UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO DE JOINVILLE
CURSO DE ENGENHARIA MECATRÔNICA

MARIANA ANTOSZ SIESLEVSKI

# APLICAÇÃO DE TÉCNICAS DE CIÊNCIA DE DADOS PARA O MOVIMENTO EMPRESA JÚNIOR

Joinville
2022

MARIANA ANTOSZ SIESLEVSKI

# APLICAÇÃO DE TÉCNICAS DE CIÊNCIA DE DADOS PARA O MOVIMENTO EMPRESA JÚNIOR

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do título de bacharel em Engenharia Mecatrônica no curso de Engenharia Mecatrônica, da Universidade Federal de Santa Catarina, Centro Tecnológico de Joinville.

Orientador: Prof. Dr. Benjamin Grando Moreira

Joinville
2022

Dedico este trabalho a meus pais, meus irmãos e minha família.

## AGRADECIMENTOS

Sempre tive comigo que a jornada é tão importante quanto a chegada, por isso, gostaria de agradecer algumas pessoas que fizeram a jornada da faculdade e do tempo de TCC mais fácil.

Agradecer meus pais, Francisco e Suzana que sempre me apoiaram nas minhas decisões e me incentivaram a ir cada vez mais longe. Aos meus irmãos Tadeu e Alícia que sempre estiveram comigo. À minha vó Terezinha e todos os meus familiares que sempre me deram exemplo de luta, determinação e me apoiaram em todos os meus passos.

Aos meus amigos e meus companheiros de jornada que além de me apoiarem estiveram comigo em cada passo, àqueles que chegaram ao longo dessa jornada e àqueles que já estão na minha vida antes disso e em especial às pessoas com que tive prazer de morar junto durante essa jornada, vocês foram essenciais para que eu chegasse até o final. Obrigada por estarem aqui em todos os momentos difíceis, desde estudos antes de uma prova, até o choro da reprovação e principalmente por estarem nos momentos bons comemorando comigo as minhas vitórias. Além disso, agradeço à todos aqueles com quem pude fazer trabalhos em conjunto e que dessa maneira puderam contribuir com a minha formação. Agradecer à todos que passaram pelo time de futsal feminino da UFSC e aos times das copinhas com quem pude compartilhar bons momentos, vocês deixaram a faculdade mais leve.

Ao Movimento Empresa Júnior que foi a minha segunda faculdade e que pode me ensinar coisas que a sala de aula nunca conseguiu, que me deu responsabilidades que eu nunca imaginaria ter com 22 anos e que me trouxe amizades e pessoas incríveis para a minha vida. Obrigada à ESATI Jr que foi minha porta de entrada e minha segunda casa na UFSC Joinville por 2 anos e todos aqueles que passaram por ela durante esse tempo, à FEJESC que me abriu os olhos para o tamanho do movimento e que me trouxe diversos aprendizados e ao Núcleo Vale Boreal que me deu o privilégio de liderar o movimento junto com duas mulheres incríveis e todo o nosso time e à todos que confiaram no meu trabalho durante esses 4 anos de jornada.

Ao meu orientador Prof. Dr. Benjamin Grando Moreira que foi peça chave para este trabalho, me trazendo grandes aprendizados e contribuições para esta entrega e a todos os professores que já fizeram parte da minha trajetória na UFSC Joinville, vocês me transformaram em uma melhor profissional e sou muito grata por todo o conhecimento partilhado.

## RESUMO

Este trabalho visa apresentar técnicas de *data science* aplicadas à empresas juniores de modo a trazer para gestores maior previsibilidade de seus dados para uma melhor tomada de decisão. Com o auxílio da ferramenta Orange que disponibiliza algoritmos de ciência de dados sem a necessidade de codificação por parte do usuário, foi possível aplicar os conceitos de previsibilidade e agrupamento por meio de algoritmos de classificação, regressão e agrupamento. O trabalho entrega um novo modelo de agrupamento das empresas juniores, os resultados das previsões de faturamento para o ano seguinte e preve se a empresa júnior alcançará suas metas ou não também, além de disponibilizar os fluxos para que os gestores possam utilizar seus próprios dados para outras previsões. O método pôde ser validado qualitativamente por atuais membros do movimento empresa júnior e mostrou-se útil e intuitivo para uso de gestores no seu dia-a-dia.



**Palavras-chave**: análise de dados; movimento empresa júnior; Orange; previsão; ciência de dados.

**ABSTRACT**

This final paper intends to show data science techniques applied to junior enterprises data, in a way to bring to managers more predictability of their own data to improve the process of decision making. With the tool Orange that provides data science algorithms without the need for user coding it was possible to apply the concepts of predictability and clustering through algorithms of classification, regression and clustering. The final paper delivers a new model of clustering the junior enterprises, the predictions of the next year billing and if the junior enterprise will reach out its goals or not, besides of making available all the workflows so the managers can use their own data bases to make other predictions. The method could be qualitatively validated by current members of the junior enterprise movement and proved to be useful and intuitive for use by managers in their daily basis.

**Keywords**: data analysis; junior enterprise movement; Orange; predictions; data science.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE SIGLAS

EJ Empresa Júnior

FEJESC Federação das Empresas Juniores de Santa Catarina

MEJ Movimento Empresa Junior

ODS Objetivos de Desenvolvimento Sustentável da ONU

**SUMÁRIO**

# 1 INTRODUÇÃO

As empresas juniores (EJs) são associações civis sem fins lucrativos, formadas e geridas totalmente por estudantes do ensino superior, que tem como principal objetivo contribuir com a formação universitária e profissional através da vivência empresarial, por meio do gerenciamento das áreas de uma empresa e execução de projetos, fazendo com que esses estudantes possam ajudar a transformar o Brasil em um país mais empreendedor (BRASIL JúNIOR, 2020).

A primeira empresa júnior surgiu na França em 1967, na L'Ecole Supérieure des Sciences Economiques et Commerciales de Paris (ESSEC ), e foi denominada como Junior Enterprise, tendo como principal objetivo proporcionar uma realidade empresarial para os alunos e prestar consultoria para empresas de mercado (BECKER; SILVA, 2017). No Brasil, a primeira empresa júnior surgiu em São Paulo, na Fundação Getúlio Vargas (FGV) e foi denominada como Júnior GV, o movimento foi se expandindo pelo país e em 1990 já existiam sete empresas juniores (BECKER; SILVA, 2017).

Em 1993 foi realizado o primeiro Encontro Nacional das Empresas Juniores (ENEJ), e a partir de então, alguns estados sentiram a necessidade de criar federações que pudessem representar as empresas juniores e garantir a unidade do movimento. Em função disso, e pela vontade de trazer o ENEJ para Santa Catarina, a Federação de Empresas Juniores de Santa Catarina (FEJESC) surgiu em 1994 (BECKER; SILVA, 2017). Em 2003 foi criada a Brasil Junior (BJ), pois, com o desenvolvimento e avanço do movimento, foi necessário criar uma entidade que pudesse representar as empresas juniores em todo o Brasil (BECKER; SILVA, 2017).

O Movimento Empresa Júnior (MEJ) é um movimento que unifica as mais de 1500 empresas juniores do Brasil e, atualmente, é baseado em um planejamento estratégico trienal construído com diversos agentes, por meio de análises e dados. As empresas juniores do Brasil estão em 299 Insituições de Ensino Superior (IES), nas 27 unidades federativas do país e presentes em mais de 3400 cursos espalhados nessas IES, os três cursos com maior número de EJs são administração, engenharia civil e engenharia de produção (BRASIL JúNIOR, 2021a).

Como há grande diversidade de cursos, IES e tamanho das EJs dentro do MEJ, os projetos realizados também são diversos, sendo os mais comuns entre as Ejs, no ano de 2021, o registro de marca, pesquisa de mercado, rotulagem nutricional, projeto arquitetônico, plano de negócios e mapeamento de processos. As empresas juniores são responsáveis por vender, gerenciar e executar todos esses projetos e em 2021 o MEJ faturou quase 69 milhões de reais e executou mais de 41mil projetos (BRASIL JúNIOR, 2021b).

Um exemplo dessas Ejs é a SMART Consultoria Jr., empresa júnior de engenheria de produção e sistemas da Universidade do Estado de Santa Catarina (UDESC) em Joinville, a SMART é uma EJ que vende serviços de pesquisa de mercado, mapeamento de processos, 5S, LEAN e planejamento estratégico. Em 2021 a SMART vendeu 40 projetos e faturou pouco mais de R$150.000,00, sendo que desses projetos, 15 foram executados com outras EJs, um princípio importante do planejamento estratégico de 2019-2021 (BRASIL JúNIOR, 2021b).

Dados, atualmente, são um dos maiores ativos de uma empresa, com isso, cada vez mais as empresas se movem para tomar todas as decisões baseadas em dados. Segundo a Forrester (FORRESTER, 2021) , 4036 executivos em 45 países e cerca de 73% das empresas brasileiras, dizem ter negócios data driven (orientado em dados), porém, só 28% das empresas consideram o tratamento dos dados como prioridade. Já no mundo, 66% das empresas se dizem data driven, mas apenas 21% realmente priorizam o tratamento. Para que isso seja possível e eficiente, é necessário ter uma boa base de dados, e que também seja feita uma análise desses dados de maneira precisa.

Como o mercado sênior ainda precisa se desenvolver na utilização dos dados, o mesmo ocorre para as EJs que, principalmente desde 2017 têm conseguido coletar dados anuais sobre metas e desempenho. Mesmo que a Brasil Júnior já tenha algumas ferramentas de tratamento desses dados, a aplicação de data science pode tornar seguramente previsíveis alguns desafios ou facilitar ainda mais as tomadas de decisão, e esse é o objetivo principal deste trabalho.

Com o crescimento do uso de dados para tomadas de decisão e outros aspectos gerenciais das empresas, é fundamental que a aplicação de data science seja acessível para que o gerente de uma empresa não precise aprender a programar para isso. Por isso, este trabalho visa o desenvolvimento de ferramentas que facilitem tal inclusão, e uma ferramenta que corresponde a essas expectativas é a orange, uma ferramenta open source que permite criar todo o fluxo de um projeto de data mining sem a necessidade de código (ORANGE DATA MINING, 2022c).

A partir da montagem desses fluxos de projeto será possível prever processos comerciais, dando mais clareza sobre as vendas de uma EJ, até prever dados gerais sobre o MEJ, como o faturamento e número de projetos do ano do movimento, o que é importante para mensurar o real impacto do que vem sendo feito. Assim como é possível a mesma aplicação para empresas do mercado sênior, que podem prever as vendas ou mesmo como seus consumidores se comportam em determinada compra.

Trabalhos relacionados a estes pode-se citar o desenvolvido por José Hercos (JUNIOR, 2014) que faz uma análise da eficiência de empresas juniores por meio de modelos matemáticos, a grande diferença entre os trabalhos está na metodologia, o autor faz o uso de uma metodologia puramente matemática, assim como há a diferença

do movimento, visto que de 2014 para 2021 houve um grande aumento no número de empresas juniores no Brasil, assim como o crescimento em maturidade de empresas juniores. Assim como, o trabalho de Hercos é uma tese para doutorado, dessa maneira o nível técnico da parte matmemática é mais complexo.

Outro trabalho que se relaciona ao tema de empresas juniores é o de Jhonata Souza e Matheus Galvão (SOUZA; GALVÃO, 2019), no trabalho os autores analisam os processos de gestão da empresa júnior UTIC, da Universidade Federal Rural do Amazonas, por meio da coleta e análise de dados qualitativos com os membros da empresa. As principais diferenças se dão pela difernça no método da análise e também na finalidade do trabalho, visto que os autores trazem uma visão de análise dos processos do negócio e não utilizam de conceitos de *data mining*. Neste trabalho os autores fazem análises qualitativas por meio de entrevistas o que nãe será foco neste trabalho, que analisará dados quantitativos de planilhas disponibilizadas pela Brasil Júnior (SOUZA; GALVÃO, 2019).

## 1.1 OBJETIVOS

### 1.1.1 Objetivo Geral

Utilizar de técnicas e algoritmos de ciência de dados para criar *workflows* que auxiliem gerentes do movimento empresa júnior a terem mais previsibilidade em seus dados.

### 1.1.2 Objetivos Específicos

- Utilizar uma plataforma na qual não seja necessário o usuário desenvolver códigos para utilizar os algoritmos de ciência de dados;
- Prever faturamento e o alcance de metas das empresas juniores para o ano seguinte;
- Propor uma nova maneira de agrupar as empresas juniores.

# 2 FUNDAMENTAÇÃO TEÓRICA

## 2.1 CONCEITOS GERAIS SOBRE CIÊNCIA DE DADOS

As informações hoje são encontradas pela internet ou relatórios internos e com esse montante de informação é necessário saber ler e esculpir essas informações para que elas valham para uma tomada de decisão assertiva. A análise de dados faz com que possamos sair do dado bruto para encontrarmos o dado elaborado, via interpretação, análise e síntese, que nos fará entender como utilizar esse dado da melhor maneira.

Um dado pode ser caracterizado como um fato que é coletado e que geralmente é armazenado. O processo de vida de um dado consiste desde o seu surgimento, com um sensor, por exemplo, o seu armazenamento e então a sua transformação. Depois dessas etapas, um dado é possível de ser analisado e visualizado e após servir para a sua necessidade, ele pode então ser descartado. A ciência de dados envolve todo este ciclo de vida e são os processos e tecnologias que estudam os dados durante este tempo (AMARAL, 2018).

Os conceitos referentes a *data science* são diversificados e aplicam-se a diferentes áreas do conhecimento. Dessa maneira, pode-se citar o conceito da tecnologia da informação, que pode ajudar a encontrar atributos informativos que sejam do interesse do cientista ou de um negócio. Basicamente, nos negócios usa-se de *data science* para extrair conhecimento útil desses dados para resolver problemas.

Dentro desses conceitos, um importante é o de mineração de dados, processo este que tem estágio muito bem definidos e decompõe a tarefa de encontrar padrões a partir de um conjunto de dados. Quando são utilizados um grande número de dados e se quer encontrar informações úteis desses dados, é neccesário entender qual a decisão que precisa ser tomada e o que é útil para essa decisão, com esse pensamento, chega-se no conceito de formular soluções de mineração de dados e avaliar os resultados envolve pensar cuidadosamente sobre o contexto em que essas informações serão utilizadas (FAWCETT; PROVOST, 2018).

A partir deste grande conceito de mineração de dados é possível entrar em alguns pontos específicos de sub-classificações para entender com mais detalhes. Os quatro grande estágios da mineração de dados são definir objetivos, prepração dos dados, aplicação de algoritmos de mineração de dados e avaliar os resultados. Estes 4 estágios citados descrevem de maneira geral tudo o que já foi citado até aqui (INTERNATIONAL BUSINESS MACHINES CORPORATION, 2021a).

O estágio de definição de objetivos do negócio é um dos mais importantes tanto para a organização quanto para facilitar o trabalho do cientista de dados, por

isso, é uma etapa que precisa ser de conversa entre todos os *stakeholders* envolvidos (INTERNATIONAL BUSINESS MACHINES CORPORATION, 2021a). Depois de definido o objetivo, serão coletados e preparados os dados que são relevantes para se chegar na resposta necessária. Os dados nessa etapa podem ser refinados e podem ser coletados alguns dados que ainda estejam faltantes.

Após a definição dos objetivos e tendo os dados prontos, é possível aplicar algoritmos de mineração de dados para que esses dados sejam tratados e transformados em informações importantes. Nesta etapa é necessário avaliar o conjunto de dados existente e é importante entender qual a análise que precisa ser feita ou qual a saída desejada para o problema, para então encontrar o melhor algoritmo para cada caso.

Na última etapa, depois que os dados foram tratados, agora eles podem ser visualizados, avaliados e interpretados e dessa maneira podem serem aliados dos negócios, mostrando inforamções importantes que podem agregar no processo de tomada de decisão. A mineração de dados é de extrema importância e algumas das principais aplicações desta técnica envolvem vendas e marketing, a educação, otimização de operações e detecções de fraudes. Como um processo complexo e cheio de definições, há algumas sub-classificações da mineração de dados que serão detalhadas a seguir.

## 2.2 ABORDAGEM SUPERVISIONADA

Há 2 grandes tipos de abordagem em *data science*, são elas a abordagem supervisionada e a não supervisionada, a grande diferença entre elas se da em como os dados estão disponíveis, se eles estão rotulados ou não. Exemplos rotulados são aqueles que quando há uma entrada, a sua saída é conhecida (INTERNATIONAL BUSINESS MACHINES CORPORATION, 2021b). Na abordagem supervisionada tem-se os dados rotulados. Esse tipo de abordagem é utilizado geralmente para treinar ou supervisionar algoritmos que classificam os dados ou preevem os resultado. Nesta abordagem o algoritmo aprende por meio do treinamento com a base de dados a partir de previsões que vão sendo feitas e ajustadas com relação a resposta correta. A abordagem supervisionada pode ser dividida em dois tipos, classificação e regressão.

### 2.2.1 Classificação

Classificação é o processo de prever uma classe por meio de alguns dados disponíveis, ou seja, os problemas de classificação usam um algotirmo para atribuir dados em categorias específicas, como por exemplo separar spam na caixa de e-mail. A classificação é um modelo preditivo que tem como tarefa a aproximação de uma função mapeada de variáveis de entrada para variáveis discretas de saída.

Existem alguns tipo de algoritmos de classificação, dentre eles, classificação linear, árvore de decisões e *random forest* estão entre alguns mais comuns. A árvore de decisão é um tipo de algoritmo que envolve uma coleção de nós de decisão, que são conectados por galhos que se estendem do nó raiz para os nós folhas. Como início há sempre o nó raíz e a partir dele os atributos são testados nos nós de decisão e são levados pelos galhos até outro nó de decisão ou ao nó folha, dependendo do seu resultado (LAROSE; LAROSE, 2014).

Para que esse algoritmo seja possível de ser aplicado, são necessários seguir alguns critérios.

- Como árvores de decisão são parte de uma aprendizagem supervisionada, ela requer de variáveis pré-classificadas.
- O banco de dados fornecido deve ser rico e variado, pois como as árovres de decisão aprendem pelo exemplo é necessário o máximo de informações para que seja possível ter previsibilidade.
- A classe de destino dos atributos devem ser discretas. A variável de destino deve assumir valores que são claramente demarcados como pertencentes a uma determinada classe ou não pertencentes.

Um exemplo de aplicação de árvores de decisão são para avaliar se os consumidores são de alto ou baixo risco para compra. Com isso, são definidos atributos importantes para se definir isso e perguntas que levem aos nós finais (LAROSE; LAROSE, 2014). O algoritmo de árvore de decisão funciona da seguinte maneira:

- Primeiro há o carregamento da base de dados e dividisão de dados entre dados de treinamento e dados de teste;
- Treinamento da árvore de decisão a partir dos dados de treinamento;
- Usar a classificação para prever o nome da classe dos dados de teste;
- Calcular a acuracidade da previsão.

## 2.3 ABORDAGEM NÃO SUPERVISIONADA

Já a abordagem não supervisionada usa algoritmos para analizar e agrupar dados que não são rotulados, ou seja, a reposta certa não é informada ao sistema. O grande objetivo dessa abordagem é descobrir padrões ou uma estrutura nos dados. Esta abordagem não necessita de uma supervisão humana durante o seu processo, mas ainda precisam de uma ajuda humana para validar as saídas. A abordagem pode ser divida em três grandes tarefas, o agrupamento, a associação e a redução de dimensão.

Assim como na abordagem supervisionada, esse trabalho trará apenas uma tarefa na abordagem não supervisionada, sendo esta a associção. A associação é um método que utiliza de diferentes regras para encontrar relações entre variáveis em um

dado banco de dados. Este tipo de método é tipicamente utilizado em sugestões de compras em *e-commerce*, em que por exemplo, se o determinado cliente tem uma variação de idade e uma renda determinada, ele terá uma porcentagem de chance de comprar determiando produto (DOMINGUES, 2003).

Do método de regras associativas será destacado o algoritmo a priori. Este algoritmo tenta reduzir a busca do problema das regras associativas para um número mais fácil de ser gerenciado. Isto é feito a partir da propriedade que diz que se um item Z não é frequente, então se um novo item B for adicionado a ele, esse item Z não se tornará mais frequente. E desta maneira, nenhum conjunto que contenha Z será frequente para ser analisado. Matematicamente, essa propriedade nos diz que:

Se um item Z não for frequente então para qualquer item A, $Z \cup B$ não será frequente.

Para este algoritmo há três conceitos importantes, o de *support*, confiança e frequência. Pode-se definir *support* como a proporção de transições que acontecem os eventos A e B. Ou seja:

$$P(A \cap B) = \frac{\text{número de transições contendo A e B}}{\text{número total de transições}} \tag{1}$$

E confiança é a medida de acuracidade é determinada pela porcentagem de transições em D que contém A e também contém B. Pode-se definir como:

$$P(A \mid B) = \frac{P(A \cap B)}{P(A)} \tag{2}$$

Um item frequente é aquele que aparece pelo menos num certo mínimo de vezes, para isso, a frequência de aparições deste item precisa ser $> \phi$, em que $\phi$ é decido por meio da observação dos dados. Por isso, para começar um algoritmo precisa-se achar os itens frequentes daquela base de dados e a partir destes itens, geram-se regras que satisfaçam o mínimo de *support* e confiança necessários. A partir disso este algoritmo pode ser útil para entender o perfil de compra de um consumidor, por exemplo, ao passo que a partir do algoritmo pode-se entender se há a compra de um produto antecedente, então qual será a consequência desta compra, ou seja, a compra de um outro produto que geralmente os consumidores tendem a comprar de maneira conjunta.

Diferenças que podem ser relevantes na comparação entre as duas abordagens pode-se citar os objetivos de cada uma, sendo que a abordagem supervisionada é prever as saídas, já na não supervisionada é tirar informações e padrões de um grupo de dados. As aplicações de cada abordagem também são diferentes, um exemplo de aplicação para supervisionada é previsão do tempo, já um exemplo para a não supervisionada é detecção de anomalias. A complexidade também é bem diferente entre si, a supervisionada é mais simples. A não supervisionada é computacionalmente

complexa e você precisa de ferramentas poderosas para conseguir trabalhar com um grande número de dados.

## 2.4 ORANGE

O orange é uma ferramenta utilizada para *data mining*, que é *open source* e permite que o usuário consiga trabalhar sem a necessidade de programar o seu próprio código, isso porque a ferramenta do tipo *drag and drop* funciona como um *workflow* e consegue produzir desde pequenas análises até algumas mais complexas, por isso, é objeto de estudo de grande valia que irá ser utilizado neste trabalho já que permite a utilização de técnicas de ciência de dados por gerentes que não tenham tanto conhecimento técnico sobre o assunto.

A visão geral que tem-se ao abrir a ferramenta é de uma tela branca onde será produzido e montado o *workflow* e do lado esquerdo uma barra com *widgets* que são as funcionalidades do sistema, o conjunto de *widgets* produz o *workflow*. Para melhor organização, os *widgets* são separados em algumas categorias padrões, *data*, *transform*, *visualize*, *model* e *evaluate*, além dessas, é possível instalar outras categorias para diferentes tipos de análise, como análise de texto e de imagens, por meio de *add-ons*.

Em um primeiro momento só irão ser utilizadas as categorias padrões para o desenvolvimento deste trabalho, pois elas já são suficientes para o tipo de análise que será feita. A categoria *data* se refere aos dados, nela encontramos *widgets* referente a entrada de dados, como um arquivo de excel. Na categoria *transform*, existem métodos para organizar os dados, selecionar somente parte dos dados e até mesmo pré-processar. Em *visualize* encontra-se maneiras de visualizar os dados, em formas de gráficos, tabelas e outras formas. Na categoria *model* tem-se os princpais algoritmos para *data mining*, como árvore de decisão, regressão logística, redes neurais, entre outros. E em *evaluate* há *widgets* para avaliar o resultado do tratamento dos dados e fazer previsões desses dados.

A ferramenta possui diversos vídeos tutoriais num canal próprio no Youtube e também é intuitiva para ser utilizada, dessa maneira, a principal dificuldade que pode surgir é entender os conceitos por trás dos algoritmos e interpretação dos resultados essa também é uma barreira para que os conceitos e aplicação de *data mining* sejam mais difundidos nas organizações.

Como exemplo, pode-se montar um *workflow* básico e inicial para demonstras como a ferramenta funciona, para a montagem pode-se arrastar os *widgets* da barra do lado esquerdo até a tela branca ou mesmo clicar na tela que irão aparecer as sugestões. O primeiro *workflow* pode ser montado da seguinte maneira, com o *widget file* aplica-se uma base de dados, e conecta-se o *widget* data table, para que os dados possam

ser vistos na forma de tabela, liga-se o *file* a um dos algoritmos escolhidos e pode-se ligar o algoritmo ao *widget predictions* dessa maneira, é possível fazer previsões para aquela base de dados.

Ao realizar alguns testes, é possível ligar mais de um algoritmo à mesma base de dados e também ao mesmo *widget* de *predictions* por exemplo, dessa maneira é possível fazer comparações entre dois algoritmos diferentes, o que pode ser bem útil na hora da realização de análises. Ainda para analisar os resultados pode-se utilizar o *widget test and score* que dará detalhes como acuracidade e precisão para classificação, e dados como erro MSE para regressão.

Para o *widget test and score* temos um conjunto de dados de análise que podem ser vistos, os campos de resultados de avaliação que aparecem para algoritmos de classificação são AUC, CA, F-1, *Precision* e *Recall* ou revocação, já para algoritmos de regressão as métricas são MSE, RMSE, MAE e $R^2$. (ORANGE DATA MINING, 2022d)

### 2.4.1 Métricas de avaliação

AUC é a sigla para *Area under ROC* em português a área sob a curva ROC, nessa curva, os dados são classificados entre 0,0 e 1,0, com um limiar de 0,5. Os números que ficam acima desse limiar são classificados como testes bons, podendo evoluir de bons para excelentes e abaixo disso os testes são inúteis, quanto mais próximo de 1 o teste é mais preciso. (ORANGE DATA MINING, 2022d)

CA é a sigla para *Classification accuracy* que é a acurácia de classificação que mostra a proporção de exemplos que foram classificados de maneira correta. F-1 é uma média harmônica ponderada entre os outros dois termos, a precisão e a revocação. (ORANGE DATA MINING, 2022d)

A precisão é a proporção de verdadeiros positivos entre instâncias classificadas como positivo, por exemplo, se houvesse uma base de dados em que seria necessário classificar espécies de animais, seria a proporção de cachorros que seriam corretamente classificados como cachorros. E a revocação parecido com a precisão, avalia a capacidade do método de detectar com sucesso resultados classificados como positivos, acertando os resultados positivos dentre as métricas de verdadeiros positivos e falsos negativos. (ORANGE DATA MINING, 2022d)

A métrica MSE significa *Mean square error*, em tradução livre, erro quadrático médio, que mede a diferença entre o que é estimado e o estimador, usa a distância Euclidiana para calcular o erro e mostra a magnitude do erro.

RMSE significa *root mean squared error* ou erro quadrático médio, que da uma medida de imperfeição do ajuste do estimador aos dados, e é calculada pela raiz quadrada da média aritmética dos quadrados de um conjunto de números.

A sigla MAE é o *mean absolute error* que é o erro absoluto médio e é usado

para avaliar o quão próximo o resultado da previsão está dos resultados esperados. É utilizada a distância Manhattan para calcular este erro.

E o $R^2$ é uma métrica que vai de 0 a 1 e pode ser interpretada como a proporção da variância na variável dependente que é previsível da variável independente, quanto mais próximo de 1 for este valor, melhor a regressão está funcionando.

### 2.4.2 Pré-processamento de dados

Uma etapa importante para que se consiga realizar boas análises da base de dados escolhidas é fazer um bom pré-processamento dos dados, e isso é necessário em alguns casos, por exemplo quando há alguma valor faltante ou valores *outliers* aqueles muito maiores ou muito menores do que o comum. Em alguns conjuntos de dados analisados nos trabalho encontraram-se alguns valores faltantes, e a partir disso há diversas maneiras para pré-processar esses dados.

Uma maneira de tratar valores faltantes é remover os registros que tenham valores nulos, o que para este trabalho não parece a melhor opção visto que pode mascarar algumas informações. Pode-se também realizar uma média ou uma mediana com os valores do mesmo atributo ou então preencher o atributo que está faltando com os valores que mais se repetem na base de dados, para que isso seja feito da maneira correta é preciso olhar para a base de dados e para o objetivo do trabalho ou do negócio para entender a melhor maneira de susbstituir esse tipo de valor.

Além dessas maneiras, pode-se ainda transformar os dados, para que os dados estejam em formatos adequados para as análises, para isso pode-se fazer normalização, seleção de atributos, geração de hierarquia de conceitos ou discretização. Os principais que poderão ser utilizados são a normalização e a discretização, a normalização dimensiona os valores dos dados para um intervalo específico e a discretização transforma funções contínuas em discretas, o que pode ser importante porque alguns algoritmos só funcionam com variáveis discretas.

### 2.4.3 Previsões

Para encontrar resultados a partir dos dados pode-se usar de previsões com *data science*, para isso geralmente usa-se de uma base de dados, em que uma parte desses dados é utilizado como dados de treino e outra parte é usado como dados de teste, em que os dados de treinamento são utilizados para treinar o modelo e dados de testes são os utilizados para comprovar que aquele modelo realmente funciona. Ainda podem ser utilizados outros dados como dados de validação, que são usados para comparação de diferentes modelos, mas que não serão foco neste trabalho.

Para modelos de previsão geralmente são utilizados os dados de maneira sequencial, então se houverem 6 meses de dados é comum que os primeiros meses

sejam utilizados como dados de treino e os últimos meses como dados de teste. Geralmente, é utilizada uma proporção de 70% para 30% entre dados de treino e dados de teste, respectivamente.

Outra maneira que os dados podem ser preparados para a previsão a partir da base de dados pronta é a partir da validação cruzada. Neste método, a base de dados é separada em X números de grupos e segue-se um método a partir da decisão do número de grupos. É feita uma iteração para cada grupo, em que, o grupo é separado para teste, enquanto os outros são utilizados para treinamento. Treina-se o modelo com os dados de treino e obtêm-se os resultados com os dados de treino, os valores das métricas são salvos e o modelo é descartado, repete-se o processo para o próximo grupo (SAVIETTO, 2021).

Alguns testes empíricos em *machine learning* aplicada mostram que alguns valores que são confiáveis para o número de grupos são X = 5 ou 10, que geralmente não trazem valores de erros altos nem de *bias* ou variância. Em que o erro de *bias* é causado por suposições errôneas ou muito simplificadas que o modelo aprende durante o treinamento, se o *bias* for alto o modelo não conseguiu identificar relação entre os dados. Já a variância apresenta o quanto o modelo muda quando treinado com conjuntos de dados diferentes (SAVIETTO, 2021).

Outra avaliação importante a ser feita para comparar os dados que estão sendo utilizados é calcular a correlação entre eles, isso é feito no Orange por meio do *widget correlation*. Com esse *widget* é possível calcular qual a correlação entre pares dos dados utilizados, quanto mais próximo de 1 o valor da correlação mais esse par de dados está relacionado entre si linearmente, isso é importante para a previsão de faturamento com a regressão linear que será feita pois mostra os dados que são mais importantes de serem considerados durante o processamento (ORANGE DATA MINING, 2022a).

Para este cálculo de correlação podem ser utilizados dois métodos a relação de Pearson ou de Spearman, neste caso irá se utilizar do método de Pearson, que pode variar de -1 até 1, em que o sinal indica a direção entre as variáveis e o valor em módulo representa quão forte é essa relação. Não há um consenso para a interpretação dessas forças, alguns autores citam que associações forte estão entre valores de 0,5 e 1 enquanto outros autores citam que associações forte estão entre valores de 0,7 a 1, neste trabalho, irá se considerar valores fortes iniciando em 0,5 (FILHO; JÚNIOR, 2009).

# 3 ANÁLISE DOS DADOS

Para uma maior facilidade no entendimento do texto, o capítulo de análise dos dados irá tratar sobre o método desenvolvido e também os resultados obtidos. Dessa maneira, o capítulo será dividido em 3 secções cada uma com os métodos que foram desenvolvidos e os seus resultados.

O processo para o levantamento de dados deu-se pela procura de quais os dados necessários para a realização do trabalho e quais os tipos de dados. Primeiro decidiu-se por focar em dados quantitativos ao invés de dados qualitativos pois estavam disponíveis uma quantidade maior de dados que poderiam ser analisados, assim como de um maior número de EJs. A Brasil Júnior disponibilizou os dados de todas as empresas juniores do Brasil dos últimos 3 anos. Com a disponibilidade dos dados esses foram organizados em um dicionário de dados.

Como resultado da obtenção dos dados obteve-se um dicionário de dados com a organização dos dados dispostos em 5 colunas, mostrados na Tabela 13 no Apêndice A. No dicionário de dados a primeira coluna é o nome da variável e também o nome da coluna presente nas planilhas de dados originais, a segunda coluna descreve o que significa as variáveis da primeira coluna, na terceira coluna há o tamanho de dígitos ou caracteres que aquela variável precisa ter, a quarta coluna mostra qual o tipo da variável, podendo ser um número, texto, data, entre outros e a quinta coluna são os valores aceitos por essa variável. Nessa tabela mostra-se o intervalo de valores aceitos somente para variáveis do tipo numérica. Todos os dados foram retirados do portal da Brasil Júnior e os dados disponibilizados no portal respeitam a LGPD.

Depois de levantados e identificados os dados, foi necessário entender de que maneira eles seriam tratados, e então foi definida uma biblioteca ou ferramenta para este propósito. Há diversas bibliotecas que podem ser utilizadas em *machine learning* como TensorFlow, Keras [1] e outras, assim como ferramentas que permitem desenvolver *machine learning* sem o uso de códigos, como a WEKA bastante conhecida em Java, e o software Orange.

Após o entendimento sobre as possibilidades de ferramentas e bibliotecas e o alinhamento com o propósito do trabalho foi entendido que a Orange seria a ferramenta que conseguiria atender os requisitos de maneira satisfatória e que também se aproximaria a uma realidade de utilização por gestores ou colaboradores em uma empresa, por motivos já apresentados na fundamentação teórica.

[1] https://insightlab.ufc.br/8-bibliotecas-de-deep-learning-mais-usadas-em-python

## 3.1 ANÁLISE DE *CLUSTER* DAS EJS

No MEJ, *clusters* servem para dar à rede bases para as taxas anuais de crescimento e foram criados para gerar estímulos diretos para contribuir com o crescimento da rede, então dividiram-se as empresas juniores em 5 níveis de maturidade diferentes. Após as etapas de estudos variadas, desde entender a mensagem do Planejamento Estratégico da rede até a definição da nova régua de *cluster* e construção das dores por trás dos *cluster*, a Brasil Júnior conseguiu definir a fórmula que traria a divisão dos *clusters* e a sua régua para comparação que podem ser vistas na Equação 3 e Tabela 1 abaixo (BRASIL JúNIOR, 2019).

$$\textit{Cluster} = \%\ \text{de membros que executam} \times \frac{\text{faturamento}}{\text{membro}} \times \text{NPS} \tag{3}$$

A Equação 3 corresponde à fórmula utilizada para o cálculo do *cluster*, em que tem-se a multiplicação da porcentagem de membros que executam (a quantidade de membros que executaram pelo menos 1 projeto ao longo do ano divido pelo número total de membros da empresa júnior), pelo faturamento total da empresa júnior dividido pelo número de membros totais, multiplicado pelo *Net Promoter Score* (NPS). O NPS, uma métrica utilizada para entender a satisfação dos clientes, e é medida de 1 a 10 por meio da pergunta "De 1 a 10 o quanto você recomendaria nossa empresa/serviço para um familiar/amigo"?

Tabela 1 – Régua de *Cluster* pela Brasil Júnior.

| ***Cluster*** | **Régua** |
|---|---|
| 1 | Até 26320,55 |
| 2 | 26320,56 a 73950,00 |
| 3 | 73950,01 a 166666,71 |
| 4 | 166666,72 a 389536,50 |
| 5 | acima de 389536,51 |

Fonte: (BRASIL JúNIOR, 2019).

Na Tabela 1 observa-se como o cálculo da Equação 3 é utilizado para a separação dos *clusters*. Como já apresentado, são 5 *clusters* e eles são divididos por meio da régua, ou seja, a partir do valor obtido para a EJ, essa é classificada em um dos grupos conforme a Equação 1.

A partir desse entendimento, algumas perguntas podem ser levantadas como: seria essa a melhor maneira de desenvolver o cálculo dos *cluster*? Será que uma EJ é realmente similar as outras EJs do *cluster*? Porque além do número de *cluster* em si obtido por meio do cálculo e classificado pela régua, é importante que os gestores das EJs e dos núcleos e federações, tenham um pensamento crítico para avaliar as EJs e entender se as dificuldades daquele *cluster* fazem sentido para a sua realidade atual.

Por isso, é importante também que os gestores possam ter uma ferramenta

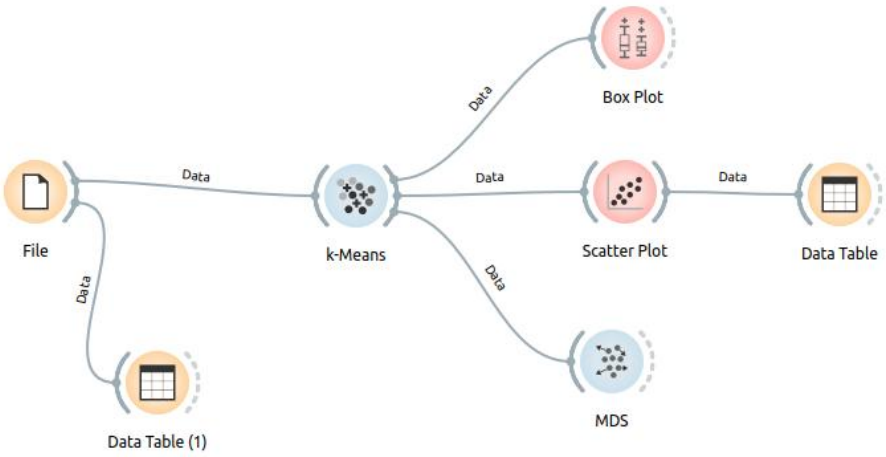
para que consigam entender como a sua EJ se comporta diante também de outros indicadores propostos pelo Planejamento Estratégico da rede e comparar a sua EJ às demais EJs para entender padrões de comportamento e ter mais clareza conforme a sua posição perante às outras EJs e assim poder traçar estratégias e planos de ações mais eficientes.

Para isso, foi criado um *worflow* no Orange para que os gestores possam fazer diversas comparações entre a sua EJ e as demais a partir de uma base de dados. Para a construção de um *workflow* 1 foi feita uma validação utilizando-se de dados das EJs da Federação das Empresas Juniores do Estado de Minas Gerais (FEJEMG) e foram utilizados 3 algoritmos para o teste de qual traria melhores resultados. Os escolhidos para testes foram o *cluster* hierárquico, *K-means* e *K-means* interativo.

O *K-means* interativo na ferramenta Orange aparece como um *widget* e ele funciona de maneira mais didática, que mostra passo-a-passo de como é o funcionamento do algoritmo *K-means*. O objetivo desse *widget* é mostrar como o *K-means* funciona a partir de uma entrada de dados. Esse *widget* aplica o *K-means* nos atributos de uma maneira passo-a-passo, assim o usuário consegue acompanhar como são posicionados os centróides e quais os movimentos que ele faz até a definição dos *clusters*. Basicamente, os resultados são iguais ao do *k-means* porém ele não é o utilizado neste trabalho pois o objetivo não é mostrar como o algoritmo funciona, mas sim os resultados obtidos a partir deste algoritmo (ORANGE DATA MINING, 2022b).

Depois da comparação entre os 3 algoritmos entendeu-se que o que melhor conseguiu trazer *insights* e resultados mais conclusivos foi o algoritmo *K-means*. Para obter os seus resultados foram utilizados 3 *widgets* associados ao algortimo, *box plot*, *scatter plot* e *Multidimensional scaling* ou (MDS), ainda foi associado o *widget data table* ao *scatter plot* para que seja possível fazer a ligação dos dados plotados ao nome das EJs e aos seus dados. O *workflow* 1 que foi montado para gerar os resultados pode ser visto na Figura 1 abaixo.

Figura 1 – *Workflow* 1 com o algoritmo *K-means*.

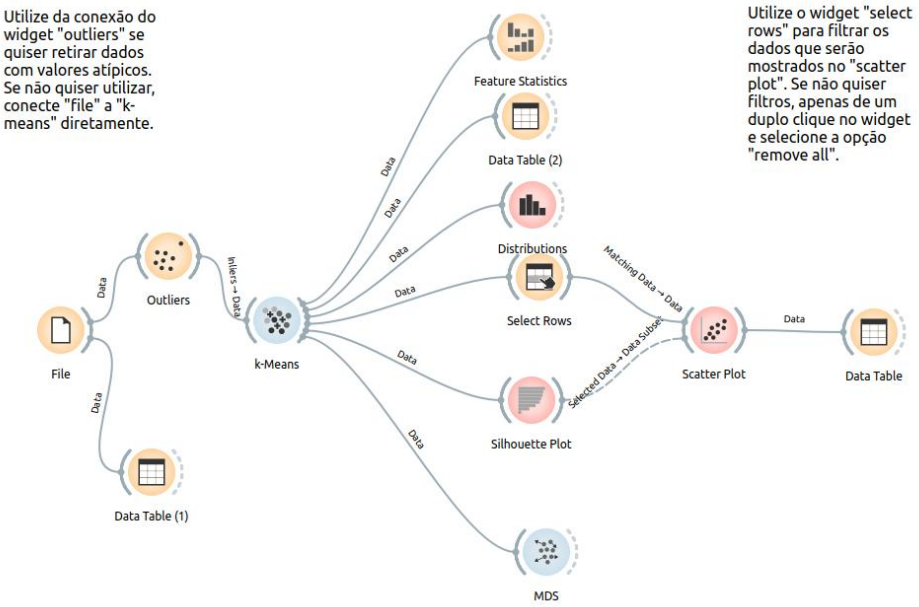



Fonte: Autora (2022).

Na Figura 1 pode-se ver o *worflow* montado na Orange em que tem-se como entrada uma base de dados depositada no *widget* File, ligada a um *Data Table* onde é possível ver os dados da base, e também ligada ao algoritmo *K-means* que é ligado em 3 *widgets* para visualizar os resultados.

A partir do *worflow* 1 da Figura 1 foram identificados alguns pontos de melhoria que poderiam ser feitos para que o gestor conseguisse extrair informações mais generalistas e ter as informações mostradas de uma maneira mais fácil para a sua análise. Por isso, alguns *widgets* foram adicionados e outros *widgets* foram retirados para facilitar a visualização e o entendimento dos dados, workflow 2 pode ser visto na Figura 2. O *Workflow* 2 foi criado para que a análise feita pelos gestores pudesse ser facilitada.

Figura 2 – *Workflow* 2 como nova versão do *Workflow* 1 com o algoritmo *K-means*.



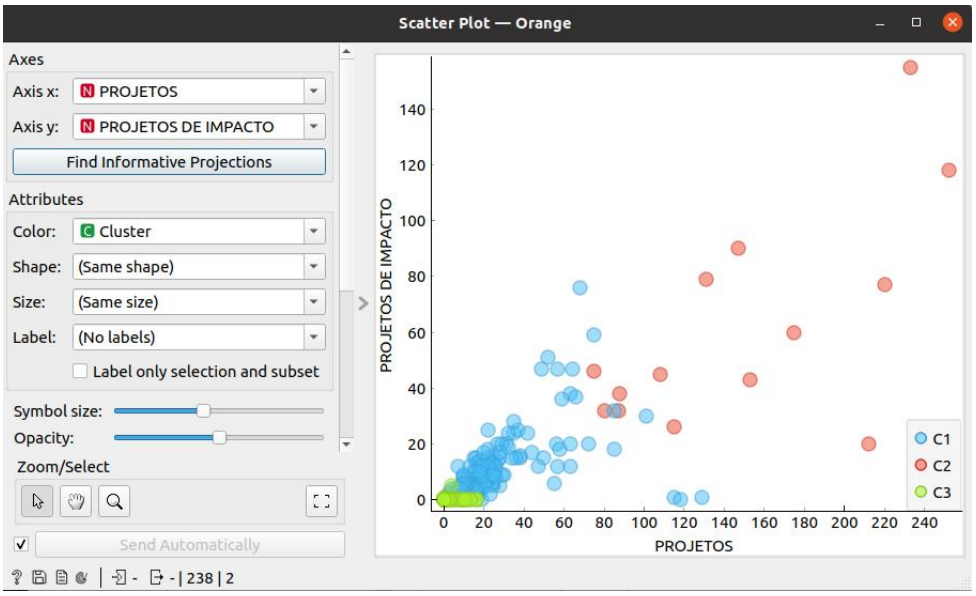

Fonte: Autora (2022).

O *workflow* 2 conta com 4 novos *widgets* de análise dos dados e um novo *widget* como um filtro antes do algoritmo de agrupamento. Os *widgets* de *Feature Statistics*, *Distribution Stats* e *Silhouette Plot* trazem uma visualização dos dados de maneira simplificada, já o *widget Select rows* trabalha como um filtro para a visualização em *scatter plot*. Esse filtro pode ser feito a partir de qualquer um dos atributos de entrada e também pelos *clusters*, assim como é possível fazer a combinação de várias condições para o filtro. Outro *widget* que serve como filtro é o *Outliers*, em que ele retira as EJs que tem dados considerados muito distintos dentre as outras.

Na Figura 2, também é possível notar que foram adicionadas algumas instruções para que o gestor possa modificar o *workflow* 2 para que seja mais adequado as necessidades. A primeira instrução mostra que o *widget outliers* pode ser desconectado do *workflow* 2 para que todos os dados sejam considerados no agrupamento, assim como também são mostradas instruções gerais para filtrar os dados com o *widget Select rows*.

Com o *worflow* da Figura 2 o gestor pode analisar como a sua EJ se encontra em relação as outras por diversas combinações de dados diferentes. Na base de dados utilizada como exemplo os dados utilizados foram de faturamento, projetos, porcentagem de membros que executam projetos, projetos que atingem alguma ODS, NPS e projetos de impacto. Ao selecionar 3 *clusters* fixos no algoritmo *K-means* pode-se ver no *widget scatter plot* os resultados mostrados na Figura 3.

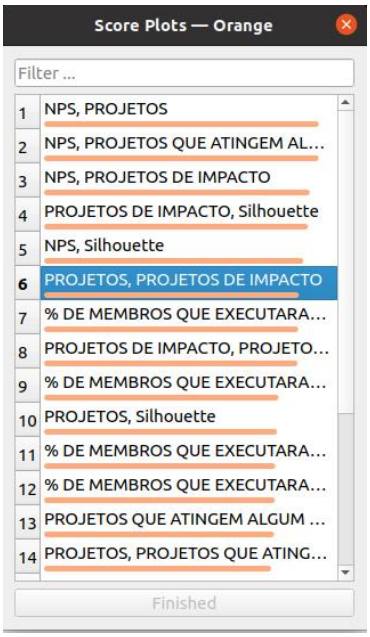
Figura 3 – *Scatter plot* aplicado ao algoritmo *K-means*.




Fonte: Autora (2022).

Caso o gestor queira fazer modificações nos parâmetros que serão analisados no gráfico ele pode mudar manualmente por meio dos botões em *"axis A"* e *"Axis B"* ou também ele pode clicar no botão *"Find Informative Projections"* que em tradução livre significa encontre projeções informativas. Dessa maneira será apresentado ao gestor diversas opções de combinações dos dados que podem gerar informações relevantes para a sua análise. Alguns exemplos de combinações apresentadas podem ser vistas na Figura 4.

Figura 4 – Combinações de informações projetivas.

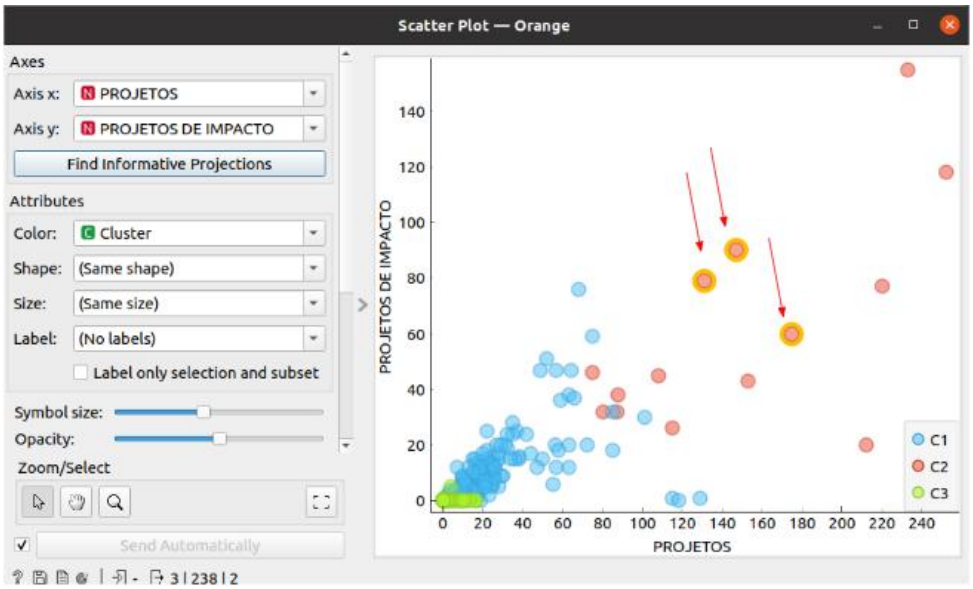



Fonte: Autora (2022).

Na Figura 4 mostram algumas combinações possíveis de serem feitas. Para selecionar alguma das combinações listadas o usuário apenas precisa clicar na combinação escolhida e fechar a caixa de *"Score Plots"*.

Para que os resultados encontrados sejam conclusivos e úteis para os gestores, os gráficos precisam ser ter ligação com o nome das EJs, ou seja, precisa ser fácil saber qual EJ é representada por qual círculo no gráfico. Para que isso seja possível, é adicionado o *widget* de *Data table* ao *Scatter Plot*, dessa maneira ao selecionar alguns círculos no gráfico, como explicitado pelas setas na Figura 5, eles serão destacadas na tabela com os dados das EJs na coluna *"Selected"*, conforme destacado na Figura 6.

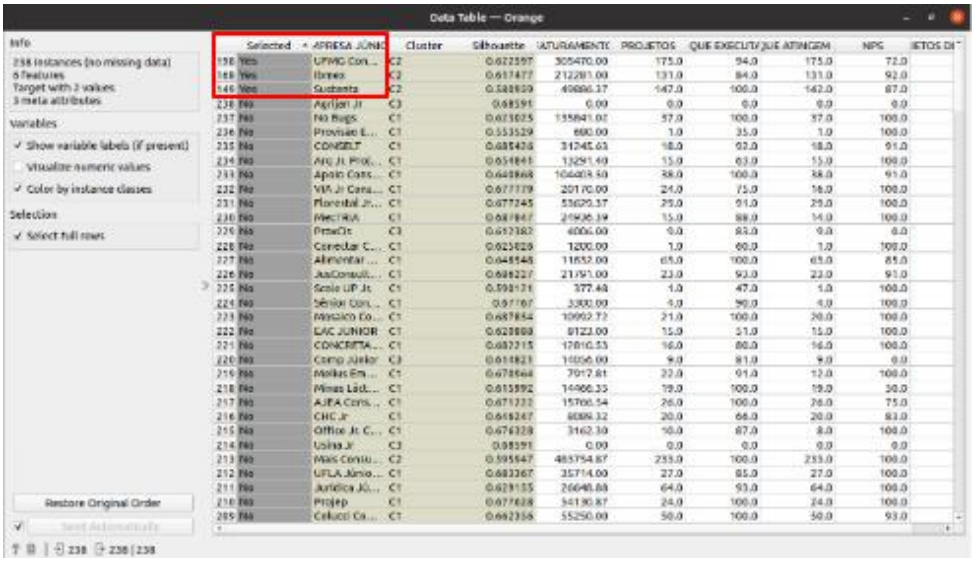
Figura 5 – *Scatter plot* com alguns círculos selecionados.




Fonte: Autora (2022).

Na Figura 5 pode-se observar 3 círculos que são do C2 selecionados, essa seleção serve para conseguir enxergar na tabela quais são as EJs selecionadas.

Figura 6 – Dados das EJs e indicação de quais foram as selecionadas na Figura 7.



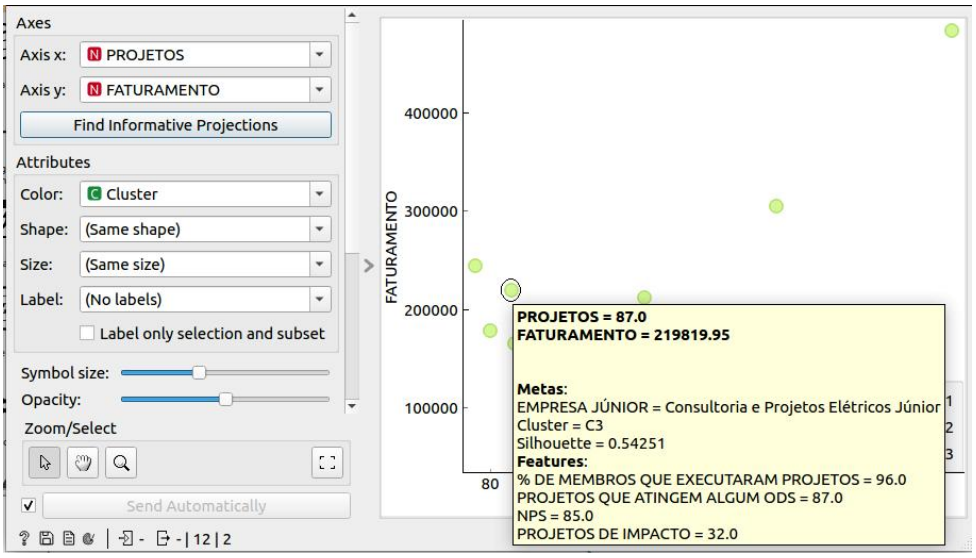

Fonte: Autora (2022).

A Figura 6 mostra que os círculos que foram selecionados no *Scatter plot* estão destacados com um *Yes* na coluna *Selected*, dessa maneira pode-se reconhecer qual a EJ que é representada pelo círculo no gráfico. Isso é importante para que o gestor

consiga identificar onde está a sua EJ e como ela se comporta perante as outras.

Outra maneira de saber qual EJ está sendo mostrada no *scatter plot* é posicionar o mouse em cima de um dos círculos, dessa maneira irá aparecer um *pop-up* com todas as informações correspondentes daquela EJ. Na Figura 7 é possível ver um exemplo, ao posicionar o mouse em cima do círculo que está marcado pode-se encontrar as informações contidas na planilha de entrada dos dados daquela EJ correspondente.

Figura 7 – *Pop-up* mostrando à EJ corresponde aquele círculo.

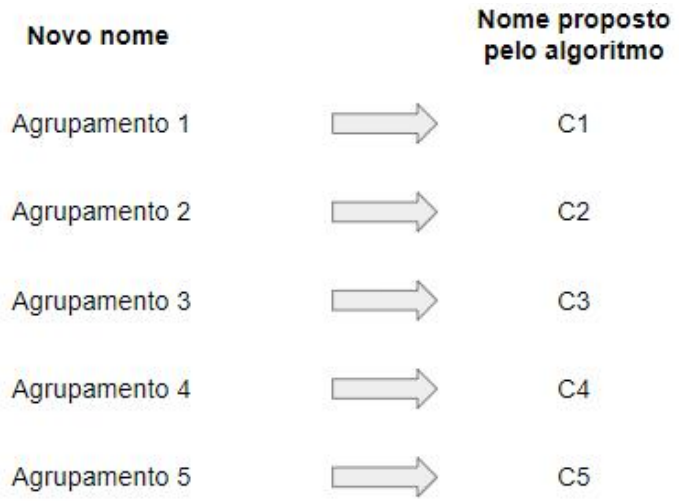



Fonte: Autora (2022).

Para compreender os resultados obtidos, foi feita a comparação entre os *clusters* atuais da Brasil Júnior e os *clusters* que foram obetidos por meio do algoritmo *k-means* e, para que a comparação fosse facilitada, o algoritmo foi programado para determinar 5 *clusters*. Este é o motivo pelo qual foram selecionados 5 novos agrupamentos ao invés dos 3 novos agrupamentos que foram apontados como ideal pelo *k-means* como apresentado na Figura 3.

Neste texto, para evitar confusão para o leitor já que tanto na classificação da Brasil Júnior e também no algoritmo utilizam a denominação de *cluster*, será feita uma alteração: os grupos propostos pelo algoritmo são chamados de agrupamento e a divisão feita pela Brasil Júnior a chamada de *cluster*, assim como explicitado na Figura 8.

Figura 8 – Correlação entre o novo nome de agrupamento e o nome proposto pelo algoritmo.



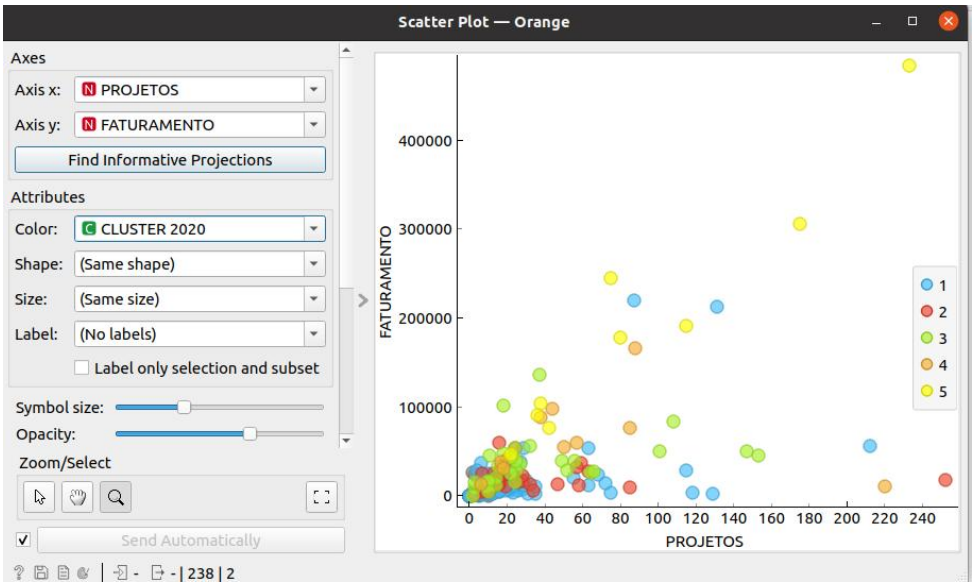

Fonte: Autora (2022).

Na Figura 8 pode-se notar qual deve ser a comparação feita do novo nome que será citado nesse texto com o nome proposto pelo algoritmo *k-means* que poderá ser notado em algumas Figuras adiante. É importante ressaltar que sempre que a palavra agrupamento for citada a autora está se referindo aos novos *clusters* sugeridos pelo algoritmo *k-means*.

Em um contexto geral os resultados foram satisfatórios pois uma parte do novo agrupamento gerado está coerente com a atual divisão dos *clusters* proposta, porém nota-se que a comparação precisa ser cuidadosa porque os agrupamentos gerados pelo algoritmo (C1,C2,C3,C4,C5) não podem ser comparados diretamento pelos seus iguais dos *clusters* atuais. Por exemplo, pelas Figuras 9 e 10 pode-se ver que as EJs de *cluster* 1 atualmente, no *cluster* gerado pelo algoritmo deve ser C1 ou C3.

A Figura 10 mostra como ficaram os 5 agrupamentos propostos a partir do algoritmo do *k-means* na visualização de *scatter plot* e a Figura 9 mostra como estão divididos os *clusters* atuais também no *scatter plot*.

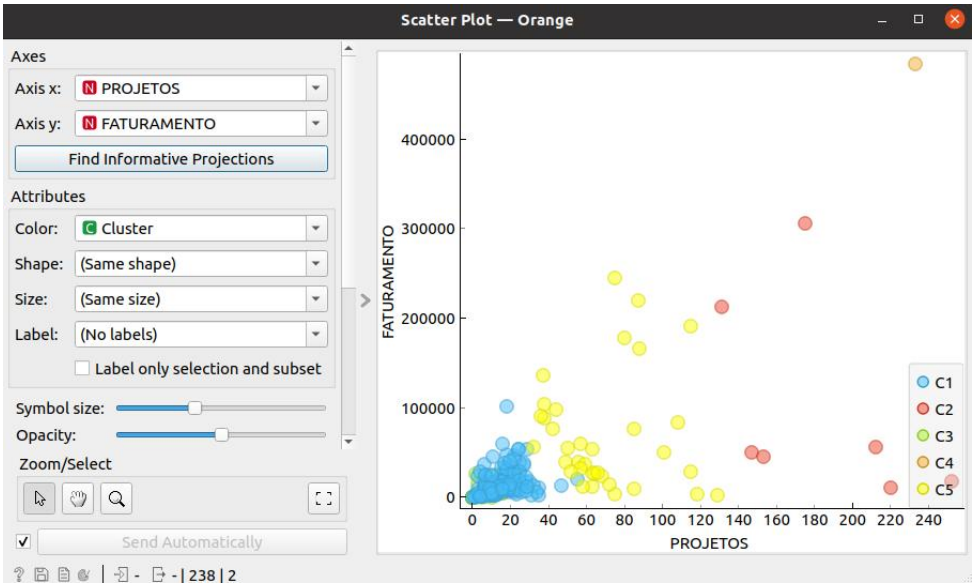
Figura 9 – *Scatter plot* mostrando a divisão de *cluster* atual




Fonte: Autora (2022).

A Figura 9 mostra no *scatter plot* a divisão dos *cluster* atuais feitos pela Brasil Júnior e corresponde à dados das EJs em 2020. É importante salientar que os *clusters* de 2020 foram calculados a partir de resultados feitos pelas EJs em 2019 e que o número de *cluster* não muda no decorrer do ano.

Figura 10 – *Scatter plot* mostrando a divisão de agrupamento proposto




Fonte: Autora (2022).

A Figura 10 mostra os novos agrupamentos propostos pelo algoritmo e mostra bastante similaridade pelos *cluster* C1 e C3, visto que o C3 é quase difícil de notar já que está encoberto pelos círculos azuis no começo do gráfico. Os agrupamentos propostos foram calculados por meio de dados e resultados feitos pelas EJs em 2020, por isso também podem haver divergências nas comparações.

Tendo essas informações, foi exportada a *Data Table* presente no *workflow* 2 que está conectada ao *scatter plot* e foram transferidos esses dados para uma planilha, para que as comparações fossem feitas com os dados encontrados nas bases de dados. Para as comparações divide-se em abas pelo *cluster* atual, e ao lado na tabela foi colocado qual o agrupamento proposto para àquela EJ e então contou-se o número de vezes que o agrupamento se repetia para entender quantitativamente os dados. De maneira qualitativa foi comparado o gráfico de *scatter plot* dos *clusters* atuais e agrupamentos propostos para entender a relação entre eles.

Um exemplo pode ser visto na Tabela 2, em que o *cluster* atual escolhido para a análise foi o *cluster* 4. Na análise dos gráficos entre o *cluster* atual e o agrupamento proposto pode-se ver que as EJs com *cluster* 4 atuais devem ser comparadas nos agrupamentos propostos às EJs agrupadas como C5 ou C2, o que de maneira geral aconteceu, visto que 70% das EJs ficaram agrupadas como C5 ou C2.

Tabela 2 – Comparação entre o *cluster* atual e o agrupamento proposto para EJs de *cluster* 4.

| ***Cluster* 4 atual** | **Agrupamento proposto** |
|---|---|
| Ideal Consultoria e Empreendimentos | C5 |
| MASCI Consultoria Jr. | C1 |
| Fator Júnior | C5 |
| Soluções Consultoria | C1 |
| EFICAP | C2 |
| PJ Consultoria | C5 |
| Esamb | C5 |
| Projet | C1 |
| EMAS Jr. Consultoria | C5 |
| Colucci Consultoria Jurídica Júnior | C5 |

Fonte: Autora (2022).

Na Tabela 2 é mostrado uma comparação entre o *cluster* atual e o agrupamento proposto para às EJs que, segundo a classificação da Brasil Júnior em 2020, eram consideradas de *cluster* 4.

É importante ressaltar que por meio do agrupamento feito pelo algoritmo *k-means* obteve-se alguns resultado que precisam ser levados em consideração nessas análises comparativas: dos agrupamentos propostos, quase 59% das EJs foram classificadas como C1, um número relativamente alto e que faz com que provavelmente hajam EJs C1 na comparação com todos os outros *clusters* atuais e apenas uma EJ foi classificada como C4, que deveria ser numa comparação direta *cluster* 5 atual. Há

também que cuidar na análise entre os *cluster* atuais 1 e 2 pois, como visto no gráfico de *scatter plot* dos *clusters* atuais, são EJs que ficam muito próximas entre si.

Para avaliar os demais resultados foi utilizada uma matriz de confusão para a interpretação. Dessa maneira é possível avaliar quais resultados foram agrupados da maneira correta e quais foram os resultados agrupados de maneira incorreta. A matriz de confusão correlaciona os resultados dos valores reais com os valores previstos, neste caso, agrupados e aponta métricas como acuracidade dos resultados, precisão e *recall* ou revocação.

Tabela 3 – Matriz de Confusão.

| | **Agrup 1** | **Agrup 3** | **Agrup 5** | **Agrup 2** | **Agrup 4** |
|---|---|---|---|---|---|
| *Cluster* 1 | 87 | 18 | 9 | 2 | 0 |
| *Cluster* 2 | 25 | 3 | 5 | 1 | 0 |
| *Cluster* 3 | 18 | 3 | 9 | 2 | 0 |
| *Cluster* 4 | 3 | 0 | 6 | 1 | 0 |
| *Cluster* 5 | 1 | 0 | 6 | 1 | 1 |

Fonte: Autora (2022).

Na Tabela 3 **Agrup n** significa agrupamento e ela mostra a matriz de confusão desenvolvida a partir dos *clusters* reais e dos agrupamentos obtidos pelo algoritmo e pode ser interpretada da seguinte maneira:

- Na diagonal principal há os resultado chamados de verdadeiros positivos, em que o que se quer prever ou classificar foi feito de maneira correta, ou seja, o *cluster* 1 foi classificado corretamente como agrupamento 3 em 18 EJs.
- Na linha horizontal fora da diagonal principal tem-se os falsos positivos, em que o *cluster* que buscamos classificar foi feito de maneira incorreta, ou seja, o *cluster* 1 teve 18 EJs agrupadas incorretamente no agrupamento 3, 9 EJs agrupadas incorretamente no agrupamento 5 e 2 EJs agrupadas incorretamente no agrupamento 2.
- Na linha vertical de cada agrupamento fora da diagonal principal tem-se os falsos negativos, quando os *clusters* atuais que não são os que estão sendo classificados no momento são previstos incorretamente, ou seja, o *cluster* 2 teve 25 EJs classificadas incorretamente como agrupamento 1, mesmo o *cluster* 2 não sendo o que busca-se prever para o agrupamento 1.

Nota-se que o maior número de erro vem dos falsos positivos, principalmente do *cluster* 1 com o agrupamento 1, isso se da pelo grande número bruto de EJs que foram agrupadas no agrupamento 1 e também pela proximidade de resultados entre o agrupamento 3 e o agrupamento 1.

Para ter uma validação mais qualitativa do processo de agrupamento das EJs, foi testado o algoritmo *k-means* na base de dados com empresas juniores da FEJESC e foi perguntado para empresários juniores o que eles achavam do resultado.

De maneira geral, houveram muitas EJs também agrupadas no agrupamento 1 e esse foi o o agrupamento que obteve mais divergências, sendo relatado que algumas EJs eram muito diferentes entre si, mas que os demais agrupamento faziam sentido qualitativamente, o que trás mais confiança para os resultados.

## 3.2 PREVISÃO DE FATURAMENTO

Após finalizada a etapa de agrupamento, passou-se a considerar a próxima etapa do trabalho que tem como objetivo aumentar a previsibilidade para os resultados das empresas juniores. A segunda etapa consiste em analisar dados de faturamento de anos anteriores das empresas juniores e prever qual será o seu faturamento no ano seguinte, o que pode dar mais assertividade para a EJ quando for traçar as suas metas anuais. Além disso são utilizados dados de número de projetos e faturamento para prever se aquela EJ em análise deve alcançar as suas metas no final do ano.

Alcançar resultados cada vez maiores de maneira constante é de extrema relevância para o movimento empresa júnior, e a meta de faturamento da rede é uma das mais desafiadoras e importantes já que é por meio dela que é medido o quanto uma empresa júnior consegue reinvestir na educação empreendedora no Brasil por meio de investimentos nos seus membros, em sua estrutura ou em diversos outros aspectos. Por isso, o faturamento da rede é medido desde 2009 quando foi criado o primeiro planejamento estratégico da rede e é uma métrica que cresce a cada ano.

Dessa maneira, conseguir compreender como essa métrica deve crescer ao longo dos anos é fundamental para garantir o crescimento sustentável do MEJ, olhando consequentemente para o crescimento em faturamento de cada uma das EJs da rede. Em um breve histórico, pode-se ver como o faturamento do MEJ cresceu nos últimos anos, podendo ser visto na Tabela 4.

Tabela 4 – Faturamento da rede nos últimos anos.

| **Faturamento 2019** | **Faturamento 2020** | **Faturamento 2021** | **Faturamento 2022 até agora** |
|---|---|---|---|
| R$44.809.297 | R$49.273.445,54 | R$69.480,922,39 | R$69.904.465,58 |

Fonte:(BRASIL JúNIOR, 2022).

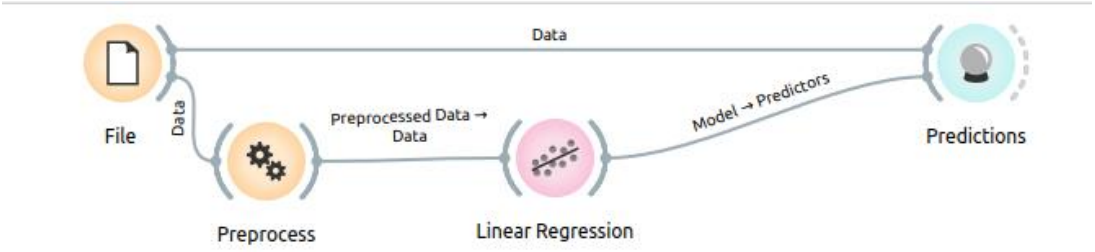
É possível ver a influência do momento de pandemia para os resultados de faturamento em 2020, no qual foi um ano de adaptações no modelo de trabalho de muitas EJs e também muitas vezes no modelo de negócio, em que algumas precisaram rever a sua oferta de serviços e qual seria o seu público alvo. Após isso, há novamente um crescimento considerável e há a meta de um crescimento ainda maior para o triênio.

Por isso, foi desenvolvido o *workflow* 3 para prever o faturamento das empresas juniores utilizando como base o faturamento dos últimos 4 anos, o *cluster* e o número de projetos vendidos nos últimos anos, esses dados se correlacionam pois teoricamente

quanto mais projetos uma EJ vendeu mais ela faturou e consequentemente o seu *cluster* deve ser maior. Dessa maneira, foi utilizado um algoritmo de regressão linear para fazer a previsão e o *workflow* 3 montado pode ser visto na Figura 11.

É importante ressaltar que parte importante deste trabalho também é facilitar o uso de técnicas de *data science* para gestores de EJs. Dessa maneira, mesmo tendo poucos dados disponíveis para fazer essa análise e mesmo com resultados que possam vir a não serem tão bons, é importante que os *workflows* estejam intuitivos, assim como a parte de análise dos resultados na Orange.

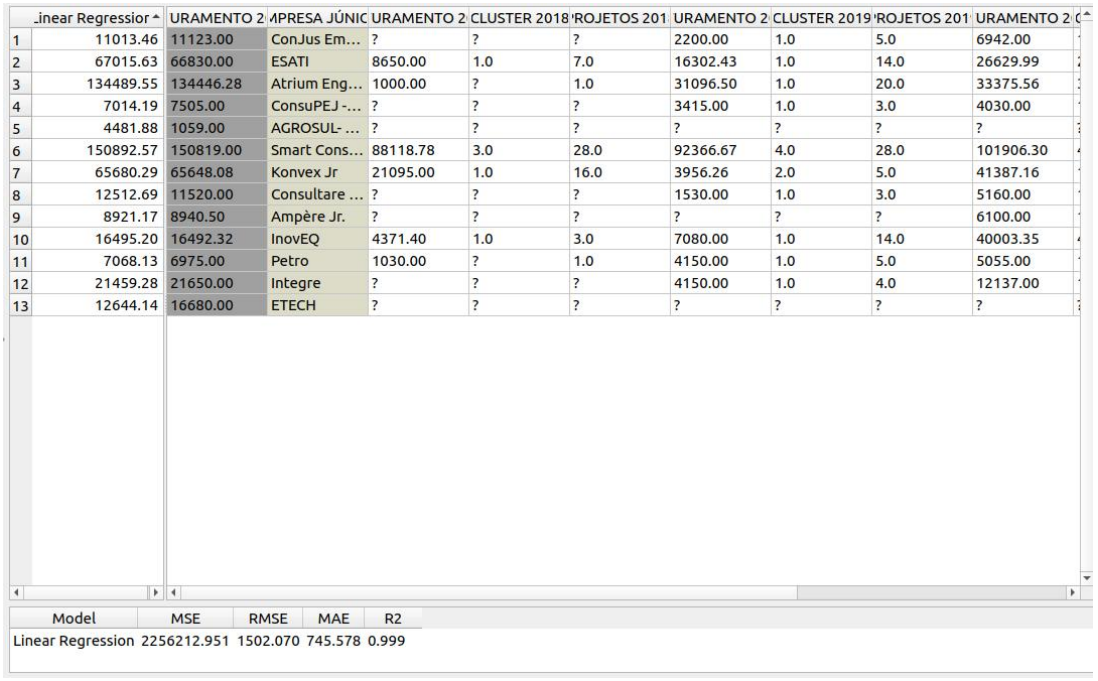
Figura 11 – *Workflow* 3 para previsão do faturamento.




Fonte: Autora (2022).

Na Figura 11 se apresenta o *workflow* 3 que consiste em um *widget* de documento em que é colocada a base de dados, e ele é ligado ao pré-processamento que irá adicionar valores aos campos que estão sem informações, geralmente em EJs que ainda não eram federadas naquele ano. Logo após conectado ao *widget* de regressão linear, que aplica o algortimo na base de dados pode-se ver o resultado da regressão linear, que prevê o faturamento individual de cada EJ e também às métricas de erro quadrático médio, o desvio médio quadrático, erro absoluto médio e o $R^2$. A Figura 12 mostra os resultados obtidos pela regressão linear.

Figura 12 – Resultado da regressão linear para previsão do faturamento.



| | inear Regressior | URAMENTO 2 | MPRESA JÚNIC | URAMENTO 2 | CLUSTER 2018 | ROJETOS 201 | URAMENTO 2 | CLUSTER 2019 | ROJETOS 201 | URAMENTO 2 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 11013.46 | 11123.00 | ConJus Em… | ? | ? | ? | 2200.00 | 1.0 | 5.0 | 6942.00 |
| 2 | 67015.63 | 66830.00 | ESATI | 8650.00 | 1.0 | 7.0 | 16302.43 | 1.0 | 14.0 | 26629.99 |
| 3 | 134489.55 | 134446.28 | Atrium Eng… | 1000.00 | ? | 1.0 | 31096.50 | 1.0 | 20.0 | 33375.56 |
| 4 | 7014.19 | 7505.00 | ConsuPEJ -… | ? | ? | ? | 3415.00 | 1.0 | 3.0 | 4030.00 |
| 5 | 4481.88 | 1059.00 | AGROSUL- … | ? | ? | ? | ? | ? | ? | ? |
| 6 | 150892.57 | 150819.00 | Smart Cons… | 88118.78 | 3.0 | 28.0 | 92366.67 | 4.0 | 28.0 | 101906.30 |
| 7 | 65680.29 | 65648.08 | Konvex Jr | 21095.00 | 1.0 | 16.0 | 3956.26 | 2.0 | 5.0 | 41387.16 |
| 8 | 12512.69 | 11520.00 | Consultare … | ? | ? | ? | 1530.00 | 1.0 | 3.0 | 5160.00 |
| 9 | 8921.17 | 8940.50 | Ampère Jr. | ? | ? | ? | ? | ? | ? | 6100.00 |
| 10 | 16495.20 | 16492.32 | InovEQ | 4371.40 | 1.0 | 3.0 | 7080.00 | 1.0 | 14.0 | 40003.35 |
| 11 | 7068.13 | 6975.00 | Petro | 1030.00 | ? | 1.0 | 4150.00 | 1.0 | 5.0 | 5055.00 |
| 12 | 21459.28 | 21650.00 | Integre | ? | ? | ? | 4150.00 | 1.0 | 4.0 | 12137.00 |
| 13 | 12644.14 | 16680.00 | ETECH | ? | ? | ? | ? | ? | ? | ? |

| Model | MSE | RMSE | MAE | R2 |
|---|---|---|---|---|
| Linear Regression | 2256212.951 | 1502.070 | 745.578 | 0.999 |

Fonte: Autora (2022).

De modo geral, é possível concluir que os resultados são satisfatórios pois grande parte dos faturamentos previstos ficou em uma margem de erro de apenas 10% como pode ser visto na comparação na Tabela 5, sendo que o erro foi calculado diminuindo o valor aproximado do exato e dividindo pelo valor exato. O maior erro pode ser visto na previsão de faturamento para a EJ Agrosul, o que pode ser dado pelo faturamento de 2021 ser baixo comparado com o das outras EJs o que pode ter sido afetado pelo pré-processamento escolhido e pela falta de dados de anos anteriores visto que a EJ foi fundada em 2020 e não tem dados de anos anteriores de faturamento.

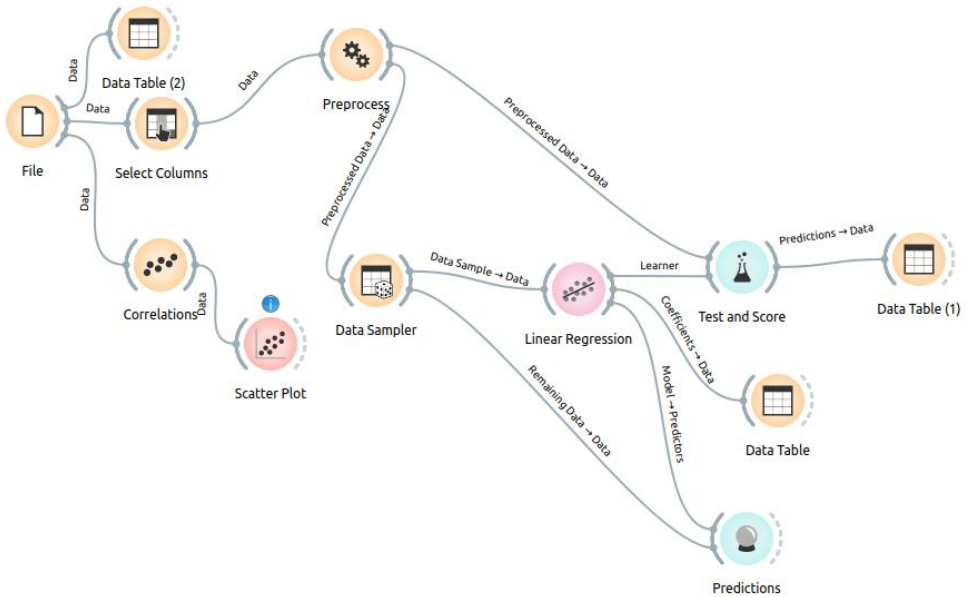
Tabela 5 – Comparação entre o faturamento previsto e feito.

| **Empresa júnior** | **Faturamento previsto (R$)** | **Faturamento 2021 (R$)** | **Erro** |
|---|---|---|---|
| ConJus Empresa Júnior | 11.013,455 | 11.123,00 | 0,9848% |
| ESATI | 67.015,63 | 66.830,00 | 0,2778% |
| Atrium Engenharia Júnior | 134.489,55 | 134.446,28 | 0,0322% |
| ConsuPEJ - Consultoria Agropecuária | 7.014,19 | 7.505,00 | 6,5397% |
| AGROSUL- Soluções Agronômicas | 4.481,88 | 1.059,00 | 323,2180% |
| Smart Consultoria Júnior | 150.892,57 | 150.819,00 | 0,04885% |
| Konvex Jr | 65.680,29 | 65.648,08 | 0,0491% |
| Consultare Empresa Júnior Jurídica | 12.512,69 | 11.520,00 | 8,6171% |
| Ampère Jr. | 8.921,17 | 8.940,50 | 0,2162% |
| InovEQ | 16.495,20 | 16.492,32 | 0,0175% |
| Petro | 7.068,13 | 6.975,00 | 1,3352% |
| Integre | 21.459,28 | 21.650,00 | 0,8809% |
| ETECH | 12.644,14 | 16.680,00 | 24,1958% |

Fonte: Autora (2022).

A partir desses resultados obtidos foram montados *workflows* adicionando uma nova base de dados com EJs diferentes para que uma base de dados servisse como teste e outra servisse como treino. Além disso, foi feito o *workflow* 4 usando a mesma base de dados e o *widget data sampler* que automaticamente divide dados para treino e para teste, a partir do que o usuário escolher como visto na Figura 13.

Figura 13 – *Workflow* 4 para previsão de faturamento.




Fonte: Autora (2022).

A Figura 13 mostra o *Workflow* 4 montado para prever o faturamento EJs.

Entra-se com a base de dados e então a partir dela consegue-se tirar as correlações e selecionar as colunas que serão utilizadas para a previsão. Depois os dados passam por um pré-processamento em que são usados os valores médios ou retirados as linhas com valores faltantes e então os dados são separados em dados de treino e teste. Em que os dados de treino são usados para treinar o modelo da regressão linear e os restantes usados para a previsão, sendo os dados mostrados em tabelas.

A partir do *workflow* 4 foram obtidos alguns resultados. Quando a base de dados continha EJs muito diferentes entre si na comparação de faturamento a regressão linear não conseguia trazer resultados coerentes com os resultados esperados. Por exemplo, ao selecionar todas as EJs de Santa Catarina numa base de dados de em torno de 70 EJs muito diferentes entre si, com faturamentos variando de 1.000,00 reais até mais de 1.000.000,00 de reais, a regressão linear obteve resultado até negativos para a predição, o que não pode ser levado em consideração.

A partir destes testes, foi entendido que as EJs precisariam ser separadas em base de dados diferentes para fazer uma previsão mais coerente com a regressão linear no Orange. Por isso, foi feita uma separação por região. No caso de Santa Catarina foram separadas as EJs na região Norte e Vale do Itajaí, região da Grande Florianópolis e regiões de Sul, Serra e Oeste, desta maneira obtiveram-se resultados mais coerentes.

Além da separação entre regiões foi utilizado do *widget correlation* para somente usar os dados que tinham uma forte correlação entre si e dessa maneira obter resultados mais precisos. As correlações entre os dados da região da Grande Florianópolis pode ser vista também na Figura 14. Além disso é possível plotar um gráfico com essas correlações por meio do *scatter plot* e com a reta de regressão é possível ver o *score* de cada par de variáveis.

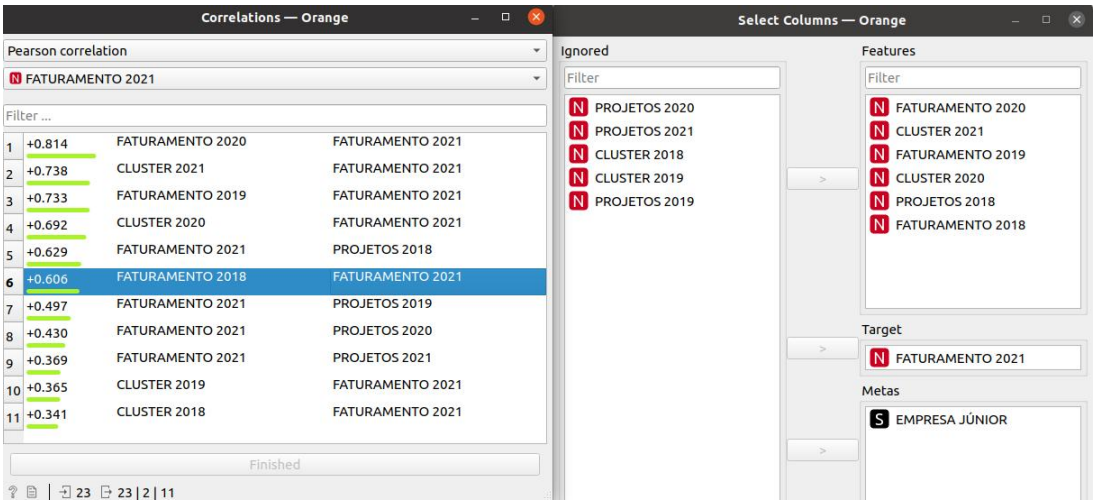
Como cada região apresenta dados diferentes e comportamentos diferentes entre si, as correlações entre os pares consequentemente também se mostra diferente, por isso é utilizado do *widget Select Columns* para que seja possível escolher só as colunas com a correlação forte com o dado objetivo (Faturamento 2021) neste exemplo. Dessa maneira, para cada região diferentes variáveis são consideradas para fazer a previsão do faturamento.

Ainda na região da Grande Florianópolis por conter uma grande diferença entre as EJs precisou-se dividir as EJs em dois grupos diferentes e propôs-se uma nova maneira de fazer a diferenciação entre as EJs. Foram retiradas da grande base de dados as EJs de *cluster* 5, do agrupamento da Brasil Júnior, visto que geralmente são as empresas com números de faturamento que podem ser considerados *outliers* com relação aos outros *Clusters* e dessa maneira obtiveram-se os resultados a seguir.

As Figuras 14, 15 e 16 mostram como foi seguido o processo para obter os resultados de previsão das EJs da região da Grande Florianópolis. As demais regiões seguiram o mesmo método, podendo divergir na quantidade de dados para teste ou

treino ou mesmo no método de divisão dos dados, que pode ser por validação cruzada ou divisão aleatória dos dados para teste e treino, assim como mudam as colunas selecionadas pela correlação.

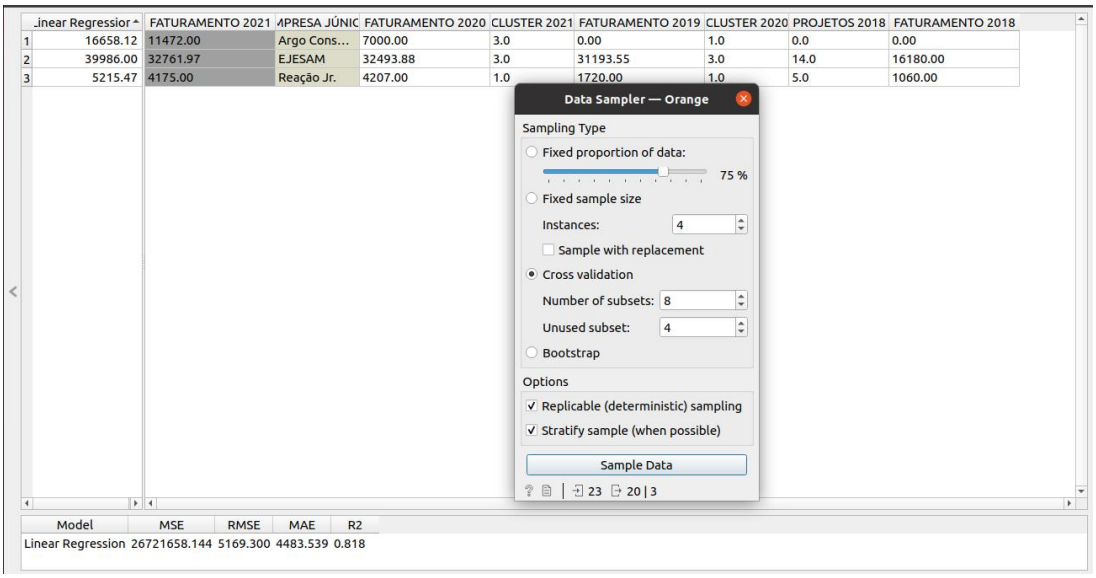
Figura 14 – Correlação entre os dados da região da Grande Florianópolis.




Fonte: Autora (2022).

A Figura 14 mostra a correlação entre as variáveis da base de dados com a variável que é a meta a ser comparada "faturamento 2021", também é mostrado que as colunas escolhidas para serem utilizadas para o modelo são aquelas que apresentam correlação forte, índice maior que 0,5.

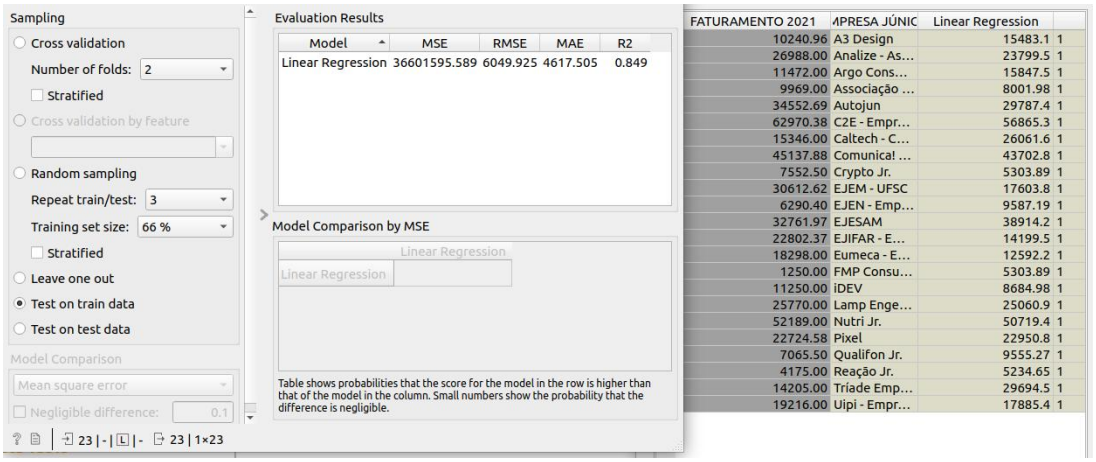
Figura 15 – Resultado da regressão linear para previsão do faturamento com validação cruzada.




Fonte: Autora (2022).

A Figura 15 mostra o *widget* de previsões e o *data sampler*. Pode-se ver que foi escolhida a validação cruzada, em que 8 subgrupos foram escolhidos e 4 não são usados. Dessa maneira, obtiveram-se os resultados de previsão que podem ser vistos na Figura 16. Os níveis de assertividade em geral acabam caindo para essa base de dados, assim como o erro acaba subindo, mas ainda podem ser valores que podem ajudar a basear uma decisão para as EJs.

Figura 16 – Resultado da regressão linear para previsão do faturamento usando a base de treino.




Fonte: Autora (2022).

Na Figura 16 vê-se o *widget text & score* e os seus resultados na tabela, aqui são utilizados os dados pré-processados com o modelo de regressão linear, e o teste é feito com os dados de treino, dessa maneira obtêm-se as previsões de faturamento vistas na coluna "*Linear Regression*". Ainda nota-se que quanto maior o número de dados obtidos, geralmente mais próximo é o valor previso do valor real.

A partir desse método obtiveram-se os resultados para outras duas regiões, a região Norte e Vale do Itajaí em Santa Catarina e os outros resutlados incluem as regiões Sul, Serra e Oeste em conjunto, visto que se fossem divididas essas regiões não haveriam dados o suficiente para fazer a análise separadamente. Ainda podem-se agrupar as regiões pois não há uma diferença tão grande entre essas EJs olhando para os seus dados brutos coletados.

Para a Região Norte e Vale do Itajaí, a partir das correlações obtidas foram retiradas as colunas de "Projetos 2020"e "Projetos 2021"e foi utilizado o pré-processamento adicionando valores médios ou mais frequentes aos dados faltantes. Para o *widget predictions* foi utilizado o *data sampler* com a validação cruzada, e para o *text & score* foi utilizado o teste com os dados de treino, assim como mostrado nos exemplos com os dados da região da Grande Florianópolis. A partir disso foram obtidos os resultados apresentados nas Tabelas 6 e 7.

Tabela 6 – Comparação entre o faturamento previsto e feito a partir do *widget predictions*.

| **Empresa júnior** | **Faturamento 2021 (R$)** | **Faturamento previsto (R$)** | **Erro** |
|---|---|---|---|
| ESATI | 66.830,00 | 51.918,86 | 22,3120% |
| AGROSUL- Soluções Agronômicas | 4.481,88 | 16.739,36 | 323,2180% |
| ETECH | 16.680,00 | 16.739,36 | 0,3559% |

Fonte: Autora (2022).

A Tabela 6 mostra os resultados obtidos depois dos dados serem divididos pela validação cruzada e serem usados apenas 3 EJs para os testes. Obtiveram-se resultados não tão bons, podendo ser devido ao método de divisão dos dados e também da quantidade de dados não ser tão grande, ainda mais quando são ignoradas algumas colunas.

A Tabela 7 mostra os dados vindos da tabela conectada ao *widget Text & Score* e trás melhores resultados, podendo ser considerados satisfatórios visto que a maioria dos erros está abaixo ou muito próximo de 10%. As divergências podem estar acontecendo também pelo baixo número de dados presentes e também pela maneira como os dados são separados como teste e treino, além do pré-processamento.

Da mesma maneira, foi seguido o mesmo método para obter os resultados apresentados para a região Sul, Serra e Oeste, é importante ressaltar que grande parte

Tabela 7 – Comparação entre o faturamento previsto e feito a partir do *widget Text & Score*.

| **Empresa júnior** | **Faturamento previsto (R$)** | **Faturamento 2021(R$)** | **Erro** |
|---|---|---|---|
| ConJus Empresa Júnior | 9.727,20 | 11.123,00 | 12,54% |
| ESATI | 65.480,07 | 66.830,00 | 2,01% |
| Atrium Engenharia Júnior | 134.886,20 | 134.446,28 | 0,3272% |
| ConsuPEJ - Consultoria Agropecuária | 8.158,27 | 7.505,00 | 8,7045% |
| AGROSUL- Soluções Agronômicas | 8.753,01 | 1.059,00 | 726,53% |
| Smart Consultoria Júnior | 150.578,70 | 150.819,00 | 0,1593% |
| Konvex Jr | 65.816,57 | 65.648,08 | 0,2566% |
| Consultare Empresa Júnior Jurídica | 9.625,73 | 11.520,00 | 16,4432% |
| Ampère Jr. | 10.554,46 | 8.940,50 | 18,0523% |
| InovEQ | 16.952,85 | 16.492,32 | 2,7923% |
| Petro | 6.159,58 | 6.975,00 | 11,6905% |
| Integre | 24.2424,48 | 21.650,00 | 11,9745% |
| ETECH | 8.753,01 | 16.680,00 | 47,5239% |

Fonte: Autora (2022).

das EJs dessas regiões são EJs novas, então é ainda mais perceptível a diferença entre os resultados pois não há um grande número de dados e os dados não são totalmente preenchidos, o que pode piorar o resultado do modelo.

Para essa região, segundo as correlações foram retiradas às colunas de "Projetos 2020", "Projetos 2021"e "*cluster* 2020"e o pré-processamento foi configurado para adicionar valores faltantes com a média ou os valores mais frequentes daquela base de dados. A separação dos dados é feita por validação cruzada assim como para as outras regiões para o *widget predictions* e para o *widget Text & Score* é usado o teste nos dados de treino. A partir disso foram obtidos os resultados presentes na Tabela 8.

Tabela 8 – Comparação entre o faturamento previsto e feito a partir do *widget predictions*.

| **Empresa júnior** | **Faturamento 2021 (R$)** | **Faturamento previsto (R$)** | **Erro** |
|---|---|---|---|
| MAJ Mecatrônica e Automação Júnior | 12.770,00 | 38.387,60 | 200,6076% |
| Safra Júnior | 3.000,00 | 4.330,40 | 44,3469% |
| Sem Fronteiras Consultoria Jr | 10.444,00 | 9.947,32 | 4,7555% |

Fonte: Autora (2022).

A Tabela 8 mostra os resultados obtidos depois dos dados serem divididos pela validação cruzada e serem usados apenas 3 EJs para os testes. Dessa maneira obtiveram-se resultados não tão bons, podendo ser devido ao método de divisão dos dados e também da quantidade de dados não ser tão grande ainda mais quando são ignoradas algumas colunas.

É possível ver pela comparação do erro que esses foram os piores resultados na

Tabela 9 – Comparação entre o faturamento previsto e feito a partir do *widget Text & Score*.

| **Empresa júnior** | **Faturamento 2021 (R$)** | **Faturamento previsto (R$)** | **Erro** |
|---|---|---|---|
| Alquimia Jr | 9.200,00 | 11.637,56 | 26,4953% |
| CAV Florestal Empresa Júnior | 5.250,00 | 13.667,65 | 160,3362% |
| Consultali | 3.610,00 | 6.487,32 | 79,7042% |
| ContrAut Jr | 13.220,5 | 11.476,83 | 13,1891% |
| ECO JR | 5.638,50 | 9.623,50 | 70,6749% |
| EJEC | 23.500,00 | 27.862,60 | 18,5642% |
| ENEjr | 108.000,00 | 106.240,06 | 1,6295% |
| Galo Jr. Comunicação | 31.109,28 | 35.815,15 | 15,1269% |
| iDealize Júnior | 31.087,5 | 25.508,89 | 17,9448% |
| iModa Jr | 10.205,03 | 1.168,07 | 88,5539% |
| Kadima | 2.287,96 | 1.423,39 | 37,7877% |
| MAJ - Mecatrônica e Automação Júnior | 12.770,00 | 16.064,40 | 25,7979% |
| PRO Consultoria Júnior | 34.275,00 | 19.262,23 | 43,8009% |
| PROJETA Ambiental Jr | 35.400,00 | 37.421,44 | 5,7102% |
| ProjeteJr | 1.150,00 | 7.953,51 | 591,61% |
| Safra Júnior | 3.000,00 | 4.221,50 | 40,7167% |
| Sem Fronteiras Consultoria Jr | 10.444,00 | 8.618,81 | 17,4759% |
| Zootec Jr. | 7.179,73 | 2.874,52 | 59,9633% |

Fonte: Autora (2022).

comparação entre as 3 regiões. Isso pode ser dado pelo fato que a maioria dessas EJs são novas então tem um histórico de dados tão grande, o que prejudica a aprendizagem pelo modelo, isso pode ser notado nos melhores resultados que foram obtidos pela ENEJr, uma EJ mais antiga e de maior maturidade, assim como a PROJETA Ambiental Jr, o que pode trazer mais confiabilidade à hipótese levantada.

Dessa maneira, foram feitas outras comparações utilizando da mesma lógica, foram pegas diferentes EJs e separadas em duas bases de dados diferentes, EJs de *Cluster* 1, 2, 3 e 4 e outra base de dados apenas com EJs de *Cluster* 5. Outro ponto que pode colaborar para bom resultados para a previsão com EJs de *Cluster* 5 é porque geralmente essas são empresas juniores mais velhas e de mais maturidade o que faz com que consequentemente elas tenham um maior volume de dados brutos, a grande maioria tem os dados de faturamento desde 2018, o que ajuda a ter melhores resultados.

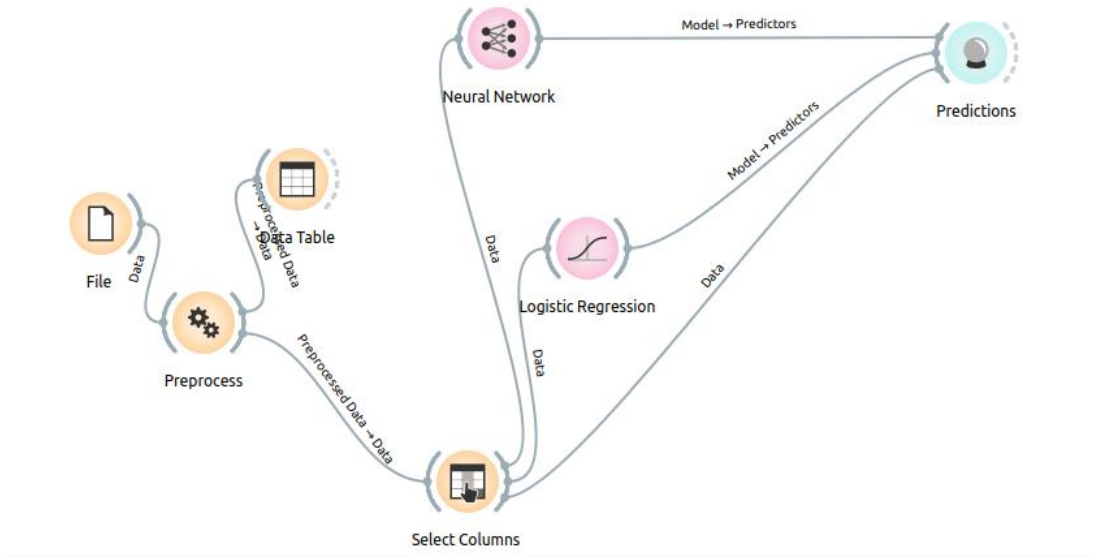
## 3.3 PREVISÃO DO ALCANCE DAS METAS

Por mais que a meta de faturamento seja importante, é ainda mais importante que ela seja atingida ao final do ano pelas EJs. Por isso, dentro do MEJ existe o conceito de "Alto Crescimento"ou AC. Uma EJ AC, até o planejamento estratégico da rede 2019-2021 era definida como uma EJ que alcançasse as suas metas de faturamento, número de projeto e porcentagem de membros que executam projetos. Assim, acompanhar se essas metas estão sendo atingidas e se tem o potencial de serem atingidas são de

muita relevância.

Dessa maneira, foi montado outro *wokflow* para tentar prever se as EJs serão alto crescimento. Para isso, também foram utilizados dados dos últimos 4 anos e foram utilizadas as seguintes métricas: faturamento, *cluster* do ano vigente, número de projetos vendidos e se a EJ foi AC durante aquele ano. Para isso, foi feita a comparação com dois algoritmos diferentes para entender qual traria o melhor resultado e o *workflow* 5 proposto pode ser visto na Figura 17.

Figura 17 – *Workflow* 5 para previsão se a EJ será AC.




Fonte: Autora (2022).

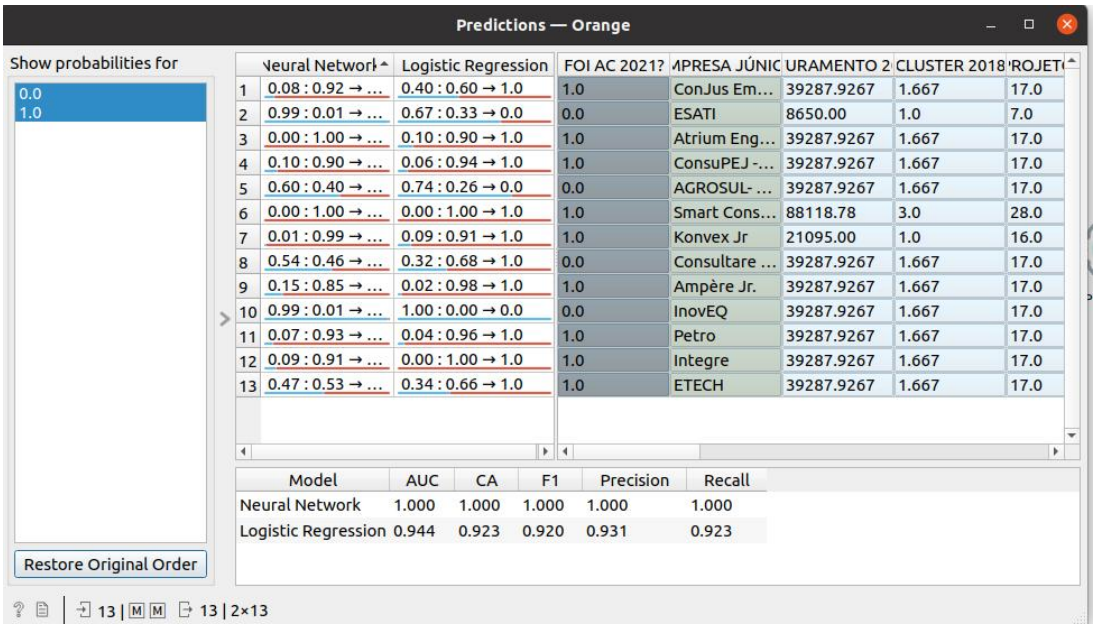
Na Figura 17 pode-se ver o *worflow* 5 proposto para a previsão se a EJ será AC. O *workflow* 5 inicia com o *widget* da base de dados que então vai passar por um pré-processamento que está configurado para adicionar os valores faltantes com valores médios a partir dos outros valores da base. Em seguida são escolhidas as colunas usadas para o treino e teste e qual a coluna meta para a previsão. Os dados são utilizado em uma Rede Neural Artificial (RNA) e na Regressão Logística, após isso pode-se ver os resultados dos algoritmos no *widget* de previsão.

Para obter os resultados é necessário configurar a RNA de maneira adequada. A representação escolhida para o *workflow* 5 foi de 3 camadas e os neurônios foram espalhados de maneira que na primeira camadas houvessem 10 neurônios e na segunda e na terceira camada houvessem 8 neurônios em cada. O método de ativação da rede é a identidade, o resolvedor foi escolhido o Adam e o número máximo de épocas de treinamento de 200 iterações.

Os resultados e a comparação entre os dois algoritmos RNA e Regressão

Logística podem ser vistos na Figura 18.

Figura 18 – Resultado da previsão se a EJ será AC.

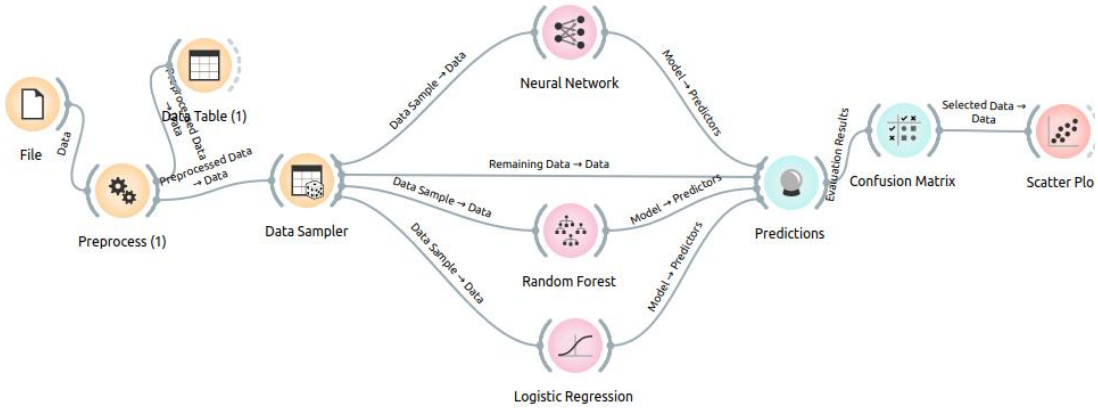



Fonte: Autora (2022).

Na Figura 18 é possível identificar que ambos os algoritmos tiveram resultados satisfatórios pois conseguiram acertar mais de 90% dos resultados, entretanto o algoritmo que obteve o melhor desempenho foi a RNA que teve um resultado perfeito, conseguindo acertar todas as previsões. A Regressão Logística também obteve um bom resultado e pode ser utilizada como base para a avaliação. O que pode ter levado ao erro foi o baixo número de dados e alguns dados, principalmente dos anos de 2018 e 2019 com vários valores faltantes, trazendo grande dependência ao pré-processamento.

O *workflow* 5 pode ser considerado um ensaio para o *workflow* 6 que foi melhorado e incrementado. Por isso, o *workflow* 5 realizou uma previsão para uma base menor de dados, já para o *workflow* 6 foi usada a base de dados de EJs de SC. Por serem dados categóricos e são utilizados algoritmos mais potentes, não há o mesmo problema que foram encontrados na previsão do faturamento, dessa maneira é válida a entrada de dados de maior número de EJs. Dessa maneira, foi adicionada uma nova base de dados assim como novos algoritmos e a matriz de confusão para avaliar os resultados, o novo *workflow* 6 pode ser visto na Figura 19.

Figura 19 – *Workflow* 6 para prever se a EJ será AC.

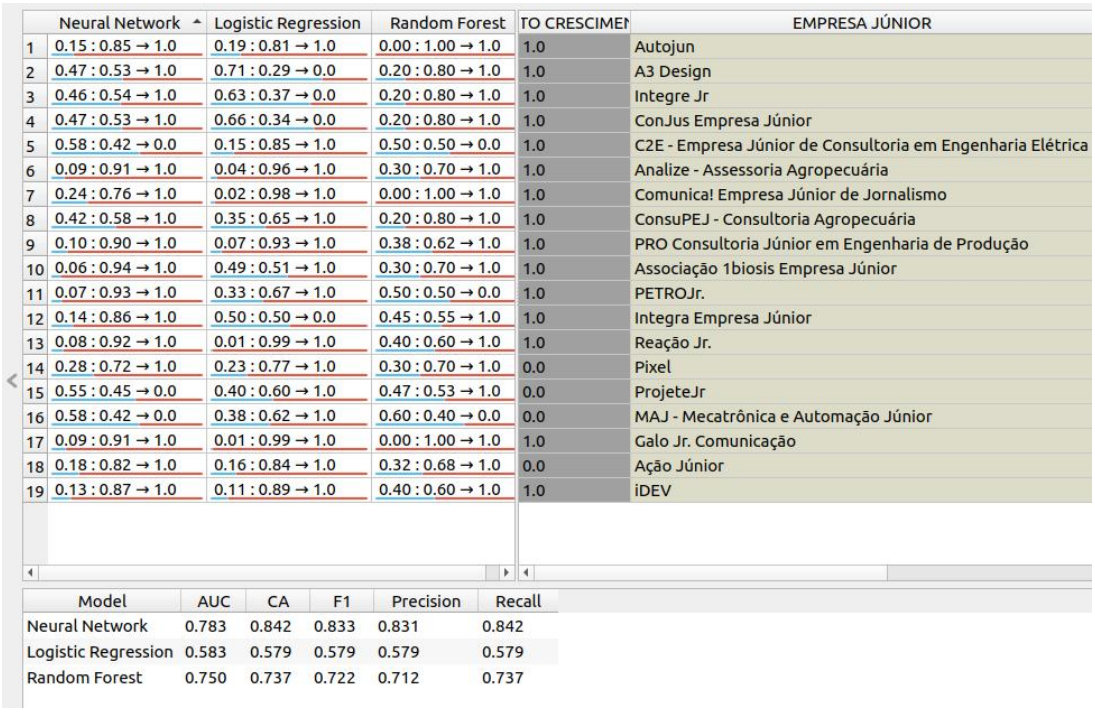



Fonte: Autora (2022).

Na Figura 19 vê-se o *workflow* 6, nele o *widget Data Sampler* foi adicionado, dessa maneira os algoritmos são treinados com 70% da base de dados e os outros 30% são mandados para o *widget Predictions* para as previsões de fato. O algoritmo de Floresta Aleatória (*Random Forest*) foi adicionado e os resultados podem ser avaliados pela matriz de confusão e os valores podem ser vistos no *Scatter plot*.

Com isso, foram obtidos diferentes resultados para cada um dos algoritmos sendo que o melhor desempenho foi da RNA, que foi configurada com 3 camadas com 10, 8 e 8 neurônios em cada uma respectivamente e sua ativação é feita pelo método ReLu. A Floresta Aleatória está configurada para um número de 10 árvores. Os resultados das previsões podem ser vistos na Figura 20 assim como os a comparação entre os 3 algoritmos.

Figura 20 – Novo *workflow* para prever se a EJ será AC.



| | Neural Network | Logistic Regression | Random Forest | TO CRESCIMEN | EMPRESA JÚNIOR |
|---|---|---|---|---|---|
| 1 | 0.15 : 0.85 → 1.0 | 0.19 : 0.81 → 1.0 | 0.00 : 1.00 → 1.0 | 1.0 | Autojun |
| 2 | 0.47 : 0.53 → 1.0 | 0.71 : 0.29 → 0.0 | 0.20 : 0.80 → 1.0 | 1.0 | A3 Design |
| 3 | 0.46 : 0.54 → 1.0 | 0.63 : 0.37 → 0.0 | 0.20 : 0.80 → 1.0 | 1.0 | Integre Jr |
| 4 | 0.47 : 0.53 → 1.0 | 0.66 : 0.34 → 0.0 | 0.20 : 0.80 → 1.0 | 1.0 | ConJus Empresa Júnior |
| 5 | 0.58 : 0.42 → 0.0 | 0.15 : 0.85 → 1.0 | 0.50 : 0.50 → 0.0 | 1.0 | C2E - Empresa Júnior de Consultoria em Engenharia Elétrica |
| 6 | 0.09 : 0.91 → 1.0 | 0.04 : 0.96 → 1.0 | 0.30 : 0.70 → 1.0 | 1.0 | Analize - Assessoria Agropecuária |
| 7 | 0.24 : 0.76 → 1.0 | 0.02 : 0.98 → 1.0 | 0.00 : 1.00 → 1.0 | 1.0 | Comunica! Empresa Júnior de Jornalismo |
| 8 | 0.42 : 0.58 → 1.0 | 0.35 : 0.65 → 1.0 | 0.20 : 0.80 → 1.0 | 1.0 | ConsuPEJ - Consultoria Agropecuária |
| 9 | 0.10 : 0.90 → 1.0 | 0.07 : 0.93 → 1.0 | 0.38 : 0.62 → 1.0 | 1.0 | PRO Consultoria Júnior em Engenharia de Produção |
| 10 | 0.06 : 0.94 → 1.0 | 0.49 : 0.51 → 1.0 | 0.30 : 0.70 → 1.0 | 1.0 | Associação 1biosis Empresa Júnior |
| 11 | 0.07 : 0.93 → 1.0 | 0.33 : 0.67 → 1.0 | 0.50 : 0.50 → 0.0 | 1.0 | PETROJr. |
| 12 | 0.14 : 0.86 → 1.0 | 0.50 : 0.50 → 0.0 | 0.45 : 0.55 → 1.0 | 1.0 | Integra Empresa Júnior |
| 13 | 0.08 : 0.92 → 1.0 | 0.01 : 0.99 → 1.0 | 0.40 : 0.60 → 1.0 | 1.0 | Reação Jr. |
| 14 | 0.28 : 0.72 → 1.0 | 0.23 : 0.77 → 1.0 | 0.30 : 0.70 → 1.0 | 0.0 | Pixel |
| 15 | 0.55 : 0.45 → 0.0 | 0.40 : 0.60 → 1.0 | 0.47 : 0.53 → 1.0 | 0.0 | ProjeteJr |
| 16 | 0.58 : 0.42 → 0.0 | 0.38 : 0.62 → 1.0 | 0.60 : 0.40 → 0.0 | 0.0 | MAJ - Mecatrônica e Automação Júnior |
| 17 | 0.09 : 0.91 → 1.0 | 0.01 : 0.99 → 1.0 | 0.00 : 1.00 → 1.0 | 1.0 | Galo Jr. Comunicação |
| 18 | 0.18 : 0.82 → 1.0 | 0.16 : 0.84 → 1.0 | 0.32 : 0.68 → 1.0 | 0.0 | Ação Júnior |
| 19 | 0.13 : 0.87 → 1.0 | 0.11 : 0.89 → 1.0 | 0.40 : 0.60 → 1.0 | 1.0 | iDEV |

| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Neural Network | 0.783 | 0.842 | 0.833 | 0.831 | 0.842 |
| Logistic Regression | 0.583 | 0.579 | 0.579 | 0.579 | 0.579 |
| Random Forest | 0.750 | 0.737 | 0.722 | 0.712 | 0.737 |

Fonte: Autora (2022).

Na Figura 20 podem ser vistos os 30% dos dados que estão sendo usados para previsão, em que podem ser vistas as proporções que os algoritmos usaram para fazer a previsão, e qual foi o valor da previsão em si, sendo os valores na coluna em cinza são os valores reais. É possível identificar que a RNA obteve os melhores resultados, seguido pela floresta e por fim a Regressão Logística, ao avaliar os resultados das métricas mostrados no final da Figura. As matrizes de confusão para cada um dos algoritmos ajuda a avaliar os resultados e serão mostradas nas Tabelas 10, 11 e 12.

Para o arranjo dos dados na tabela foram trocados os valores "Sim"e "Não"para valores binários 1 e 0, de maneira que fossem possíveis analisar os resultados. Para a facilidade de interpretação nas matrizes de confusão serão retomados o "Sim"e "Não".

Tabela 10 – Matriz de Confusão para a RNA.

| | **Não** | **Sim** | **Somatório** |
|---|---|---|---|
| Não | 2 | 2 | 4 |
| Sim | 1 | 14 | 15 |
| Somatório | 3 | 16 | 19 |

Fonte: Autora (2022).

Na Tabela 10 vê-se a matriz de confusão para a RNA, em que na diagonal principal encontra-se os valores que foram preditos corretamente, sejam eles 0 ou 1, na linha 1 coluna 2, têm-se o valores de falso positivo em que o valor predito foi 1, ou

seja, a RNA disse que àquela EJ seria AC, mas na verdade ela não foi e na linha 1 coluna 2, têm-se os falsos negativos, em que àquelas EJs na realidade não foram AC e o algoritmo disse que elas foram. Num geral obtiveram-se bons resultados com a RNA, com um acerto de 16 EJs, o que resulta num acurácia de 84,2%.

Tabela 11 – Matriz de Confusão para a Regressão Logística.

| | **Não** | **Sim** | **Somatório** |
|---|---|---|---|
| Não | 0 | 4 | 4 |
| Sim | 4 | 11 | 15 |
| Somatório | 4 | 15 | 19 |

Fonte: Autora (2022).

Na Tabela 11 vê-se a matriz de confusão para a Regressão Logística, em que na diagonal principal encontra-se os valores que foram preditos corretamente, sejam eles 0 ou 1 totalizando 11 EJs, com somente verdadeiros positivos, ou seja, o algoritmo não acertou nenhuma EJ que não foi AC. Num geral e comparando com os outros algoritmo não obtiveram-se bons resultados para a Regressão Logística, tendo a menor taxa de acertos, com uma acurácia de 57,9%.

Tabela 12 – Matriz de Confusão para a Floresta Aleatória.

| | **Não** | **Sim** | **Somatório** |
|---|---|---|---|
| Não | 1 | 3 | 4 |
| Sim | 2 | 13 | 15 |
| Somatório | 3 | 16 | 19 |

Fonte: Autora (2022).

Na Tabela 12 vê-se a matriz de confusão para a Floresta Aleatória, em que na diagonal principal encontra-se os valores que foram preditos corretamente, sejam eles 0 ou 1 totalizando 14 EJs, além disso foram 3 falsos positivos e 2 falsos negativos. Num geral e comparando com os outros algoritmo os resultados são bons, visto que há uma taxa de acertos alta, acurácia de 73,7% e mais parecida com a RNA.

Os 3 algoritmos poderiam ainda ter sido configurados de maneiras diferentes para testar o alcance de melhores resultados, porém, como o trabalho também visa a facilidade do entendimento dos *workflows* e dos resultados por gestores que podem vir a não ter o conheicimento técnico sobre os algoritmos, foi optado por deixar valores padrões que acompanham os algoritmos ou que são mais fáceis de entender.

Ao final do desenvolvimento de todos os *workflows* foi feita uma pesquisa qualitativa com alguns membros que ainda continuam ativamente no Movimento Empresa Júnior, em que foram apresentados todos os *workflows* e seus resultados e em seguida foram feitas 3 perguntas:

- Você acha que a utilização desse *workflow* é intuitiva?
- Você acha que esses resultados são úteis?

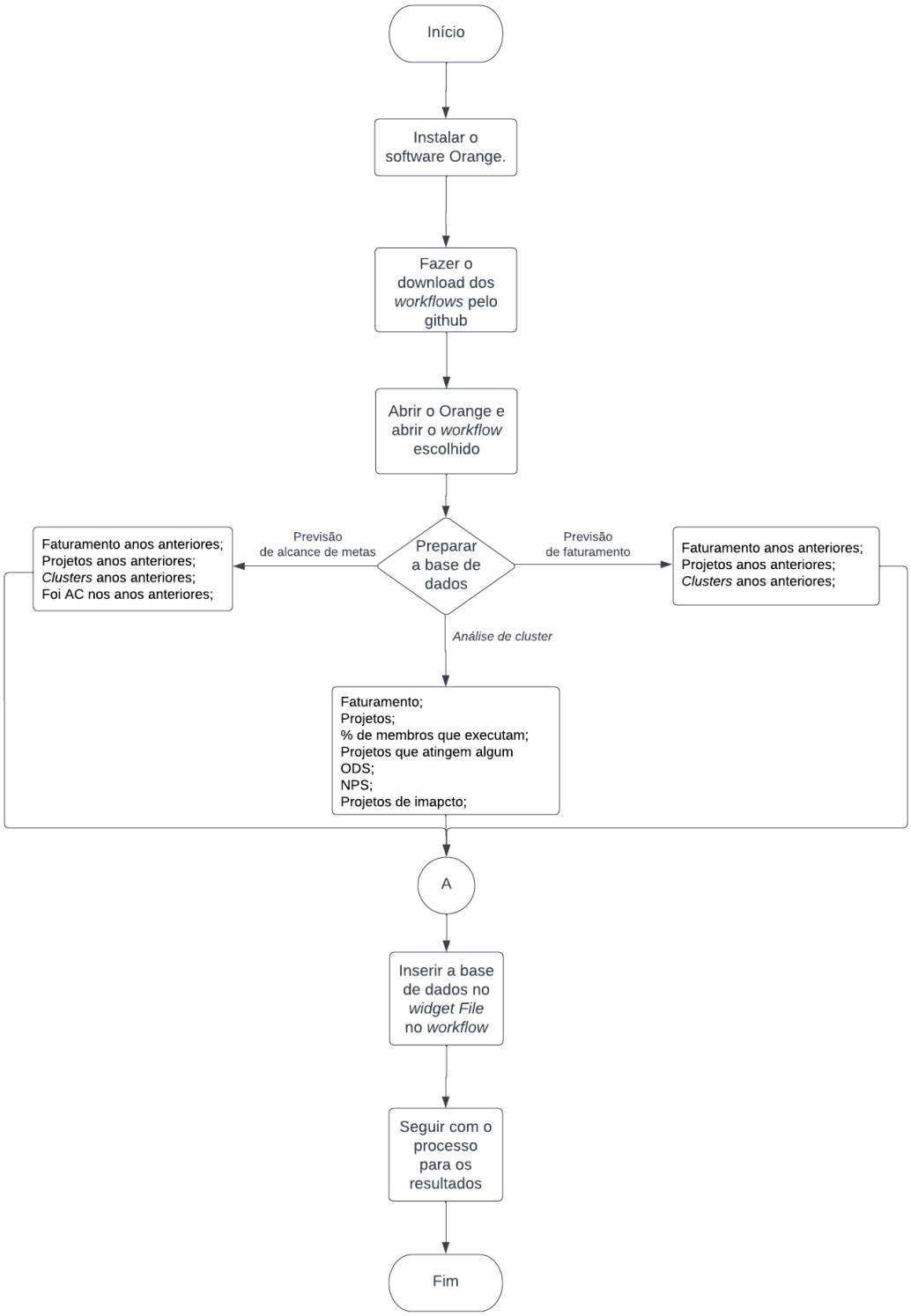
- Você acredita que conseguiria e se beneficiaria de utilizar esses workflows no seu trabalho?

De maneira geral, os respondentes acreditaram que o *workflow* 4 era o mais difícil de ser utilizado, mas acreditam que com o correto direcionamento todos os *workflows* poderiam ser utilizados. Sobre os resultados, os gestores entenderam que os resultados de alcance de metas eram mais fáceis de serem analisados e também ajudariam muito no início e ao longo do ano para os núcleos e federações direcionarem mais esforço para àquelas EJs que o algoritmo aponta como prováveis para não baterem as suas metas.

Além disso, os resultados do *workflow* 3 chamaram a atenção dos gestores pelo baixo errro que ele apresenta. De modo geral, os gestores disseram sentir dificuldades em ajustar todos os dados necessários para colocar no *workflow*, mas acreditam que isso sendo feito os resultados obtidos podem ajudar para definição de metas das EJs e acompanhamento do seu desempenho ao longo do ano. De maneira geral, os gestores aprovaram os módulos mas precisariam de uma ajuda mais profunda para conseguir implementar no seu dia-a-dia e usar como uma ferramenta cotidiana.

Para que os gestores possam utilizar da ferramenta foi criado um manual de uso, visto no Apêndice B e um fluxograma que mostra o que deverá ser feito para que os gestores possam utilizar dos *workflow*. O fluxograma é apresentado na Figura 21.

Figura 21 – Fluxograma de uso para os *workflows*.




Fonte: Autora (2022).

Na Figura 21 apresenta-se o manual para utilizar dos *workflows* como um passo-a-passo para que os gestores consigam de maneira fácil utilizá-los.

# 4 CONSIDERAÇÕES FINAIS

Serviços de *data science* e análise de dados estão cada vez mais presentes nas empresas, uma vez que dados serão cada vez mais importantes para que as empresas se mantenham competitivas no mercado. Dentro do MEJ isso não será diferente: as empresas juniores precisam aprender a utilizar mais dos seus dados e conseguir fazer análises cada vez mais profundas para que o movimento consiga crescer de maneira mais sustentável.

O estudo sobre *data science* também é uma área que vem crescendo muito dentro das pesquisas em universidades e empresas, assim como há uma grande difusão de conhecimentos sobre o assunto na internet, o que faz com que cada vez mais pessoas consigam entender sobre o assunto. A inteligência artificial vem hoje sendo usada em diversos ramos da indústria e também em ramos de pesquisa e desenvolvimento e tem tudo para ser uma grande área de desenvolvimento nos próximos anos.

As técnicas e algoritmos sobre *data science* vem sendo cada vez mais estudados e vem se tornando mais forte ao longo dos anos, conseguindo hoje em dia fazer análises em bancos de dados com mais de milhões de dados. Por isso, vem a importância de dominar as técnicas mas sobretudo conseguir coletar os melhores dados que conseguirão entregar as respostas para o que o negócio quer responder.

Parte importante dos desenvolvimentos vem de coletar e pré-processar os dados certos, por isso também há uma grande necessidade de um conhecimento profundo sobre o negócio que quer se fazer a análise, dessa maneira os empreendedores ou gerentes se tornam essenciais para que sejam feitas as perguntas certas e sejam coletados os dados mais assertivos. Grande parte da responsabilidades das empresas também vem em coletar esses dados de maneiras a respeitar a Lei Geral de Proteção de Dados (LGPD) e tentar coletar a maior quantidade de dados possíveis.

Neste trabalho, foram aplicados diversos algoritmos ligados à *data science* para responder diversas perguntas diferentes, e foram apenas alguns exemplos de como essas técnicas podem ser usadas dentro de análise no MEJ, mas que também podem ser levadas ao contexto de micro e pequenas empresas que podem ter mais previsibilidade para o seu negócio. Foi possível ver no trabalho que alguns algoritmos trouxeram resultados melhores, enquanto outros não conseguiram trazer resultados tão bons, variando também conforme a base de dados disponíveis.

De maneira geral, pode-se dizer que os resultados do trabalho foram na maioria satisfatórios, conseguindo trazer pelo menos uma direção melhor para os resultados do MEJ. Em avaliações mais qualitativas com pessoas ligadas ao movimento, foi

possível identificar que os resultados da análise *cluster* por exemplo faziam sentido qualitativamente, em sua maioria na criação dos novos agrupamentos.

Ainda, durante o trabalho poderiam ter sido melhor explorados os dados a nível Brasil, juntando diferentes EJs do Brasil inteiro para entender qual seria o comportamento do algoritmo, assim como fazer previsões à nível Brasil, o que foi prejudicado pela manipulação que era necessária de ser feita antes de passar os dados para o Orange. Outros algoritmos, principalmente para a previsão, também poderiam ter sido testados, assim como diferentes configurações para a regressão linear. Porém, como a ideia era deixar algo fácil para os gerentes entenderem, optou-se por deixar de maneira padrão.

O Orange também possui diversas outras ferramentas adicionais que podem ser incorporadas à plataforma de maneira a trazer outros *widgets*, algoritmos e maneiras de avaliar os dados e os resultados. Um exemplo disso é o pacote adicional de associação, onde há o *widget* de regras de associação que poderia ser usado para obter mais precisão sobre os dados para fazer as previsões de faturamento. Uma vez que adicionar esses elementos ao trabalho aumentaria sua complexidade essa não foi uma das escolhas para o trabalho.

Para projetos futuros poderiam ser explorados dados textuais, como atas de empresas juniores ou diversos outros relatórios para adicionar mais dados à nossa base. Além disso, conforme os anos forem passando e a base de dados das EJs forem aumentado, há uma tendência de os resultados melhorarem por conta do aumento do número de dados, dessa maneira, num trabalho futuro se utilizadas essa técnicas descritas, provavelmente obteriam-se melhores resultados para as previsões de faturamento e de EJ AC.

Outro ponto para se levar em consideração é que o MEJ é um ambiente muito dinâmico, levando em consideração que o planejamento estratégico da rede muda a cada 3 anos e com isso algumas metas gerais podem ser excluídas e outras novas podem entrar, isso pode levar a não ter uma quantidade linear de dados, e pode também influenciar na análise e em resultados. Por isso, esse trabalho também focou em métricas que vem se mantendo ao longo dos planejamentos estratégico, como o faturamento e a métrica de avaliação se a EJ bateria sua meta de alto crescimento ou não.

Um ponto que também poderia ter sido melhor trabalhado está em levar esses *workflows* para gerentes de EJs ou de núcleos e federações para entender se eles seriam intuitivos para essas pessoas assim como o uso do Orange e se os resultados seriam de fácil entendimento. É importante que esse seja um conhecimento difundido no MEJ e que as pessoas consigam extrair bons *insights* a partir dos resultados apresentados.

Todos os *workflows* apresentados neste trabalho podem ser encontrados no

repositório do GitHub por meio do link: https://github.com/marianaaas/orange-tcc.git.

# REFERÊNCIAS

AMARAL, F. **Introdução à ciência de dados: mineração de dados e Big Data**. Alta Books, 2018. ISBN 9788550804163. Disponível em: https://books.google.com.br/books?id=A5uODwAAQBAJ.

BECKER, G.; SILVA, M. A. D. O. C. D. **A CONTRIBUIÇÃO DA EMPRESA JÚNIOR NO PROCESSO DE FORMAÇÃO DO PROFISSIONAL DE ENGENHARIA DE PRODUÇÃO**. Monografia (Trabalho de Conclusão de Curso) — UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ, 2017.

BRASIL JúNIOR. **Fundamento dos Cluster 2.0**. 2019. Disponível em: https://fejers.org.br/wp-content/uploads/2019/10/Sistema-de-Clusters-2019-2021-Articula\%C3\%A7\%C3\%A3o-pr\%C3\%A9-ENEJ.pdf. Acesso em: 10 jul. 2022.

BRASIL JúNIOR. **Como funciona uma Empresa Júnior? Descubra agora**. 2020. Disponível em: https://brasiljunior.org.br/conteudos/como-funciona-uma-empresa-junior-descubra-agora. Acesso em: 13 fev. 2022.

BRASIL JúNIOR. **Dashboard da rede**. 2021. Disponível em: bit.ly/DASHDAREDE. Acesso em: 13 fev. 2022.

BRASIL JúNIOR. **Painel de resultados da rede**. 2021. Disponível em: bit.ly/dashrede. Acesso em: 13 fev. 2022.

BRASIL JúNIOR. **Portal da transparência**. 2022. Disponível em: https://brasiljunior.org.br/portal-da-transparencia. Acesso em: 06 out. 2022.

DOMINGUES, M. L. C. S. Dissertação submetida à avaliação, como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação., **Mineração de dados utilizando aprendizado não-supervisionado: um estudo de caso para bancos de dados da saúde**. 2003.

FAWCETT, T.; PROVOST, F. **Data Science para Negócios: O que você precisa saber sobre mineração de dados e pensamento analítico de dados**. Alta Books, 2018. ISBN 9788550803906. Disponível em: https://books.google.com.br/books?id=1rZwDwAAQBAJ.

FIGUEIREDO FILHO, D. B.; SILVA JÚNIOR, J. A. Desvendando os mistérios do coeficiente de correlação de pearson (r). **Revista Política Hoje**, v. 18, n. 1, p. 115–146, 2009. ISSN 0104-7094. Disponível em: https://periodicos.ufpe.br/revistas/politicahoje/article/view/3852.

FORRESTER. **Unveiling Data Challenges Afflicting Businesses Around The World**. 2021. Disponível em: https://www.delltechnologies.com/asset/en-us/solutions/industry-solutions/industry-market/data-paradox-forrester-thought-leadership-paper.pdf. Acesso em: 13 fev. 2022.

INTERNATIONAL BUSINESS MACHINES CORPORATION. **Data Mining**. 2021. Disponível em: https://www.ibm.com/cloud/learn/data-mining. Acesso em: 26 jan. 2022.

INTERNATIONAL BUSINESS MACHINES CORPORATION. **Supervised vs. Unsupervised Learning: What's the Difference?** 2021. Disponível em: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning. Acesso em: 26 jan. 2022.

HERCOS JUNIOR, J. B. **Análise da eficiência relativa das Empresas Juniores no Brasil**. Tese (Tese submetida ao Programa de Pós-Graduação em Ciências Econômicas da Universidade Estadual de Maringá como requisito parcial para a obtenção do título de Doutor em Economia.) — Universidade Estadual de Maringá, 2014.

LAROSE, D.; LAROSE, C. **Discovering Knowledge in Data: An Introduction to Data Mining**. [S.l.]: Wiley, 2014. (Wiley Series on Methods and Applications in Data Mining). ISBN 9781118873571.

ORANGE DATA MINING. **Correlações**. 2022. Disponível em: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/data/correlations.html. Acesso em: 30 out. 2022.

ORANGE DATA MINING. **K-means interativo**. 2022. Disponível em: https://orangedatamining.com/widget-catalog/educational/interactive-kmeans/. Acesso em: 10 ago. 2022.

ORANGE DATA MINING. **Orange Data Mining**. 2022. Disponível em: https://orangedatamining.com/. Acesso em: 28 fev. 2022.

ORANGE DATA MINING. **Test and Score**. 2022. Disponível em: https://orangedatamining.com/widget-catalog/evaluate/testandscore/. Acesso em: 20 maio. 2022.

SAVIETTO, J. V. **Machine Learning: Métricas, Validação Cruzada, Bias e Variância**. 2021. Disponível em: https://medium.com/@jvsavietto6/machine-learning-m\%C3\%A9tricas-valida\%C3\%A7\%C3\%A3o-cruzada-bias-e-vari\%C3\%A2ncia-380513d97c95. Acesso em: 24 ago. 2022.

SOUZA, J. C. O. D.; GALVÃO, M. R. M. **MODELAGEM DE GESTÃO DE PROCESSOS DE NEGOCIOS NA AREA DA TECNOLOGIA: UM ESTUDO DE CASO DA EMPRESA JÚNIOR UTIC**. Trabalho de Conclusão de Curso — UNIVERSIDADE FEDERAL RURAL DA AMAZÔNIA, 2019.

# APÊNDICE A - DICIONÁRIO DE DADOS

Tabela 13 – Dicionário de dados

| Coluna | Descrição | Tamanho | Tipo | Valores aceitos |
|---|---|---|---|---|
| EMPRESA JÚNIOR | Nome da empresa júnior | 0 - 255 | Texto | ... |
| FEDERAÇÃO | Nome da federação a qual a empresa júnior é federada | 0 - 30 | Texto | ... |
| NÚCLEO | Nome do núcleo no qual a empresa júnior é nucleada | 0 - 30 | Texto | ... |
| CLUSTER 2020 | Nº do índice de cluster em 2020 | 1 | Numérica | 1 \| 2 \| 3 \| 4 \| 5 |
| CLUSTER 2021 | Nº do índice de cluster em 2021 | 1 | Numérica | 1 \| 2 \| 3 \| 4 \| 5 |
| CLUSTER 2022 | Nº do índice de cluster em 2022 | 1 | Numérica | 1 \| 2 \| 3 \| 4 \| 5 |
| É DE ALTO CRESCIMENTO? | A empresa júnior alcançou as suas metas referentes a meta de alto crescimento? | 3 | Binária | Sim \| Não |
| META de Faturamento | meta de faturamento | 10 | Numérica | 0 - |
| META de % de membros que executaram | meta de % de membros que executam projetos | 3 | Numérica | 0 - 100 |
| META de Número de Projetos | meta de nº de projetos | 10 | Numérica | 0 - |
| META de % de participação em eventos | meta de % de participação em eventos, com relação ao nº de membros | 3 | Numérica | 0 - 100 |
| META de NPS | meta de nps | 3 | Numérica | 0 - 100 |
| META de Número de Projetos de Impacto | meta de número de projetos de impacto (nps 9 ou 10 e atingindo uma ODS) | 10 | Numérica | 0 - |
| FATURAMENTO | faturamento atual da empresa júnior | 10 | Numérica | 0 - |
| PROJETOS | nº de projetos atual da empresa júnior | 10 | Numérica | 0 - |
| % DE MEMBROS QUE EXECUTARAM PROJETOS | % de membros que executaram pelo menos um projeto | 3 | Numérica | 0 -100 |
| PROJETOS QUE ATINGEM ALGUM ODS | projetos que atingem algum ods | 10 | Numérica | 0 - |
| Nº DE AÇOES COMPARTILHADAS | nº bruto de ação compartilhada | 10 | Numérica | 0 - |
| NPS | nps dos contratos | 3 | Numérica | 0 -100 |
| PROJETOS DE IMPACTO | nº de projetos de impacto | 10 | Numérica | 0 - |

Fonte: Autora.

# APÊNDICE B - MANUAL DE USO

Passo-a-passo para a aplicação dos métodos apresentados neste trabalho:

1. Fazer o download do software Orange a partir do seu site: https://orangedatamining.com/download/#windows.

2. Fazer o download dos *workflows* a partir do github: https://github.com/marianaaas/orange-tcc.git.

3. Abrir o software Orange e abrir o documento desejado.

4. Preparar a base de dados correspondente para o *workflow* desejado com os dados necessários que foram apresentados no método.

5. Inserir a base de dados no *widget File* por meio de um documento .xlsx e fazer as previsões ou o novo agrupamento seguindo o método apresentado.