UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Lucas Rômulo de Souza Resende

Predição salarial de profissionais da área de dados no mercado brasileiro

São João Del Rei

2023

UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Lucas Rômulo de Souza Resende

Predição salarial de profissionais da área de dados no mercado brasileiro

Monografia apresentada como requisito da disciplina de Projeto Orientado em Computação I do Curso de Bacharelado em Ciência da Computação da UFSJ.

Orientador: Carolina Ribeiro Xavier

Universidade Federal de São João del-Rei — UFSJ

Bacharelado em Ciência da Computação

São João Del Rei

2023

Lucas Rômulo de Souza Resende

Predição salarial de profissionais da área de dados no mercado brasileiro

Monografia apresentada como requisito da disciplina de Projeto Orientado em Computação I do Curso de Bacharelado em Ciência da Computação da UFSJ.

Trabalho aprovado. São João Del Rei, 13 de fevereiro de 2023

Carolina Ribeiro Xavier
Orientadora

João Gabriel Rocha Silva
Avaliador

Lucas Gabriel da Silva Félix
Avaliador

São João Del Rei
2023

Agradecimentos

Meu primeiro agradecimento e, de longe, o mais especial, vai para minha mãe, Edna Aparecida de Souza, que, mesmo sendo originária de família pobre e tendo estudado apenas até a 4ª série do ensino fundamental conseguiu compreender o valor da educação e me apoiou por toda essa trajetória na UFSJ.

Agradeço também aos meus amigos, de mais de década, da minha cidade natal. A vida foi passando, cada um seguiu seus rumos e, ainda assim, a gente segue junto, firme e forte. Gostaria que vocês soubessem que são parte da minha família.

Quero agradecer também a todos os professores com os quais pude aprender coisas esplêndidas, em especial a Carol, que foi minha orientadora neste trabalho e com a qual eu tive a oportunidade de cursar matérias que agregaram muito valor à minha formação.

Por último, mas não menos importante, agradeço a todos os amigos que pude fazer e com os quais convivi nesses anos vividos em São João Del-Rei até então. Essa caminhada teria sido muito mais árdua sem vocês.

Só sei que nada sei.

(Sócrates)

Resumo

Este trabalho tem como objetivo realizar predições salariais e discorrer sobre os perfis dos profissionais que compõe o mercado de dados brasileiro. Poder calcular a remuneração salarial de profissionais de certa área, de acordo com suas características, é uma tarefa de grande auxílio para pessoas e empresas, especialmente quando essa é uma área que vem chamando interesse pela sua capacidade de crescimento. Buscando compreender os perfis dos profissionais que estão empregados hoje, foram aplicadas técnicas de análises de dados, buscando por características que possam influenciar no cálculo salarial. Foram, também, desenvolvidos modelos de aprendizado de máquina com o intuito de gerar predições salariais dado o perfil de um profissional. Entre as conclusões obtidas, foi possível observar que o mercado de dados brasileiro não possui os perfis de seus trabalhadores bem definidos, o que pode ser resultado de uma falta de regulamentação da área. Também foram obtidos modelos que conseguem predizer os salários dos profissionais com até 54% de precisão e RMSE baixo/médio, o que indica que esses modelos podem auxiliar as pessoas e empresas realizando predições que alcançam valores próximos da remuneração ideal dado o perfil de um profissional.

Palavras-chaves: dados, aprendizado de máquina, classificação, predição, Random Forest, Extra Trees, AdaBoost, Gradient Boosting, análise de dados, ciência de dados, engenharia de dados, salário, profissionais, desenvolvedores, gestores

Abstract

This work aims to make salary predictions and discuss the profiles of professionals that work in the Brazilian data market. Being able to calculate the salary of professionals in a certain area, according to their characteristics, is a task of great help for people and companies, especially when this is an area that has been attracting interest due to its capacity for growth. Seeking to understand the profiles of the professionals who are employed today, data analysis techniques were applied, looking for characteristics that could influence the salary calculation. Machine learning models were also developed in order to generate salary predictions given the profile of a professional. Among the conclusions obtained, it was possible to observe that the Brazilian data market does not have well-defined worker profiles, which may be the result of a lack of regulation in the area. Models were also obtained that can predict the salaries of professionals with up to 54% accuracy and low/medium RMSE, which indicates that these models can help people and companies by making predictions that reach values close to the ideal remuneration given the profile of a professional.

Key-words: data, machine learning, classification, prediction, Random Forest, Extra Trees, AdaBoost, Gradient Boosting, data analysis, data science, data engineer, salary, wage, earnings, professionals, developers, managers

Lista de ilustrações

Lista de abreviaturas e siglas

DA	Data Analysis
DE	Data Engineer
DS	Data Science
ML	Machine Learning
RMSE	Root Mean Squared Error

Sumário

1 Introdução

O setor de Tecnologia da Informação (TI) é um dos que mais cresceu no país nos últimos anos[1]. Esse crescimento foi mais rápido que o previsto, entre outros motivos, pela dependência de soluções tecnológicas que pudessem manter as pessoas conectadas, mesmo que à distância, durante a pandemia do COVID-19, o que acarretou em um aumento no número de vagas no setor.

Dentro do setor, uma área que também vem se destacando é a ciência de dados, que contou com um crescimento de mais de 450%[2] no número de vagas entre 2019 e 2020, e promete manter um crescimento acelerado.

Um dos grandes atrativos da área são seus salários. No Brasil, segundo o Glassdoor[3], um site de vagas e recrutamento online mostra que, o salário médio de um profissional cientista de dados é cerca de R$8.165/mês, podendo atingir valores acima dos R$60.000/mês.

O cálculo salarial pode ser uma questão que se mostra de grande dificuldade para várias pessoas dentro mercado, sejam profissionais ou empregadores. Essa dificuldade pode ainda se mostrar mais complexa para pessoas que estão ingressando no mercado atualmente; e pode ainda ter sido influenciada pelo grande crescimento do setor.

Com o aumento da comunidade da área é comum observar membros novatos pedindo auxílio aos mais experientes, inclusive nas redes sociais, com relação a estimativa de uma remuneração justa; fatores nos quais se basear no momento da realização do cálculo (equipamentos necessários, tempo investido, entre outros); como são entregues os projetos; entre outros quesitos.

O entendimento dos perfis dos trabalhadores e a realização de análises preditivas da remuneração atual na área podem servir de grande ajuda à esses profissionais e empresas que estão tentando se incluir nesse mercado, ou até mesmo àqueles que estão apenas tentando se manterem atualizados à situação atual.

Esse trabalho tem como objetivo a realização de análises exploratórias sobre dados desse mercado atualmente, levando à compreensão dos perfis dos profissionais no mercado brasileiro, e a criação e avaliação de modelos de ML que procuram predizer os salários de profissionais da área de dados.

Neste trabalho foram realizadas análises em uma base de dados que conta com informações de mais de 2000 profissionais, dentre eles desenvolvedores e gestores, obser-

[1] https://is.gd/setor_de_tecnologia_cresce
[2] https://is.gd/aumento_de_vagas_em_dados
[3] https://is.gd/salario_cintista_de_dados

vando-se dados demográficos, dados sobre carreira e seus afazeres e conhecimentos na área, e pode-se obter mais conhecimento sobre os perfis desses profissionais.

Os modelos de ML foram treinados e analisados utilizando essa mesma base de dados, e com eles foi possível realizar predições salariais com uma precisão de 53.6% e um RMSE de 2.13 para os profissionais desenvolvedores, e 48.6% de precisão e RMSE de 5.3 para os profissionais em cargos de gestão.

O restante deste trabalho está organizado da seguinte forma: no capítulo 2 é apresentado o referencial teórico, o capítulo 3 discorre sobre os trabalhos relacionados, no capítulo 4 pode ser compreendida a metodologia trabalhada, o capítulo 5 conta com observações realizadas sobre os resultados obtidos nos modelos de ML e no capítulo 6 são realizadas conclusões a cerca do tema e a apresentação dos trabalhos futuros.

Com este trabalho é possível concluir que a falta de regulamentação do mercado, entre outros fatores, pode tornar a tarefa de predição salarial dos profissionais –desenvolvedores e gestores– da área de dados no mercado brasileiro um tanto quanto complicada, mas, ainda assim, os modelos apresentados nesse trabalho podem auxiliar pessoas e empresas a terem uma ideia da remuneração salarial desses profissionais de acordo com seus perfis.

2 Referencial Teórico

2.1 Machine Learning

De acordo com Mitchell em (1), *Machine Learning* (ML) é o estudo e desenvolvimento de algoritmos que conseguem compreender conceitos e se aprimorar com base em suas experiências. Essa é uma área multidisciplinar que envolve inteligência artificial, probabilidade e estatística, teoria da informação, filosofia, psicologia, neurobiologia, entre outras.

Os algoritmos de ML, como *Decision Tree*, *Naive Bayes* e redes neurais, podem ser utilizados em várias áreas, com por exemplo, reconhecimento de fala, veículos autônomos, medicina e jogos de tabuleiro.

2.2 Aprendizado supervisionado

O aprendizado supervisionado é uma classe de ML, e trata de um tipo de aprendizado onde os dados que os algoritmos utilizam para aprender são demarcados com *labels*, que podem ser utilizados na avaliação dos resultados, o que induz a ideia da existência de uma supervisão nas classificações realizadas.

Os modelos gerados a partir dos algoritmos de aprendizado supervisionado buscam compreender as características dos dados utilizados na aprendizagem para poderem ser utilizados na classificação de novas instâncias, como é apontado por Cunningham, P, *et al.* em (2).

2.3 Random Forest

De acordo com Leo Braiman, o algoritmo *Random Forest* consiste na criação de várias de árvores de decisão, sendo essas árvores geradas de formas diferentes através da utilização de fatores randomizados. No momento da classificação essas árvores decidem por alguma classe através de uma votação onde ganha a maioria.

Os fatores randomizados utilizados na construção das árvores são vários. Breiman ainda cita alguns desses fatores em (3), como por exemplo, a utilização de bootstrap, pelo próprio autor em (4) e a utilização de uma amostragem randômica das *features* no momento do *split*, por Dietterich em (5).

2.4 Extra Trees

Geurts et al. propuseram em 2005 um novo algoritmo baseado em árvores, como pode ser visto em (6). Esse novo algoritmo, o *Extremelly Randomized Trees* ou *Extra Trees*, assim como o *Random Forest*, utiliza várias árvores de decisão, e a classificação se dá pela votação da maioria.

No algoritmo proposto pelos autores, no momento do *split* na construção das árvores, ambas as seleções dos atributos candidatos e do ponto de corte (atributo selecionado), são realizadas de forma parcialmente ou totalmente randomizada. Essa ideia veio de uma pesquisa realizada por Mingers, na qual ele comparava o desempenho entre árvores de decisão clássicas e árvores de decisão randomizadas, em (7).

2.5 AdaBoost

O algoritmo *Adaptive Boosting* ou *AdaBoost* foi criado por Freund et al. e pode ser encontrado em (8), e em 2009, Hastie et al. desenvolveram (9), uma versão do algoritmo que consegue trabalhar com múltiplas classes.

O algoritmo criado por Freund et al. utiliza um vetor de pesos que são calculados a partir dos erros de um estimador. Ele realiza um *loop* no qual em cada iteração é gerado estimador fraco, os erros são calculados, e o vetor de pesos é atualizado. Por fim, ele retorna uma hipótese que é utilizada na classificação de novas instâncias.

2.6 Gradient Boosting

Criado em 1999 por Friedman e descrito em (10) e (11), o *Gradient Boosting* é um algoritmo que cria vários modelos, e a cada novo modelo busca reduzir a perda do modelo anterior.

Quando um novo modelo é gerado, ele utiliza o cálculo da perda do modelo anterior para atualizar sua função de estimativa, buscando reduzir ao máximo a perda no fim das iterações. O *Gradient Boosting* trabalha com otimizações numéricas na atualização de suas funções de estimativa, e geralmente tem um desempenho melhor que o *AdaBoost*, um de seus "antecessores".

3 Trabalhos Relacionados

As análises de predição buscam auxiliar pessoas e instituições em tomadas de decisões, e são utilizadas em uma gama muito variada de áreas. Essas análises são feitas utilizando técnicas de classificação, quando se busca definir à qual classe as instâncias pertencem, ou regressão, quando é necessário inferir valores numéricos, de acordo com a natureza do problema.

Para o problema tratado neste trabalho, será necessário a aplicação de métodos de classificação, que podem ser observados na literatura em diversas áreas.

Em (12), os autores Sousa, Ronieri et al. utilizam técnicas de classificação, como o *Random Forest*, com o objetivo de predição de dislexia em crianças. A base utilizada foi projetada realizando-se exercícios linguísticos onde se observa indicadores relacionados à dislexia. Entre as crianças participantes dos exercícios cerca de 10% possuiam o distúrbio. Por fim, conseguiram um modelo que demonstrava 98% de acurácia na classificação como com/sem dislexia.

Rodrigues, Ebony et al., em (13) buscam predizer o tempo de permanência dos estudantes de graduação do ensino público brasileiro utilizando algoritmos de ML. Os dados utilizados são das edições de 2016 a 2018 do ENADE, e possuem três classes que indicam o tempo que cada estudante estava na graduação: entre 2 e 4 anos, entre 5 e 7 anos e 8 anos ou mais. Entre os algoritmos utilizados no trabalho, o *XGBoost* foi o que se destacou em todos experimentos executados.

Ainda no âmbito escolar, em (14), Colpo M. et al. utilizam modelos de *Decision Tree* e *Random Forest* buscando realizar predições sobre a evasão estudantil. Foram utilizados dados do sistema SIGAA do Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar), desde 2014. Os autores chegam a conclusão de que os alunos que possuem melhor desempenho, frequência e mais tempo de curso são aqueles que tem menos tendência à evasão.

Em (15), os autores Capanema, C et al. utilizam redes neurais para identificar possíveis pontos de interesse de visita de usuários de dispositivos móveis. O trabalho contou com dados reais de quase 200 usuários. Os autores classificam os pontos de interesse visitados pelos usuários como casa, trabalho e outros, e a partir do ponto atual de um usuário, buscam predizer qual o sue próximo destino. O trabalho trouxe melhorias em relação aos resultados conhecidos até então pela utilização do intervalo de inatividade dos usuários, o que auxiliou na classificação.

4 Metodologia

4.1 Dados

As bases utilizadas nesse trabalho foram geradas à partir de análises e processamento de uma base de dados obtida através do *kaggle*[1]: *State of Data Brazil 2021*[2], que conta com informações de profissionais que atuam na área de dados no mercado brasileiro.

Integrantes da comunidade *Data Hackers*[3] e da empresa *Bain & Company*[4] elaboraram a base *State of Data Brazil 2021* por meio de uma pesquisa na qual os participantes respondiam um questionário contendo nove seções sobre suas vivências no ramo. A divisão do questionário foi realizada da seguinte maneira:

- Dados demográficos;
- Dados sobre carreira;
- Desafios dos gestores de times de dados;
- Conhecimentos na área de dados;
- Objetivos na área de dados;
- Conhecimento em Análise de Dados/DA;
- Conhecimento em Ciência de Dados/DS;
- Conhecimento em Engenharia de Dados/DE;
- Sobre a comunidade Data Hackers.

Durante a fase de processamento e análise exploratória, foi concluído que a base de dados possui seções com características específicas de profissionais que atuam como gestores, seções específicas contendo informações de desenvolvedores, e características gerais - que ambos os grupos de gestores e desenvolvedores possuem-. Dessa forma, foram geradas duas bases de dados à partir da base original: dados sobre gestores e dados sobre desenvolvedores.

Na base de dados sobre gestores existem informações das seções de dados demográficos e dados sobre carreira - características gerais -, e também a variável *Quais dessas*

[1] https://www.kaggle.com/
[2] https://www.kaggle.com/datasets/datahackers/state-of-data-2021
[3] https://www.datahackers.com.br/
[4] https://www.bain.com/pt-br/

responsabilidades fazem parte da sua rotina atual de trabalho como gestor? da seção desafios dos gestores de times de dados. Essa base possui 479 instâncias de profissionais, representando aqueles que responderam como gestores.

A base de dados sobre desenvolvedores também possui as variáveis das seções de dados demográficos e dados sobre carreira - características gerais -, adicionalmente à variável *Quais das linguagens listadas abaixo você utiliza no trabalho?* da seção de conhecimentos na área de dados. Essa base é composta pelas 1564 instâncias daqueles que responderam as seções de desenvolvedores do questionário.

4.1.1 Processamento dos dados

Como o intuito do trabalho é realizar análises e predições sobre os profissionais do mercado brasileiro, foram necessárias algumas etapas de processamento sobre a base adquirida para que os dados resultantes fossem utilizados durante o desenvolvimento.

A primeira etapa do processamento dos dados consiste na exclusão daqueles profissionais profissionais que moram e/ou trabalham fora do Brasil. Nessa etapa foram excluídas cerca de 105 instâncias.

Na etapa seguinte buscou-se remover aquelas instâncias de pessoas que se declararam estudantes, desempregados, acadêmicos/pesquisadores e aqueles que preferiram não informar seu regime de trabalho, sendo excluídas aproximadamente 274 instâncias, restando aqueles que se declararam como CNPJ, PJ, servidores públicos, *freelancers* e estagiários.

Depois dessas etapas de eliminação de instâncias, foram avaliadas as *features* presentes na base, em busca daquelas que representavam alguma relação com o objetivo do trabalho, para comporem os dados a serem utilizados nos modelos.

Após a seleção das *features* foram realizadas outras análises exploratórias em busca de dados que pudessem levantar dúvidas, como: *outliers*, dados preenchidos incorretamente, instâncias sem valores em *features* importantes, informações que iam contra a situação do mercado, entre outros. Após essas análises foram excluídos cerca de 223 instâncias.

Por fim, os valores e nomes de algumas *features* foram renomeados com o intuito de facilitar o desenvolvimento do trabalho e a leitura dos gráficos, e essa base foi então dividida em duas novas bases: desenvolvedores e gestores.

4.1.2 Bases de dados: desenvolvedores e gestores

A base de dados completa –após o pré-processamento– conta com 2043 instâncias que representam todos os profissionais, sendo 479 gestores e 1564 desenvolvedores, e é

composta por informações das seções de dados demográficos, dados sobre carreira, desafios dos gestores de times de dados e conhecimentos na área de dados, como pode ser visto na figura 1, sendo a faixa salarial a variável que representa as classes.

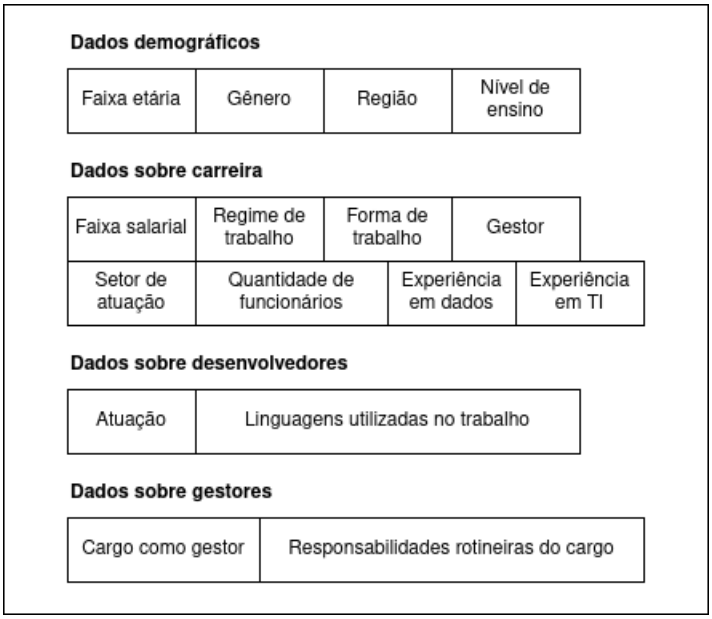

Figura 1 – Seções que compõe a base de dados e suas variáveis

As imagens 2 e 3 representam as matrizes de correlação entre as variáveis da base de desenvolvedores e da base de gestores, respectivamente.

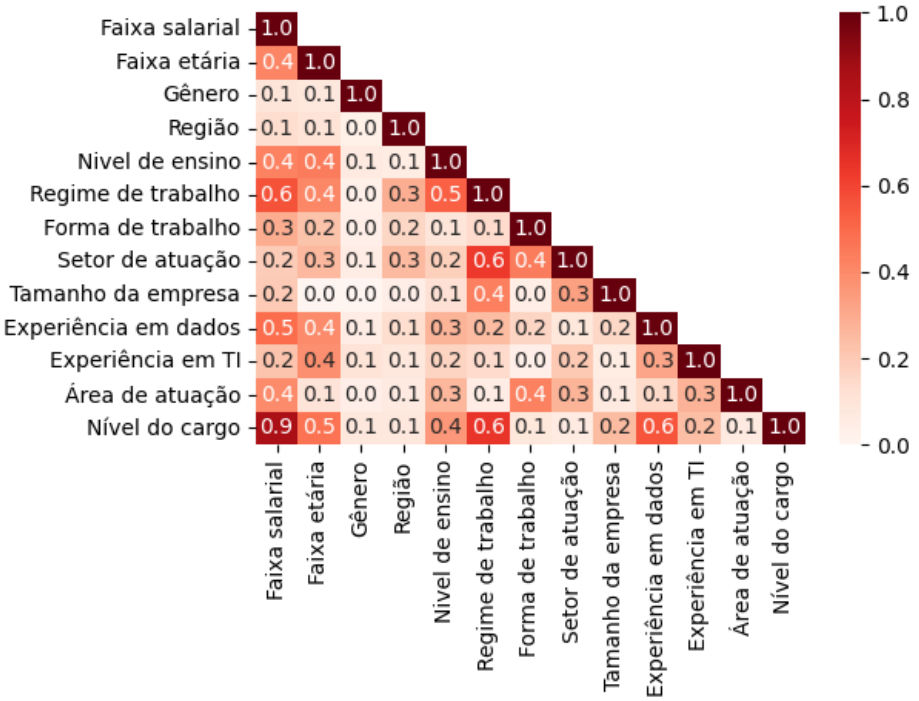

Figura 2 – Matriz de correlação das variáveis da base de desenvolvedores

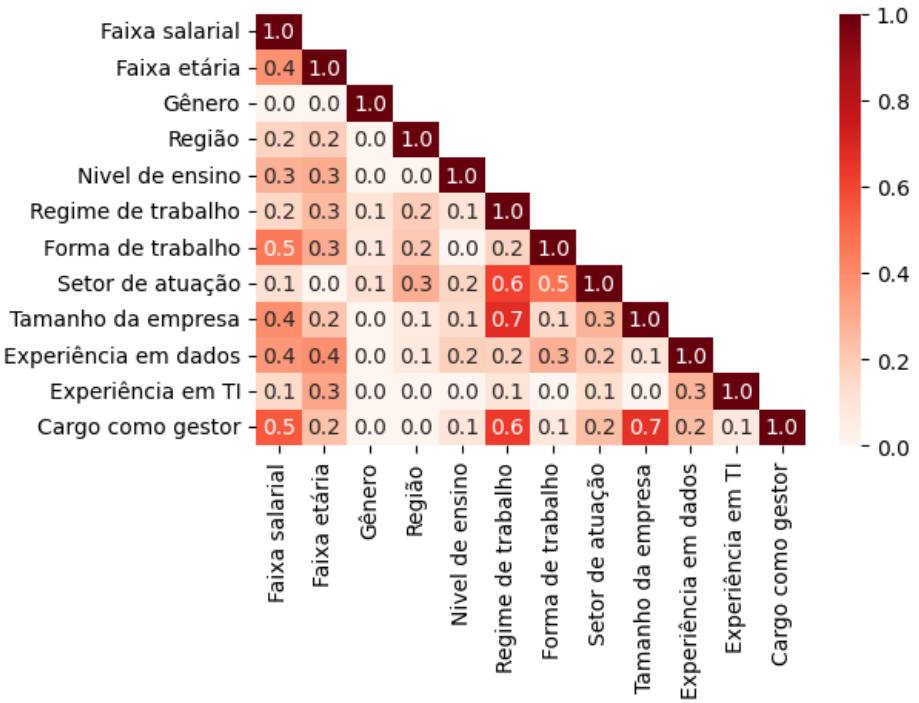

Figura 3 – Matriz de correlação das variáveis da base de gestores

Para gerar conhecimento sobre os perfis dos profissionais da área de dados foram realizadas análises exploratórias sobre as algumas *features* que foram selecionadas por possuir uma bom coeficiente de correlação com as duas bases, desenvolvedores e gestores, e algumas específicas de cada uma, sendo elas:

- Desenvolvedores: Faixa salarial, faixa etária, anos de experiência trabalhando com dados, área de atuação e nível de cargo e linguagens utilizadas no trabalho;
- Gestores: Faixa salarial, faixa etária, anos de experiência trabalhando com dados, cargo como gestor e responsabilidades do cargo.

4.1.2.1 Faixa salarial

A faixa salarial dos profissionais representa a variável objetivo (classe) do trabalho. É uma variável com característica qualitativa ordinal, e que descreve os intervalos de ganho mensal -em milhares de reais (mil R$)- dos profissionais. Ela é composta por 12 valores, que são eles: 1-2, 2-3, 3-4, 4-6, 6-8, 8-12, 12-16, 16-20, 20-25, 25-30, 30-40, 40+.

De acordo com os dados utilizados, a média salarial dos desenvolvedores é de quase R$7.600,00, e a média dos gestores chega a quase R$15.100,00. Porém, essas médias são aproximações, já que os salários são descritos por intervalos e as informações que compõe a base são do ano de 2021.

Essa diferença entre a média salarial dos desenvolvedores e gestores podem ser compreendidas na Figura 4, onde é possível observar a distribuição das faixas salariais entre os desenvolvedores e gestores.

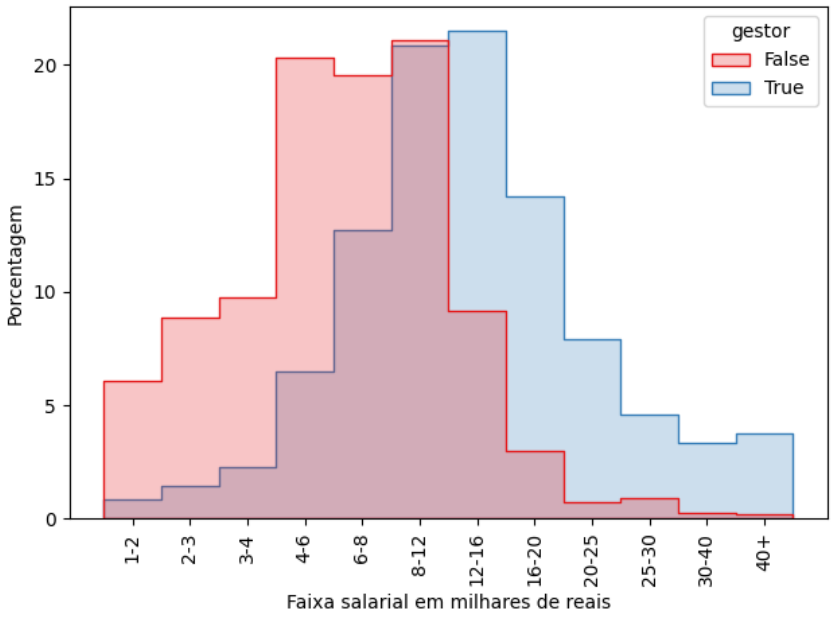

Figura 4 – Distribuição dos profissionais em relação a faixa salarial

É possível perceber que as faixas salariais com a maior concentração de desenvolvedores vão de R$4.000,00 a R$12.000,00, e mais de 85% dos desenvolvedores recebem até R$12.000,00, enquanto a maior concentração de gestores são nas faixas salariais entre R$8.000,00 e R$16.000,00, sendo que acima dos R$12.000,00 a quantidade de gestores é muito maior que a de desenvolvedores.

São apenas 5% os desenvolvedores que recebem acima de R$20.000,00, enquanto quase 4% dos gestores recebem salários acima dos R$40.000,00.

Essa diferença era esperarada, pois comumente os gestores recebem salários melhores que seus desenvolvedores.

4.1.2.2 Faixa etária

A faixa etária é uma *feature* que apresenta uma boa correlação com a faixa salarial, tanto para os desenvolvedores quanto para os gestores, sendo a única que possui o mesmo coeficiente (0.4) nas duas bases. Ela é uma variável qualitativa ordinal que representa a idade dos profissionais em intervalos e possui nove valores, sendo eles: 17-21, 22-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55+.

Na Figura 5 é possível observar a distribuição dos desenvolvedores e gestores de acordo com suas faixas etárias.

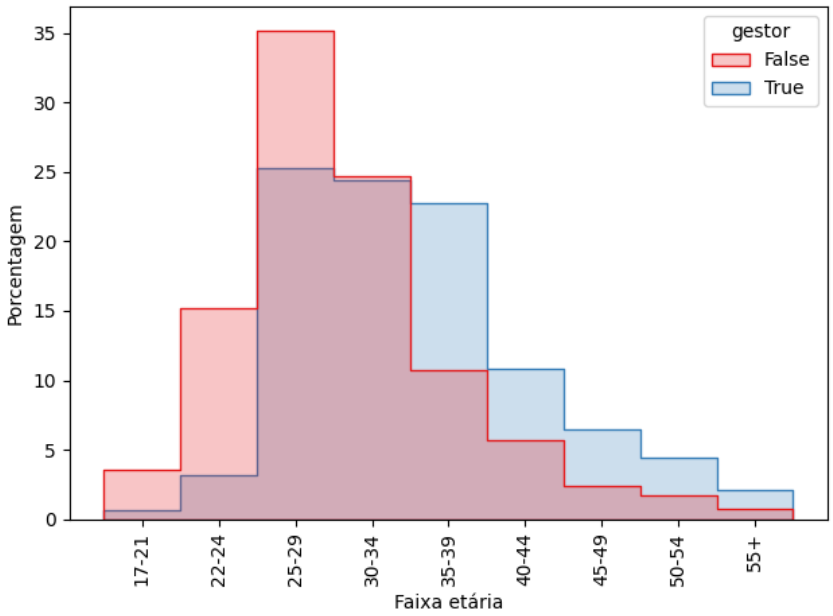

Figura 5 – Distribuição dos profissionais em relação a faixa etária

Como exibido na Figura 5, os profissionais desenvolvedores estão mais concentrados nas faixas etárias até os 34, enquanto a faixa etária dos profissionais gestores se inicia por volta dos 25 anos, e a partir dos 35 anos já existem mais gestores do que desenvolvedores. Com essas observações levantamos a hipótese de que após alguns anos de trabalho os profissionais vão sendo promovidos e chegam a cargos de gestão.

A Figura 6 exibe a distribuição etária, apenas dos desenvolvedores, em relação à faixa salarial, mostrando que o salário aumenta de acordo com as faixas etárias.

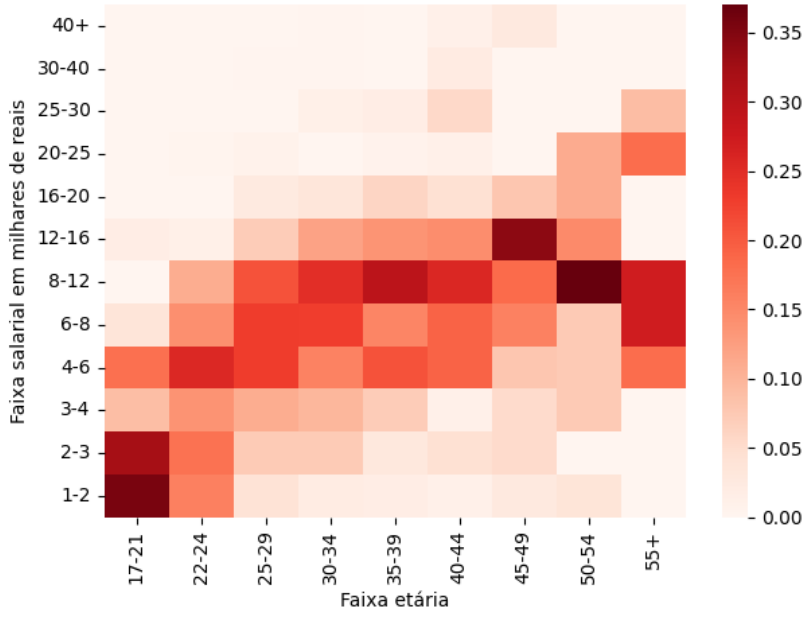

Figura 6 – Distribuição salarial dos desenvolvedores em relação a faixa etária

A grande maioria dos desenvolvedores com idade entre 17 e 21 anos recebe até R$3.000,00/mês, enquanto o salário dos profissionais mais velhos, com 55 anos ou mais, varia entre R$3.000,00 e R$30.000,00, sendo que eles representam a maior concentração dos profissionais que recebem na faixa de R$25.000,00 e R$30.000,00.

Também é possível notar que os poucos desenvolvedores que recebem acima de R$40.000,00 possuem entre 40 e 49 anos, porém os salários dos profissionais nessa faixa etária estão mais concentrados entre R$8.000,00 e R$16.000,00 por mês.

Na Figura 7 é exibida a distribuição da faixa etária dos profissionais que atuam em cargos de gestão, em relação à faixa salarial.

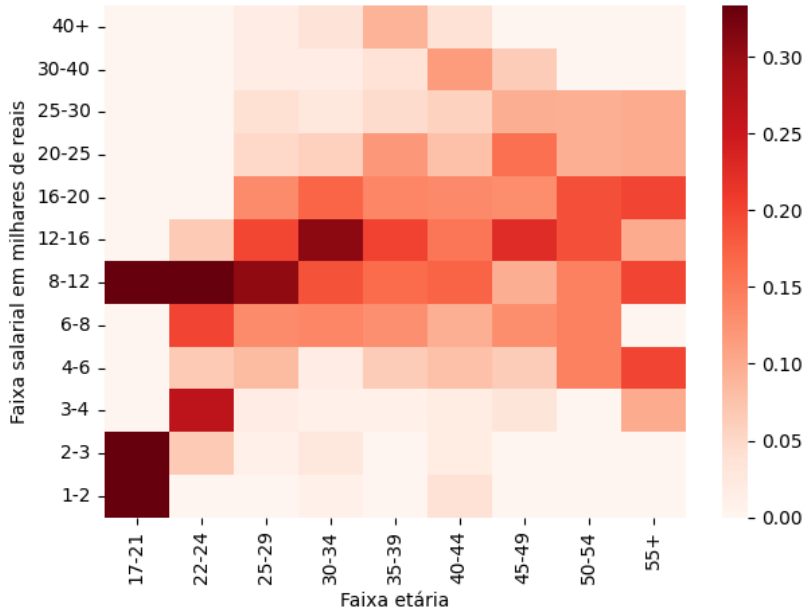

Figura 7 – Distribuição salarial dos gestores em relação a faixa etária

Assim como na distribuição dos desenvolvedores, existe um aumento salarial de acordo com a idade dos gestores. Enquanto os poucos gestores entre 17 e 21 anos, em sua maioria de empresas pequenas, recebem no máximo R$12.000,00, a maior parte dos profissionais acima dos 44 anos recebem salários entre R$25.000,00 e R$30.000,00, e os profissionais mais bem pagos, aqueles que recebem acima de R$30.000,00, estão entre os 35 e 44 anos.

4.1.2.3 Experiência com dados

A *feature* que representa o tempo de experiência na área de dados é de característica quantitativa ordinal, e seus valores representam faixas de anos, sendo eles os sete seguintes: 0 (sem experiência), 0-1, 1-2, 2-3, 4-5, 6-10, 10+.

A experiência com dados apresenta uma boa correlação para ambos, desenvolvedores e gestores, mas se destaca entre os desenvolvedores com o valor de correlação de

0.5.

A Figura 8 representa a distribuição dos profissionais, desenvolvedores e gestores, de acordo com seu tempo de experiência com dados.

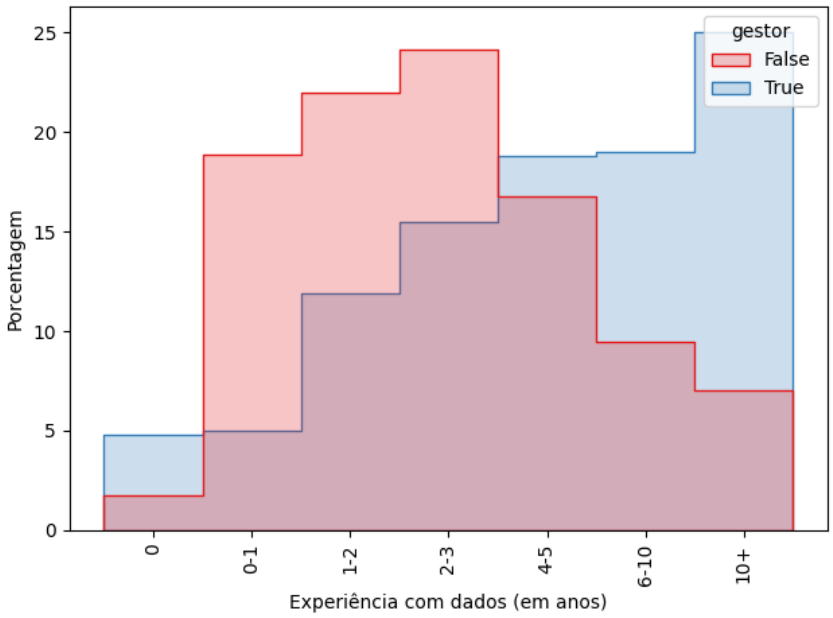

Figura 8 – Distribuição dos profissionais em relação a experiência com dados

Os desenvolvedores possuem uma distribuição que cresce rapidamente até os três anos de experiência, e que começa a decrescer após três anos, enquanto a distribuição dos gestores começa a crescer à partir de um ano de experiência, ultrapassando a quantidade de desenvolvedores nas faixas após os três anos.

Assim como na análise da faixa etária, é possível presumir que os desenvolvedores vão sendo promovidos à cargos de gestão quando vão adquirindo mais experiência.

Outra observação é de que a quantidade de desenvolvedores sem nenhuma experiência com dados e que estão trabalhando na área é muito pequena, enquanto aproximadamente 10% dos gestores possuem nenhuma ou no máximo um ano de experiência com dados, o que pode indicar que alguns cargos de gestão não necessitam de experiência na área para serem ocupados.

Na Figura 9 observa-se a distribuição das faixas de anos de experiência com dados em relação à faixa etária dos profissionais. é possível notar o crescimento salarial dos desenvolvedores de acordo com o aumento do tempo de experiência na área de dados.

Enquanto os profissionais com menos tempo de experiência, até dois anos, possuem seus salários mais concentrados nas faixas de até R$6.000,00, os profissionais com mais de 10 anos de experiência tem seus salários mais concentrados entre R$8.000,00 e R$16.000,00, e essa também é a faixa de experiência que possui a maior concentração de profissionais que recebe acima dos R$40.000,00/mês.

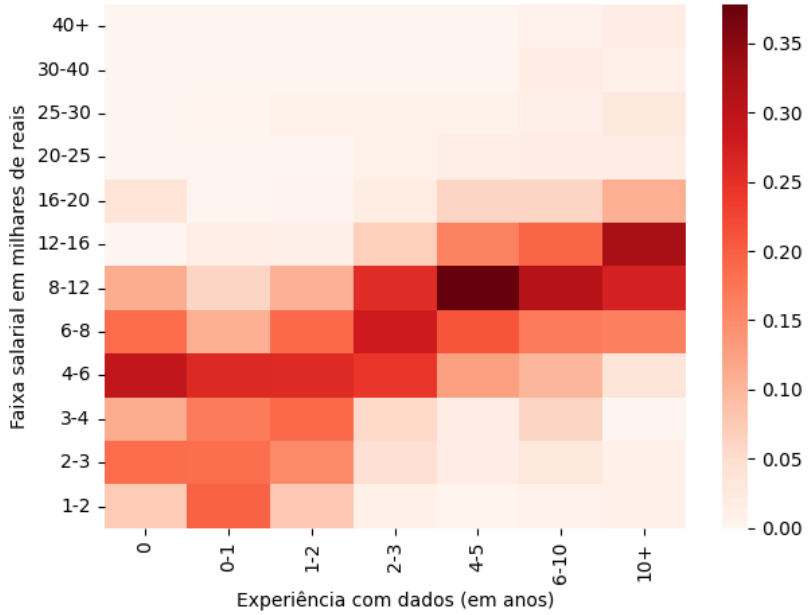

Figura 9 – Distribuição salarial dos desenvolvedores em relação à experiência com dados

A Figura 10 exibe a distribuição salarial dos profissionais gestores em relação ao tempo de experiência na área de dados. Entre os gestores sem experiência prévia na área de dados o salário predominante vai de R$6.000,00 até R$8.000,00, enquanto que os profissionais que recebem os melhores salários, aqueles acima de R$40.000,00, todos possuem quatro anos ou mais de experiência com dados.

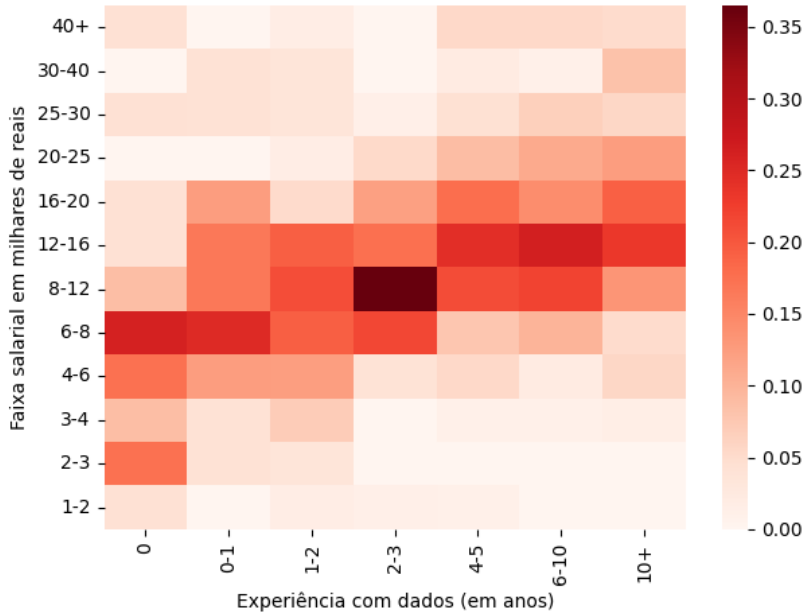

Figura 10 – Distribuição salarial dos gestores em relação à experiência com dados

4.1.3 Bases de dados: desenvolvedores

4.1.3.1 Nível do cargo e área de atuação

As *features* que representam as áreas de atuação e o nível do cargo são somente da base de dados dos desenvolvedores. Ambas são de característica qualitativa, sendo a *feature* área de atuação do tipo qualitativa nominal, e a *feature* nível de cargo do tipo qualitativa ordinal.

A variável área de atuação representa qual a área de dados que o empregado está envolvido, ou seja, se a pessoa é analista de dados, cientista de dados ou engenheiro de dados, e a variável nível de cargo apresenta a hierarquia dos cargos entre os valores estagiário, júnior, pleno e sênior.

A Figura 11 mostra a distribuição dos desenvolvedores de acordo com sua área de atuação no mercado de dados brasileiro.

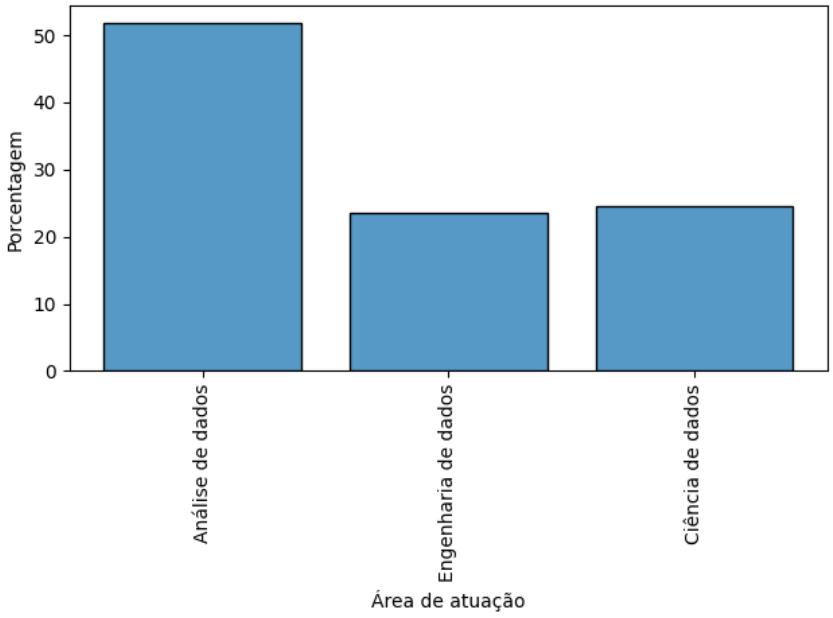

Figura 11 – Distribuição dos desenvolvedores em relação à área de atuação

É possível perceber que os profissionais analistas de dados dominam o mercado, representando mais de 50%do total de desenvolvedores. Em segundo lugar aparece os profissionais de ciência de dados, com apenas alguns poucos desenvolvedores a mais que aqueles que atuam como engenheiros de dados, que aparecem em terceiro.

A Figura 12 exibe a distribuição salarial dos profissionais desenvolvedores em relação as suas áreas de atuação: análise de dados, ciência de dados e engenharia de dados.

Os profissionais analistas de dados apresentam salários inferiores àqueles dos cientistas e engenheiros de dados, estando o salário mais bem pago na faixa de R$16.000,00 a R$20.000,00, e tendo sua maior concentração entre R$4.000,00 e R$12.000,00.

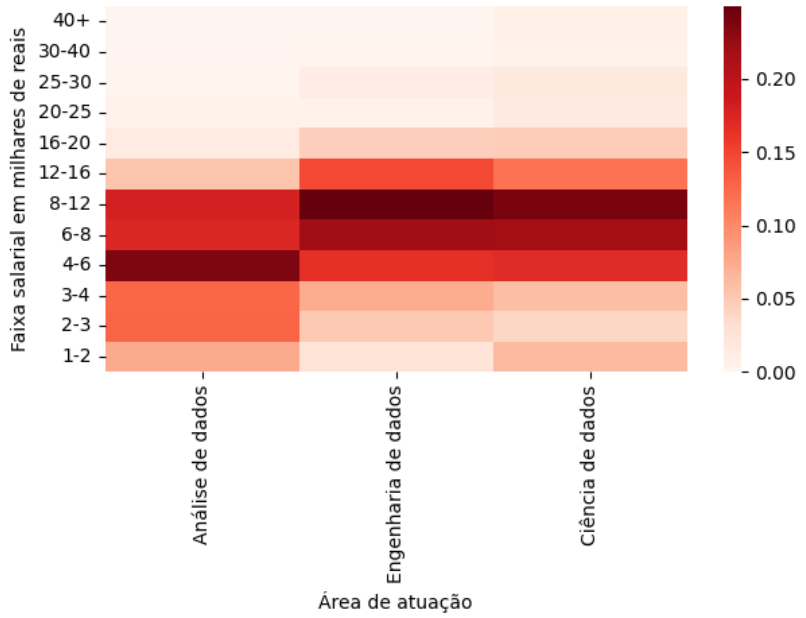

Figura 12 – Distribuição salarial dos desenvolvedores em relação à área de atuação

Os profissionais que atuam com engenharia de dados vem depois dos analistas em relação à remuneração, possuindo uma grande concentração de desenvolvedores nas faixas salarias de que vão de R$4.000,00 até os R$16.000,00, podendo chegar até os R$30.000,00/mês. Os engenheiros de dados são ainda os profissionais com maior salário mínimo, iniciando em R$2.000,00, enquanto o salário dos analistas e cientistas se iniciam em R$1.000,00.

Em relação aos profissionais desenvolvedores que atuam como cientistas de dados, esses possuem a distribuição salarial bem parecida com a distribuição dos engenheiros de dados, também possuindo uma maior concentração de desenvolvedores recebendo entre R$4.000,00 até os R$16.000,00, porém são os únicos profissionais da área de dados que recebem salários acima dos R$30.000,00/mês, sendo então os mais bem pagos.

O salário recebido pelos profissionais desenvolvedores é muito bem definido pelo nível do cargo de atuação, o que pode ser visto na Figura 13, que mostra que mais de 60% dos estagiários recebem salários de até R$2.000,00, enquanto o salário dos juniores se inicia nos R$2.000,00, tendo sua maior concentração entre R$4.000,00 e R$6.000,00.

O salário dos plenos se concentra nos valores entre R$4.000,00 e R$12.000,00, possuindo poucas amostras fora dessa concentração. Já o salário dos seniores possui sua maior distribuição entre os valores R$8.000,00 e R$16.000,00, e esses são os únicos profissionais que recebem uma remuneração salarial acima de R$16.000,00/mês.

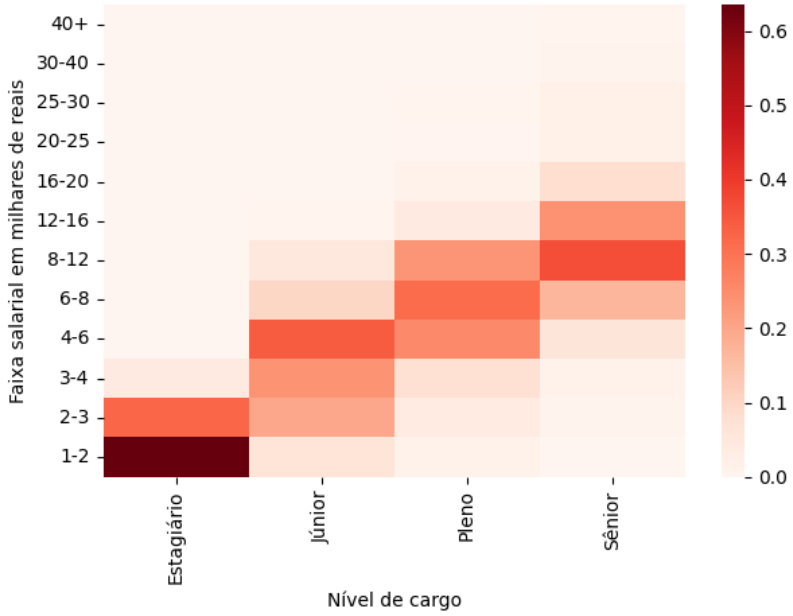

Figura 13 – Distribuição salarial dos desenvolvedores em relação ao nível do cargo

4.1.3.2 Linguagens utilizadas

Como mostra a Figura 14, as três linguagens mais utilizadas entre os profissionais desenvolvedores na área de dados são SQL, uma linguagem amplamente utilizada em bancos de dados relacionais, Python, uma linguagem de alto nível de uso geral, e R, linguagem muito utilizada por profissionais que trabalham com estatística computacional. Sendo SQL e Python utilizadas por aproximadamente 60% desses profissionais, enquanto R é utilizada apenas por pouco menos de 15%.

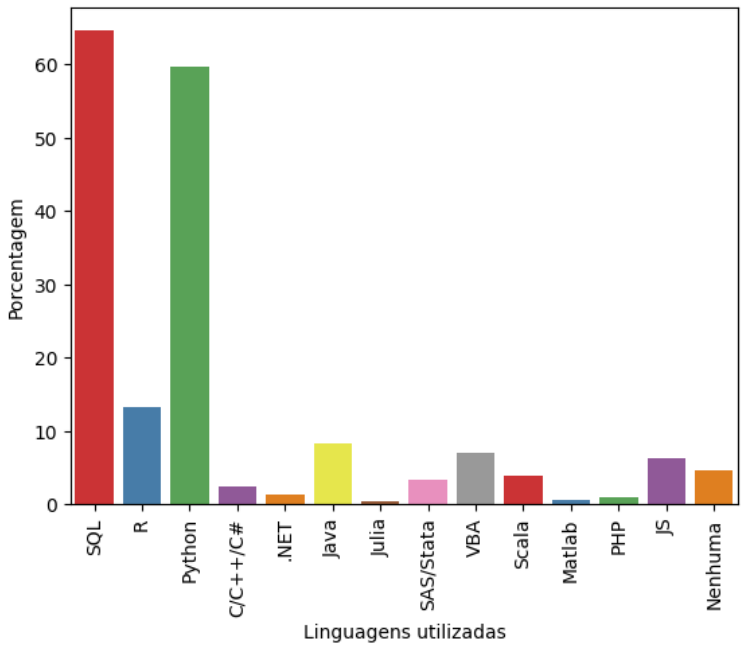

Figura 14 – Linguagens utilizadas pelos profissionais desenvolvedores

Entre as outras linguagens, as mais utilizadas são Java, Visual Basics (VBA) e JavaScript, que são linguagens comumente utilizadas no desenvolvimento de aplicações desktop, web, entre outras.

Cerca de 5% dos profissionais declararam não utilizar nenhuma linguagem no seu cotidiano na área de dados, o que reflete o crescimento do desenvolvimento utilizando ferramentas low/no-code.

4.1.4 Bases de dados: gestores

4.1.4.1 Cargo como gestor

A *feature* cargo como gestor está presente apenas na base de dados dos profissionais de gestão e é do tipo qualitativa ordinal. Ela representa a hierarquia dos cargos de gestores em empresas e possui os valores Team/Tech Leader, Supervisor/Coordenador, Gerente/Head/Diretor/VP e Sócio/C-level.

A Figura 15 mostra a distribuição dos gestores em relação aos cargos de gestão que os mesmos ocupam.

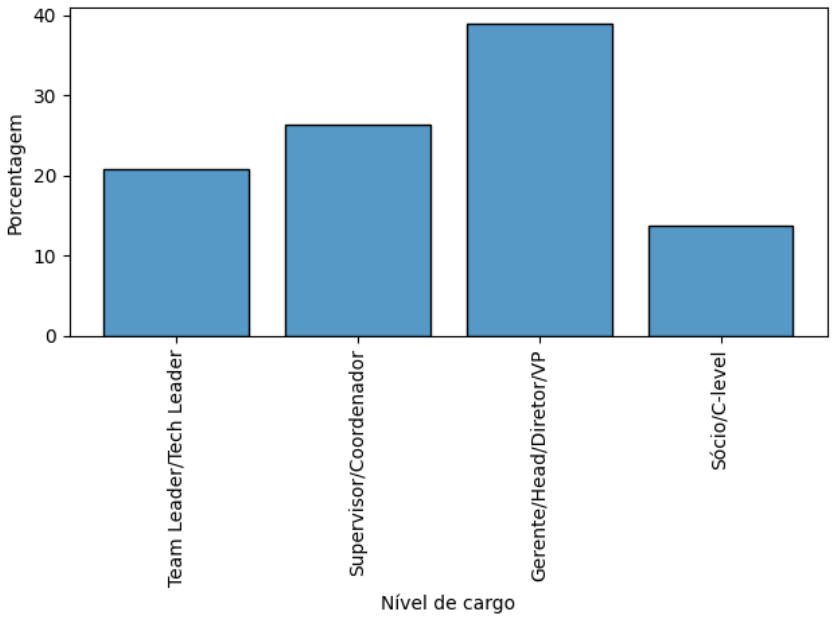

Figura 15 – Distribuição dos gestores em relação ao nível do cargo

O cargo mais recorrente do mercado é o de gerentes e diretores, que concentra quase 40% dos cargos de gestão. Os supervisores e coordenadores se mostram na segunda posição, com uma quantidade entre 25 e 30% de gestores, e são seguidos pelos Leaders, que representam pouco mais de 20%, enquanto os sócios e C-level são, aproximadamente, apenas 13% dos cargos de gestão.

Na Figura 16 é exibida a distribuição das faixas salariais dos profissionais gestores em relação nível do cargo que estão alocados.

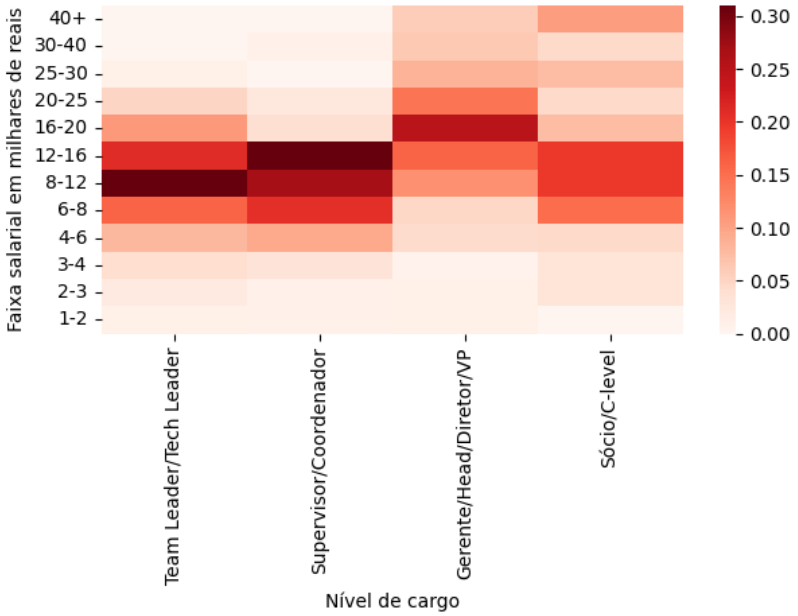

Figura 16 – Distribuição salarial dos gestores em relação ao nível do cargo

É possível compreender que o salário dos profissionais gestores se iniciam em R$4.000,00, pois existe uma muito pequena quantidade de gestores que recebe salários menores que esse valor.

Entre os Team/Tech Leaders, o salário varia entre os R$6.000,00 e R$16.000,00, sendo mais concentrado na faixa salarial entre R$8.000,00 e R$12.000,00. Os salários pagos aos supervisores e coordenadores também estão na faixa entre R$6.000,00 e R$16.000,00, porém, diferente dos Leardes, possui uma maior concentração nos valores entre R$12.000,00 e R$16.000,00/mês.

Entre os gerentes e diretores a faixa salarial entre R$16.000,00 e R$20.000,00 se mostra como moda, mas esses profissionais podem chegar a receber salários de mais de R$40.000,00, enquanto isso, os sócios e C-level possuem seus salários mais concentrados entre os R$6.000,00 e R$16.000,00, assim como os Leaders e os supervisores e coordenadores, mas é entre esses profissionais que se mostra a maior concentração dos maiores salários, aqueles acima dos R$40.000,00.

4.1.4.2 Responsabilidades

Na seção de desafios dos gestores de times de dados, os participantes tiveram de selecionar quais as responsabilidades que eles tem que lidar no cotidiano como gestores nas empresas que trabalham com dados. A lista dessas responsabilidades pode ser vista abaixo:

- R0: Pensar na visão de longo prazo de dados da empresa e fortalecimento da cultura analítica da companhia;
- R1: Organização de treinamentos e iniciativas com o objetivo de aumentar a maturidade analítica das áreas de negócios;
- R2: Atração, seleção e contratação de talentos para o time de dados;
- R3: Decisão sobre contratação de ferramentas e tecnologias relacionadas a dados;
- R4: Sou gestor da equipe responsável pela engenharia de dados e por manter o Data Lake da empresa como fonte única dos dados, garantindo a qualidade e confiabilidade da informação;
- R5: Sou gestor da equipe responsável pela entrega de dados, estudos, relatórios e dashboards para as áreas de negócio da empresa;
- R6: Sou gestor da equipe responsável por iniciativas e projetos envolvendo Inteligência Artificial e Machine Learning;
- R7: Apesar de ser gestor ainda atuo na parte técnica, construindo soluções/análises/modelos etc;
- R8: Gestão de projetos de dados, cuidando das etapas, equipes envolvidas, atingimento dos objetivos etc;
- R9: Gestão de produtos de dados, cuidando da visão dos produtos, backlog, feedback de usuários etc;
- R10: Gestão de pessoas, apoio no desenvolvimento das pessoas, evolução de carreira.

Na Figura 17 pode ser observada a ocorrência desses desafios entre os gestores. Por motivos de visualização foram utilizadas legendas na imagem de acordo com a lista de responsabilidades vista acima.

É possível perceber que o desafio mais enfrentado pelos gestores é o de *pensar na visão de longo prazo de dados da empresa e fortalecimento da cultura analítica da companhia*, que foi selecionado por cerca de 15% dos profissionais, seguido por *gestão de pessoas, apoio no desenvolvimento das pessoas, evolução de carreira* e *sou gestor da equipe responsável pela entrega de dados, estudos, relatórios e dashboards para as áreas de negócio da empresa*, que são desafios enfrentados por, aproximadamente, 12% dos profissionais gestores.

Vale observar que o quarto desafio mais enfrentado pelos gestores é *apesar de ser gestor ainda atuo na parte técnica, construindo soluções/análises/modelos etc*, o que talvez os coloque em uma posição mais próxima dos profissionais que se declararam desenvolvedores do que dos próprios profissionais de gestão.

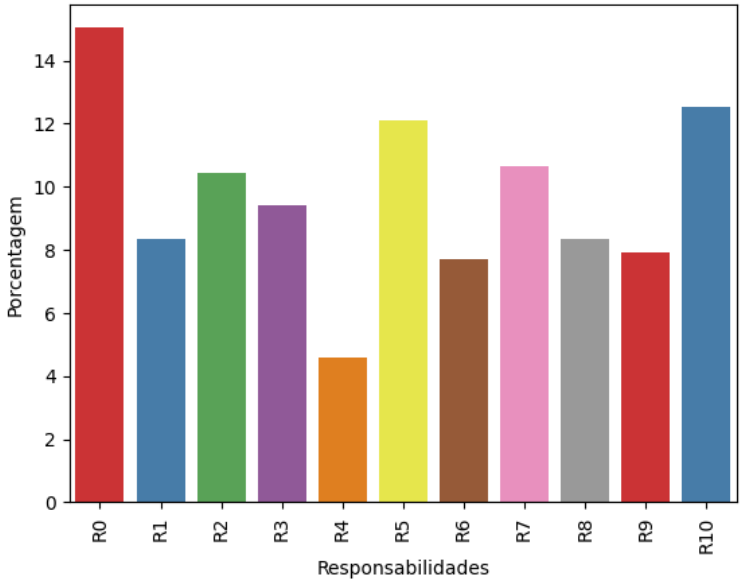

Figura 17 – Responsabilidades dos profissionais que atuam em cargos de gestão

4.2 Machine Learning

Toda a parte de ML foi desenvolvida utilizando a biblioteca *scikit-learn*(16), uma biblioteca para ambientes de desenvolvimento *Python*, que tem o intuito de ser simples e intuitiva, facilitando sua utilização, e que é *open-source*, tendo seu código disponível no GitHub[5] [6].

Os algoritmos utilizados na produção desse trabalho foram selecionados com o auxílio das bibliotecas PyCaret(17) e AutoML | Auto-Sklearn(18). A seleção foi feita de acordo com a acurácia dos algoritmos de classificação, sendo selecionados os três melhores para cada base de dados, listados abaixo:

- Desenvolvedores: Extra Trees e Random Forest;
- Gestores: Ada Boosting e Gradient Boosting.

Embora a seleção dos algoritmos foi realizada com base na acurácia, a métricas utilizadas na avaliação dos algoritmos foram o *root mean square error*, ou RMSE, e as matrizes de confusão. Todos os resultados dos modelos de ML apresentados nesse trabalho foram gerados utilizando a técnica de validação cruzada *StratifiedKFold*.

A primeira observação do desempenho dos algoritmos foi feita após o processamento da base original, que resultou nas bases dos desenvolvedores e gestores, sendo que o algoritmo que obteve o melhor desempenho para os desenvolvedores foi o *Random Fo-*

[5] https://github.com/
[6] https://github.com/scikit-learn/scikit-learn

rest, com um RMSE de 3.69, e o algoritmo que apresentou melhor desempenho sobre a base dos gestores foi o *Gradient Boosting*, com um RMSE de 8.36.

As figuras 18 e 19 apresentam como os algoritmos *Random Forest* e *Gradient Boosting* classificaram as instâncias de desenvolvedores e gestores. Pode-se observar como as predições tendem a se concentrar nas classes que possuem mais instâncias nas bases de dados.

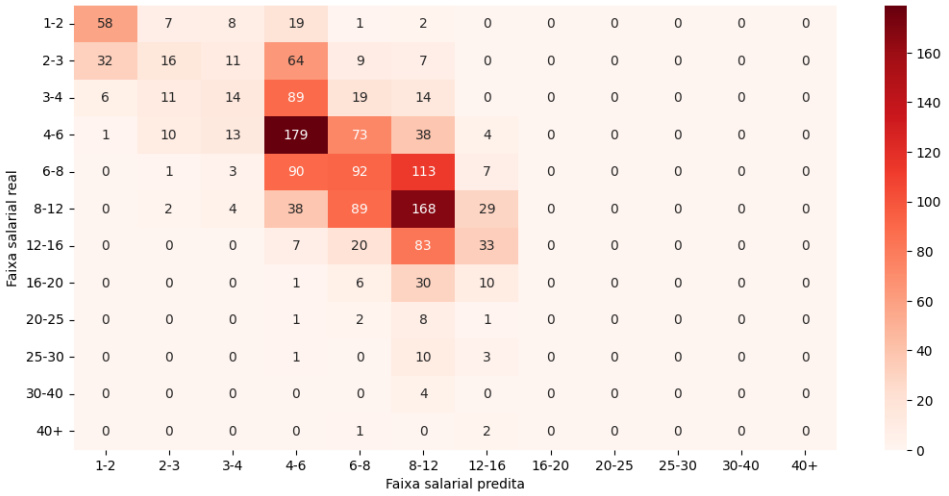

Figura 18 – Matriz de confusão da base de desenvolvedores

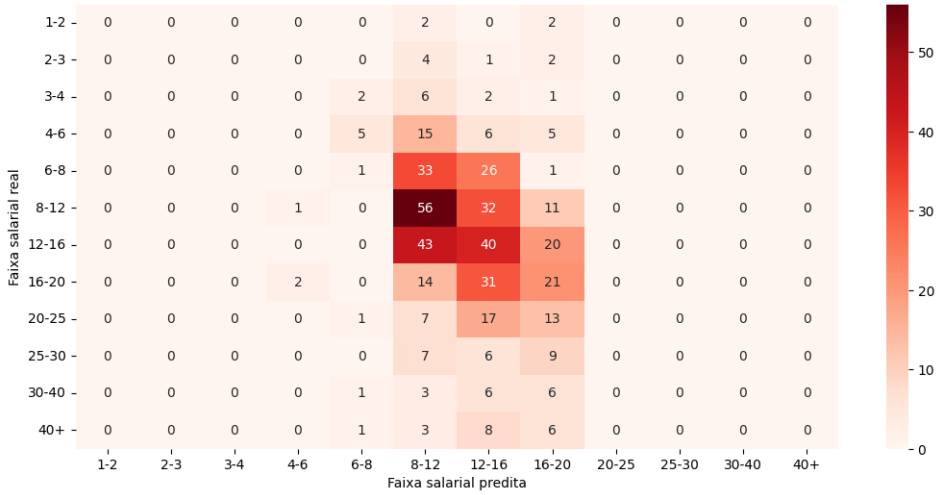

Figura 19 – Matriz de confusão da base de gestores

4.2.1 Pré-processamento

Em busca de melhorar os resultados obtidos na classificação, os dados foram submetidos à três etapas de pré-processamento:

4.2.1.1 Seleção das classes

Como pode ser observado nas matrizes de confusão das figuras 18 e 19 existe um desbalanceamento muito grande nas bases, afinal, são muito poucos os desenvolvedores que recebem salários acima dos R$16.000,00 e os gestores que recebem menos que R$4.000,00.

Buscando uma melhoria nas classificações, a primeira etapa do pré-processamento foi a de selecionar quais classes de faixa salarial representam mais as bases utilizadas. Para isso foram selecionadas, em cada uma das bases, apenas as classes que possuem uma razão maior ou igual a 1:5 em relação com a classe majoritária, resultando nos seguintes valores (os valores entre () representam a quantidade de instâncias que a classe possui):

- Desenvolvedores: (95) 1-2, (139) 2-3, (153) 3-4, (318) 4-6, (306) 6-8, (330) 8-12, (143) 12-16;
- Gestores: (31) 4-6, (61) 6-8, (100) 8-12, (103) 12-16, (68) 16-20, (38) 20-25, (22) 25-30.

4.2.1.2 Oversampling

Mesmo após a etapa de seleção das classes, as bases ainda continuavam bastante desbalanceadas, para isso foi aplicado um método de geração de instâncias sintéticas em ambas as bases, o *SMOTE*(19), com o objetivo de diminuir a razão entre a classe majoritária e todas as outras, levando a classe à uma razão de 1:2.

A quantidade de instâncias criadas para cada classe foi estabelecida pela equação 4.1

$$sintéticas_i = \alpha(n\_majoritárias - n\_classe_i) \tag{4.1}$$

onde sintéticas$_i$ representa a quantidade de instâncias criadas para a classe i, $n\_majoritárias$ é a quantidade de instâncias que a classe majoritária possui, n_classe_i é a quantidade de instâncias na classe i e α é uma variável, no intervalo [0,1], que controla a quantidade de instâncias sintéticas geradas baseada na diferença entre a quantidade de instâncias das classes majoritária e i.

No trabalho desenvolvido, onde buscou-se uma razão de 1:2 entre as classes minoritária e majoritária, para a classe de desenvolvedores foi utilizado α=0.3, e para a classe de gestores α=0.36.

Após a geração das instâncias sintéticas, a quantidade de instâncias nas bases passou a ser a seguinte:

- Desenvolvedores: (165) 1-2, (196) 2-3, (206) 3-4, (321) 4-6, (313) 6-8, (330) 8-12, (199) 12-16;
- Gestores: (56) 4-6, (76) 6-8, (101) 8-12, (103) 12-16, (80) 16-20, (61) 20-25, (51) 25-30.

4.2.1.3 Feature selection

A terceira etapa do pré-processamento das bases foi a de *feature selection*, na qual foi utilizada a técnica *SelectFromModel*, que seleciona *features* baseando-se na importância das mesmas para um modelo de ML. Foram utilizados os métodos *Random Forest* para a base de desenvolvedores, e *Gradient Boosting* para os gestores, ambos com as configurações padrão.

As *features* selecionadas pelos modelos foram as seguintes:

- Desenvolvedores: faixa etária, região, nível de ensino, regime de trabalho, forma de trabalho, setor de atuação, tamanho da empresa, tempo de experiência na área de dados, tempo de experiência prévia na área de TI, área de atuação e nível.
- Gestores: faixa etária, região, nível de ensino, regime de trabalho, forma de trabalho, setor de atuação, tamanho da empresa, tempo de experiência na área de dados, tempo de experiência prévia na área de TI e cargo de gestão.

Após a fase de pré-processamento as novas bases passaram por uma outra avaliação no objetivo de escolher quais algoritmos seriam utilizados na próxima etapa. Os métodos foram selecionados de acordo com seus valores de RSME, que podem ser vistos abaixo:

- Extra Trees: 2.16;
- Random Forest: 2.17;
- Ada Boosting: 5.91;
- Gradient Boosting: 5.56;

Para a base de gestores foi selecionado o método *Gradient Boosting*, e para a base de desenvolvedores foram selecionados os métodos *Random Forest* e *Extra Trees*, por apresentarem resultados similares, e esses métodos foram então submetidos a um processo de *Hyperparameter Tuning* em busca dos melhores modelos de ML para as bases.

4.2.2 Hyperparameter Tuning

As técnicas de aprendizado de máquina selecionadas no final do pré-processamento foram submetidas a um *Hyperparameter Tuning* exaustivo, o que tem como objetivo encontrar as melhores configurações dos algoritmos afim de gerar os modelos utilizados nas classificações finais das bases de desenvolvedores e gestores.

Para os algoritmos *Random Forest* e *Extra Trees*, os hiper parâmetros selecionados para a otimização foram a quantidade de estimadores, o critério de seleção de *feature* no momento do *split*, a quantidade de *features* analisadas no *split*, a utilização de *bootstrap* e *out-of-bag score*, *warm start* e atribuição de peso às classes de acordo com sua quantidade de instâncias.

No algoritmo *Gradient Boosting* foram selecionados os parâmetros quantidade de estimadores, a função de perda (*loss*), taxa de aprendizado, critério de seleção de *feature* no *split*, altura/profundidade máxima dos estimadores, quantidade máxima de *features* consideradas no *split* e *warm start*.

Após a geração de milhares de modelos no processo de *tuning*, os modelos utilizados nas classificações finais foram os algoritmos com as configurações de hiper-parâmetros abaixo:

- Extra Tress: quantidade de árvores: 1.250, critério de seleção de *features* no *split*: entropia, utilização de *bootstrap*, atribuição de peso às classes de acordo com sua quantidade de instâncias: balanceado e quantidade máxima de *features* analisadas no *split*: 0.6, sendo o resto dos hiper-parâmetros mantidos com seus valores padrão.
- Gradient Boosting: taxa de aprendizado: 0.175, altura/profundidade máxima dos estimadores: 5, quantidade máxima de *features* analisadas no *split*: log2, sendo o resto dos hiper-parâmetros mantidos com seus valores padrão.

5 Resultados

O resultado apresentado pelo modelo do *Extra Trees* utilizado indica que é possível predizer os salários de profissionais desenvolvedores com uma precisão de aproximadamente 53.6% e pouco menos de 2.13 de RMSE.

Pela figura 20 é possível perceber uma grande melhoria na classificação em relação à primeira observada, pois agora todas as predições estão concentradas próximas a diagonal principal, e todas as classes possuem a maioria de suas instâncias preditas corretamente.

Ainda assim o modelo não consegue aprender muito bem sobre as classes, podendo-se observar que existem instâncias sendo classificadas além de suas classes vizinhas, principalmente nas que são classificadas como faixas salariais maiores.

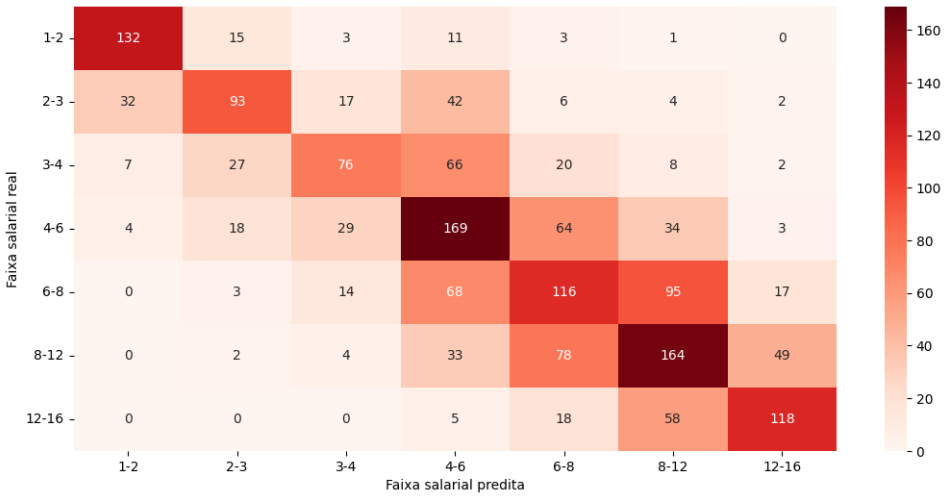

Figura 20 – Matriz de confusão da base de desenvolvedores

As predições para os profissionais gestores já se mostram mais complicadas do que as dos desenvolvedores, tendo o modelo atingido uma precisão próxima de 48.6%, um RMSE de pouco mais de 5.3.

A classificação dos gestores obteve uma melhora mais significativa que a dos desenvolvedores, estando agora as predições mais próximas da diagonal principal, sendo que na classificação inicial as predições se concentravam quase todas nas faixas entre R$8.000,00 e R$20.000,00, porém ainda se mostra pior, possuindo concentrações maiores de predições errôneas espalhadas pelo mapa, como pode ser visto na figura 21.

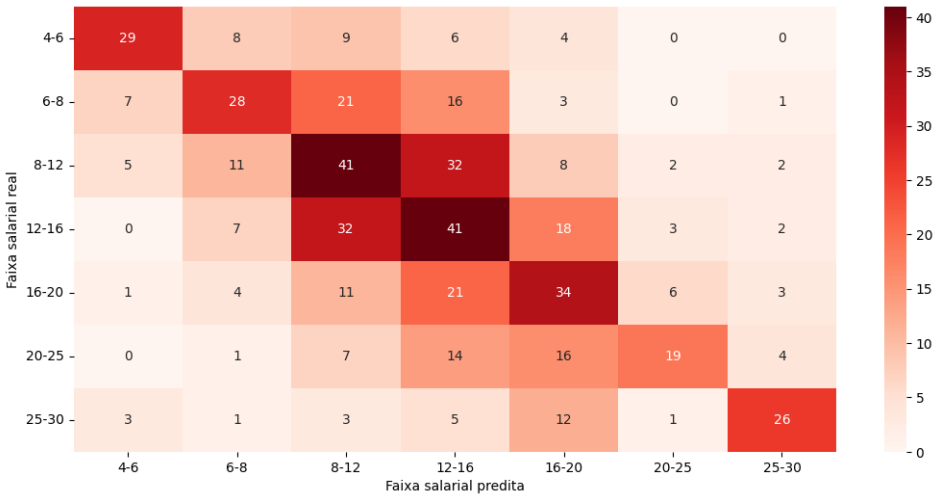

Figura 21 – Matriz de confusão da base de gestores

6 Conclusões e trabalhos futuros

É possível concluir que, apesar de terem ocorrido melhorias nas classificações ainda não se pode realizar predições muito precisas de acordo com o perfil dos profissionais, desenvolvedores e gestores, no mercado brasileiro.

O fato de essas profissões não serem regulamentadas no Brasil pode ser um fator que leva à essa falta de padrão nos cálculos das remunerações dos profissionais, que muitas vezes tem suas carteiras assinadas como outras profissões. Essa falta de regulamentação tem sido bastante discutida nos últimos anos, e já foi até pauta no senado nacional.

Outros fatores que podem ser considerados são, por exemplo, a diferença existente entre as remunerações para profissionais CLT e CNPJ; o tamanho das empresas: pode-se observar no mercado que, várias vezes, as empresas *startups* não possuem tanto capital para investir em seus empregados como empresas maiores e mais consolidadas; a disponibilidade de vagas na região onde os candidatos moram, fator que afeta principalmente àqueles que procuram por vagas presenciais; entre outros.

Parte da conclusão fica por conta das análises de perfis dos profissionais do mercado brasileiro, que, embora as classificações não tenham obtidos resultados "bons", essas análises podem auxiliar as pessoas que almejam uma vaga no mercado, ou até mesmo aquelas que já trabalham na área.

Entre os desenvolvedores, algumas *features* qualitativas ordinais exibiram boas correlações com a faixa salarial, como o nível de ensino, por exemplo. De acordo com as informações adquiridas dos profissionais que fazem parte do mercado, investir em educação e se qualificar pode ser uma boa ideia para aqueles que procuram salários mais altos.

Na base dos gestores, a *feature* que indica o nível de ensino já não exibe tanta correlação com o a faixa salarial, mas pode-se observar que a variável de tamanho da empresa, que quase não apresenta correlação com o salário na base dos desenvolvedores, é uma das que se destaca, sendo possível compreender que para os profissionais gestores pode ser interessante buscarem por empresas maiores caso estejam em busca de salários melhores.

Como trabalho futuro é pretendido adquirir dados periodicamente, para que possam ser realizadas novas análises no intuito de manter as predições salariais sempre atualizadas e poder compreender as mudanças que ocorrem no mercado.

Também é pretendido desenvolver uma aplicação, utilizando as análises e modelos de ML gerados no decorrer do trabalho, na qual os usuários possam preencher um formulário com informações sobre seu perfil profissional e recebam como *feedback* uma

predição salarial aproximada, gráficos e *dashboards*, para que possam compreender como estão relacionados com o a área de dados do mercado atual.

Referências

1 MITCHELL, T. M.; MITCHELL, T. M. *Machine learning*. [S.l.]: McGraw-hill New York, 1997.

2 CUNNINGHAM, P.; CORD, M.; DELANY, S. J. Supervised learning. *Machine learning techniques for multimedia: case studies on organization and retrieval*, Springer, p. 21–49, 2008.

3 BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, p. 5–32, 2001.

4 BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, p. 123–140, 1996.

5 DIETTERICH, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine learning*, v. 32, p. 1–22, 1998.

6 GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. *Machine learning*, Springer, v. 63, p. 3–42, 2006.

7 MINGERS, J. An empirical comparison of selection measures for decision-tree induction. *Machine learning*, Springer, v. 3, p. 319–342, 1989.

8 FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, Elsevier, v. 55, n. 1, p. 119–139, 1997.

9 HASTIE, T.; ROSSET, S.; ZHU, J.; ZOU, H. Multi-class adaboost. *Statistics and its Interface*, International Press of Boston, v. 2, n. 3, p. 349–360, 2009.

10 FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001.

11 FRIEDMAN, J. H. Stochastic gradient boosting. *Computational statistics & data analysis*, Elsevier, v. 38, n. 4, p. 367–378, 2002.

12 SOUSA, R.; SOUSA, R.; BRITO, R.; XIMENES, J. Utilização de modelos computacionais baseados em classificadores para predição da dislexia em crianças. In: *Anais do XIV Encontro Unificado de Computação do Piauí e XI Simpósio de Sistemas de Informação*. Porto Alegre, RS, Brasil: SBC, 2021. p. 113–119. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/enucompi/article/view/1776>.

13 RODRIGUES, E.; GOUVEIA, R. Técnicas de machine learning para predição do tempo de permanência na graduação no Âmbito do ensino superior público brasileiro. In: *Anais do VI Congresso sobre Tecnologias na Educação*. Porto Alegre, RS, Brasil: SBC, 2021. p. 128–137. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php-/ctrle/article/view/1755>.

14 COLPO, M.; PRIMO, T.; AGUIAR, M. Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. In: *Anais do XXXII Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2021. p. 873–884. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/1811>.

15 CAPANEMA, C.; SILVA, F. Detecção de pontos de interesse e predição de próximo local de visita de usuários móveis com base em dados esparsos. In: *Anais Estendidos do XXXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. Porto Alegre, RS, Brasil: SBC, 2021. p. 129–136. ISSN 2177-9384. Disponível em: <https://sol.sbc.org.br/index.php/sbrc_estendido/article/view/1716>.

16 PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

17 ALI, M. *PyCaret: An open source, low-code machine learning library in Python*. [S.l.], April 2020. PyCaret version 1.0. Disponível em: <https://www.pycaret.or>.

18 FEURER, M.; KLEIN, A.; EGGENSPERGER, K.; SPRINGENBERG, J.; BLUM, M.; HUTTER, F. Efficient and robust automated machine learning. In: CORTES, C.; LAWRENCE, N.; LEE, D.; SUGIYAMA, M.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. v. 28. Disponível em: <https://proceedings.neurips.cc/paper/2015/file/11d0e6287202fced83f79975ec59a3a6-Paper.pd>.

19 CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002.