram importados os pacotes pandas, numpy, matplotlib e seaborn. Depois foi atribuído ao objeto denominado ‘dados_ml’ toda a base de dados dos empregados e empresas respondentes. Houve a separação das variáveis independentes que mensuravam os atributos dos empregados e das empresas em que atuavam, conforme Quadro 1, a seguir :

Quadro 1- Descrição das variáveis independentes

Código	Descrição
150M	Estar ou não na lista das 150 Melhores Empresa
tipinstitu	Tipo de instituição
gen	Geração do empregado
origcapital	Origem do capital da empresa
capitaberto	Se é de aberto ou não
setor	Setor de atuação da empresa
anos de operação no Brasil	Anos de operação no Brasil
totempr	Total de empregados
prevquadro	Previsão sobre aumento, manutenção ou redução do quadro de funcionários
rotgeral	Índice de rotatividade geral
clt	Se o empregado é CLT ou não
tempempr	Tempo de empresa do empregado
cargo	Cargo do empregado
fxsalarial	Faixa salarial do empregado
loctrab	Local de Trabalho do empregado
sexo	Sexo do empregado
retnaemp	Fator de retenção do empregado
escolaridade	Escolaridade do empregado
meb	Escore de percepção do empregado sobre employer branding
mci	Escore de percepção do empregado sobre gestão da comunicação interna
mpa	Escore de percepção do empregado sobre gestão da participação e autonomia
msd	Escore de percepção do empregado sobre gestão da sustentabilidade e diversidade
mri	Escore de percepção do empregado sobre gestão das relações interpessoais
mgc	Escore de percepção do empregado sobre gestão de carreira
mpo	Escore de percepção do empregado sobre gestão de processos e organização

mqvt	Escore de percepção do empregado sobre gestão de saúde, segurança e qualidade de vida
medc	Escore de percepção do empregdo sobre gestão do conhecimento e educação corporativa
mlid	Escore de percepção do empregado sobre gestão do perfil da liderança
mrec	Escore de percepção do empregado sobre gestão do reconhecimento e recompensa
mestr	Escore de percepção do empregado sobre gestão estratégica e objetivos
rot_vol	Índice de rotatividade voluntária
Categ_futnaemp	Faixa de futuro de trabalho na empresa

Todas essas variáveis foram consideradas preditoras da variável dependente denominada 'fx_desemp' ou faixa de desempenho. Também houve transformações de variáveis com o intuito de balancear a quantidade de respondentes para cada variável em questão e identificar valores discrepantes.

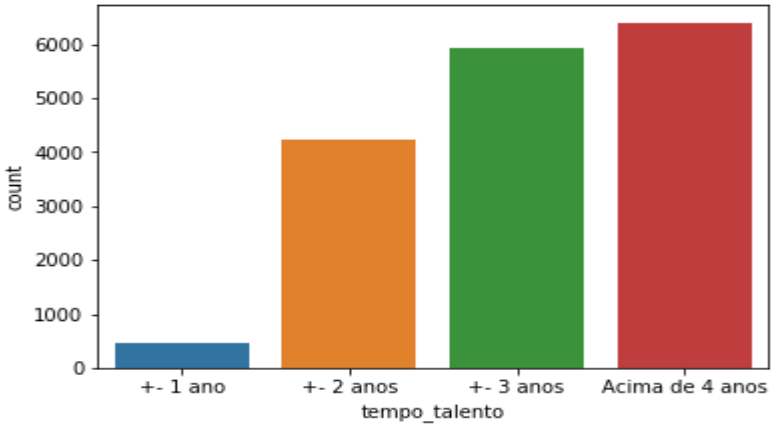
A base de dados foi dividida em duas, uma para treinamento com 80% dos respondentes e 20% para teste, com localização random_state de 101. Então, partimos para o delineamento do modelo de machine learning, optando pelo algoritmo de Randon Forest, que foi obtido ao carregar o pacote *skylearn*. Para o treinamento foram considerados 700 estimadores.

Após o cômputo do treinamento e comparação com a base de dados foi verificado uma taxa de acerto de 78% na previsão do empregados se tornar um talento, sendo ele situado na faixa superior desempenho. Além disso, procuramos tratar de maneira probabilística a chance disto ocorrer na base de teste, por meio da função 'predictions = rfc.predict_proba(x1_test)'. Com isto, incluímos uma coluna com as probabilidades de cada respondentes ser ou não talento no ambiente de trabalho em que estão trabalhando.

Adiante, consideramos que quanto maior a probabilidade de virar talento, menor o tempo de se tornar de fato. Assim, os respondentes com probabilidade maior do que 90% se tornariam talentos em mais ou menos um ano, entre 71% e 90% em mais ou menos dois anos, entre 51% e 70% em mais ou menos três anos e, abaixo de 51%, acima de 4 anos.

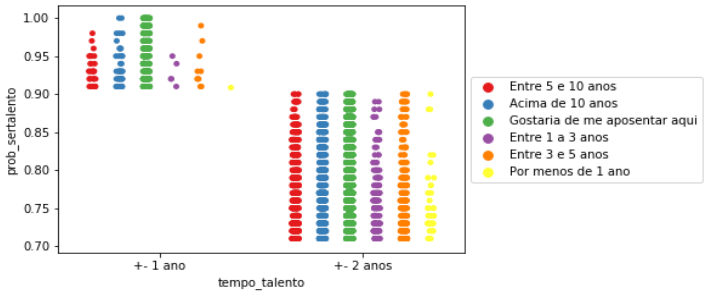
No Gráfico 1, está a distribuição dos respondentes da base de teste, conforme o tempo para se tornarem talentos. Nota-se a menor proporção daqueles que se tornarão talentos em mais ou menos um ano, o que faz sentido tendo em vista o caráter das qualidades diferenciadas desses profissionais.

Gráfico 1- Distribuição dos respondentes pelo tempo de se tornarem talentos



Procuramos também aferir o desejo dos talentos sobre o tempo de permanência na em presa, indo daqueles que pretendem ficar por mais ou menos um ano até aqueles que querem s e aposentar na empresa atual. Para esta análise selecionamos apenas os profissionais que se to rnarão talentos entre um e dois anos, mais ou menos, por serem as pessoas com o maior poten cial em termos probabilísticos.

Gráfico 2- Distribuição entre o tempo de permanecer na empresa atual X probabilidade de virar talento



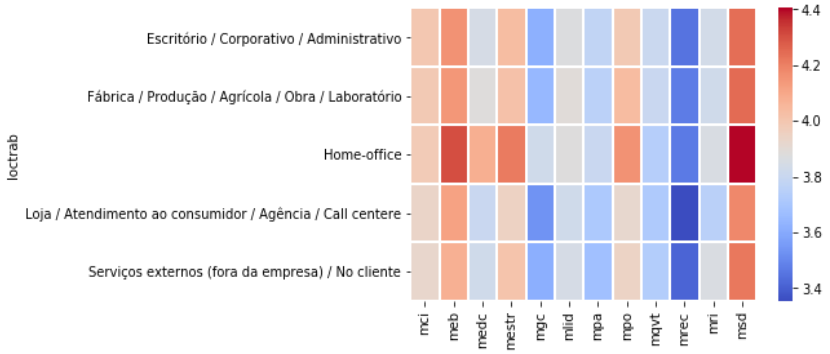
O Gráfico 2 mostra a maior distribuição de respondentes situados nas faixas daqueles que pretendem continuar trabalhando acima dos 10 anos ou até mesmo se aposentar na empresa atual. Vale ressaltar que os dados consideram os profissionais que atuam ou se inscreveram para concorrer ao prêmio das Melhores Empresas Para Você Trabalhar, ou seja, são ou possuem condições de ser empresas referências em gestão de pessoas no Brasil.

Para efeito de gestão, também é importante analisar os casos situados nas faixas inferiores a um ano. Portanto, o modelo aqui gerado serve como ferramenta de gestão ao identificar tais situações de maneira preventiva antes do pedido de demissão do talento.

Dando sequência as análises buscamos relacionar a percepção do clima organizacional com as características de perfil dos respondentes, considerando aqueles nas faixas com mais

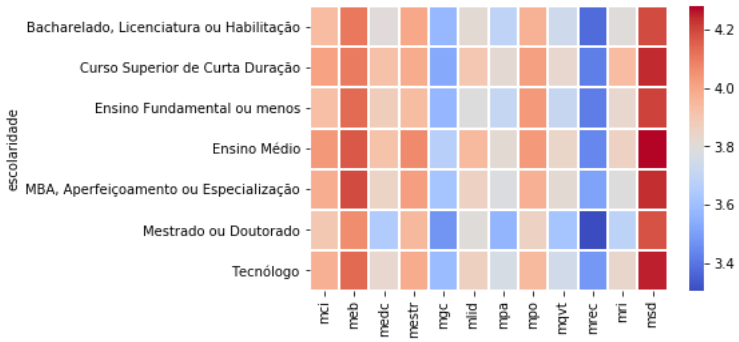
ou menos 1 e 2 anos de se tornarem talentos. É importante salientar que as categorias adotadas são as mesmas que o Guia das Melhores Empresas Para Você Trabalhar adota anualmente em suas seções. A seguir, no Gráfico 2, estão dois exemplos, ao associarmos o local de trabalho e escolaridade do respondente com as categorias de clima organizacional. O tipo de gráfico escolhido foi o de mapa de calor.

Gráfico 3- Associação entre local de trabalho do respondente e sua percepção sobre categorias do clima organizacional.



No Gráfico 3 constata-se as menores médias (tonalidades na cor azul) nas categorias de recompensas e reconhecimento, gestão de carreira, qualidade de vida e participação e autonomia. As maiores (tonalidades da cor vermelha) estão nas categorias de sustentabilidade e diversidade e *employer branding*. Especificamente, por local de trabalho, as pessoas que trabalham *home office* apresentam as maiores médias em *employer branding*, educação corporativa, estratégia e objetivos, processos e organização e sustentabilidade e diversidade.

Gráfico 4- Associação entre escolaridade do respondente e sua percepção sobre categorias do clima organizacional.



No Gráfico 4, por escolaridade, os profissionais com mestrado e doutorado apresentam menores médias em educação corporativa, gestão de carreira, participação e autonomia, recompensas e reconhecimento e relações interpessoais.

Na Tabela 1, a seguir, estão as importâncias das variáveis independentes sobre a variável desempenho. Destaca-se as cinco maiores importâncias, que são das variáveis do índice de rotatividade voluntária da empresa, setor de atuação, rotatividade geral, anos de operação no Brasil e fator de retenção do empregado.

Tabela 1- Importâncias das variáveis preditoras

Variável preditora	Importância
Rotatividade voluntária	0,098
Setor	0,054
Rotatividade geral	0,046
Anos de Operação no Brasil	0,041
Fator de retenção na empresa	0,037

Por fim, com o desenvolvimento desse sistema, é possível incluir dados das empresas participantes do Guias das Melhores Empresas para Você Trabalhar e gerar previsões que permitam desenvolver e gerenciar a retenção de talentos. Como próxima etapa, pretendemos desenvolver uma interface mais direcionada ao usuário final.

Bibliográfia

Arora, N., & Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing*, *86*, 105936. https://doi.org/10.1016/j.asoc.2019.105936

Aulia, A., Jeong, D., Saaid, I. M., Kania, D., Shuker, M. T., & El-Khatib, N. A. (2019). A Random Forests-based sensitivity analysis framework for assisted history matching. *Journal of Petroleum Science and Engineering*, *181*, 106237. https://doi.org/10.1016/j.petrol.2019.106237

Baba, B., & Sevil, G. (2019). Predicting IPO initial returns using random forest. *Borsa Istanbul Review*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Pan, S., & Zhou, S. (2019). Evaluation Research of Credit Risk on P2P Lending based on Random Forest and Visual Graph Model. *Journal of Visual Communication and Image Representation*, 102680. https://doi.org/10.1016/j.jvcir.2019.102680

Silver, D. L., Yang, Q., & Li, L. (2013). Lifelong Machine Learning Systems: Beyond Learning Algorithms. *2013 AAAI Spring Symposium Series*. Apresentado em 2013 AAAI Spring Symposium Series. Recuperado de https://www.aaai.org/ocs/index.php/SSS/SSS13/paper/view/5802

Vezza, P., Muñoz-Mas, R., Martinez-Capel, F., & Mouton, A. (2015). Random forests to evaluate biotic interactions in fish distribution models. *Environmental Modelling & Software*, *67*, 173–183. https://doi.org/10.1016/j.envsoft.2015.01.005