Probabilidade de uma empresa júnior / instância bater sua meta a partir da base de dados da FEJESP

1st Luiza Araujo de Oliveira Caram Saliba
Science and technology institute (UNIFESP)
São José dos Campos, Brazil
luiza.saliba@unifesp.br

2nd Harrison Caetano Candido
Science and technology institute (UNIFESP)
São José dos Campos, Brazil
h.candido20@unifesp.br

Resumo*—Este estudo visa analisar a probabilidade de empresas juniores atingirem suas metas utilizando dados fornecidos pela FEJESP (Federação das Empresas Juniores do Estado de São Paulo). A FEJESP, que promove o empreendedorismo e a capacitação profissional, possui uma rica base de dados que pode revelar fatores determinantes para o sucesso dessas empresas. Utilizando técnicas de mineração de dados e análise estatística, este estudo emprega um modelo de árvore de regressão para prever a probabilidade de sucesso das empresas juniores. A metodologia inclui etapas de conhecimento de domínio, pré-processamento dos dados, extração de padrões, pós-processamento e utilização do conhecimento. Espera-se que os insights obtidos ajudem gestores a tomar decisões estratégicas, fortalecendo o movimento de empresas juniores no estado de São Paulo.

***Palavras chave* — árvore de regressão, empresa júnior, FEJESP, faturamento, inovação, aprendizado de máquina**

I. Introdução e Motivação

A análise da probabilidade de uma empresa júnior atingir suas metas é essencial para entender os fatores que contribuem para o sucesso dessas organizações. A FEJESP (Federação das Empresas Juniores do Estado de São Paulo) possui uma rica base de dados que pode ser explorada para obter insights valiosos sobre o desempenho dessas empresas. A FEJESP desempenha um papel crucial no desenvolvimento e suporte das empresas juniores, promovendo o empreendedorismo, a capacitação profissional e a integração entre estudantes e o mercado, sendo responsável por faturar mais de vinte e um milhões de reais no ano passado.

Este estudo tem como objetivo analisar os dados fornecidos pela FEJESP para calcular a probabilidade de uma empresa júnior alcançar suas metas e, com isso auxiliar no processo aquelas que tem probabilidades mais baixas de atingir os objetivos.

Através da utilização de técnicas de mineração de dados e ferramentas de análise estatística, será possível identificar padrões, tendências e variáveis que influenciam o sucesso das empresas juniores no estado de São Paulo. Com isso, espera-se fornecer informações que possam auxiliar na tomada de decisões estratégicas, contribuindo para o fortalecimento do movimento de empresas juniores e promovendo a eficiência e eficácia na gestão dessas organizações.

-

.

II. Conceitos Fundamentais

A. Mineração de Dados

O avanço das tecnologias de aquisição e armazenamento de dados, como a internet, levou a criação da área de ciência de dados. Definimos a mineração de dados como uma área que estuda a descoberta de padrões e conhecimento para inteligência a partir de dados estruturados ou não [1].

No processo de estudo e análise dos dados nós seguimos uma trajetória que inicía em conhecimento de domínio (coleta dos dados), Pré-processamento (limpeza, transformação, formatação, filtragem etc), Extração de padrões (Treino e teste de modelos para extração de padrões) e Pós-Processamento (Análise dos padrões descobertos nos dados) [1].

B. Aprendizado de Máquina

O aprendizado de máquina é uma linha de pesquisa da área de inteligência artificial que trata de ensinar modelos com dados de exemplos e assim melhorar a acurácia em uma aplicação alvo [2].

Segundo a noção de aprendizado indutivo, dizemos que o aprendizado é expresso por uma função objetivo f que mapeia um conjunto X (amostra de distribuição desconhecida) a um conjunto Y, sendo X um conjunto de dados que usamos para treinar e testar o modelo e Y o valor a ser previsto [2].

Nosso objetivo aqui é encontrar uma hipótese h que melhor se ajuste a função f dentre as infinitas hipóteses que temos no chamado espaço de hipóteses, que são as infinitas possibilidades de performance por configuração do nosso modelo utilizado [2].

Sempre que estamos trabalhando com Modelos de aprendizado de máquina devemos nos atentar a escolha de hipóteses que sejam consistentes (hipóteses que se ajustam a todos os dados da amostra), menos complexas (dentre as possibilidades de hipóteses consistentes, escolher a menos complexa, seguindo a Lâmina de Ockham) e generalistas (se ajustam bem a um conjunto de treino diferente)[2].

C. Aprendizado Supervisionado

O aprendizado de máquina supervisionado consiste na ideia de conhecer os valores de Y (saída) e treinar e testar o modelo em X para reconhecer erros e acertos e apontar ao modelo qual

resultado está mais próximo do desejado. Aqui nós temos dois tipos de saída: Y discreta (tarefa de classificação) e Y contínua (tarefa de regressão) [2].

D. *Árvores de Regressão*

Este tipo de modelo de aprendizado supervisionado trabalha com a tarefa de regressão, que representa uma função que recebe um vetor de atributos e retorna um valor contínuo como saída [2]. Visto que queremos dar a probabilidade de uma empresa junior atingir seus objetivos com uma facilidade de interpretação, nós usamos este modelo como base do projeto. Definimos os nós internos como conjuntos de dados divididos por atributos de teste, cada aresta como um possível valor do atributo testado e cada nó folha como um valor a ser retornado pela função [2]. Aqui cada exemplo de teste é medido por algum método como o MSE (mean squared error), que considera todas os subconjuntos criados a partir de cada atributo em cada nó e escolhe aquele que resulta na maior redução no MSE quando este é comparado com o MSE do pai [2]. Isso é repetido recursivamente até que uma condição de parada como a profundidade máxima ou o número mínimo de amostras por folha seja alcançado [2]. Fazemos isso com o MSE para obter uma precisão maior na previsão dos valores contínuos de saída. Este modelo se ajusta perfeitamente, pois é relativamente simples de implementar e interpretar.

E. *Overfitting e Prunning*

Árvores de regressão são muito passivas de sofrerem overfitting (sobreajuste) nos dados, isso por que nossa base de dados contém valores em atributos que variam muito, o que leva a uma alta variância, ou seja, leva a maior incidência de overfitting [2]. Um motivo minimizador de overfitting é quando temos uma base de dados com maior quantidade de exemplos de treino do que atributos, o que é garantido na nossa base de dados [2].

O Prunning (ou Poda) é a técnica que entra para combater o overfitting [2]. Quando encontramos nós teste que contenham apenas um único nó folha, nós iremos pegar o conjunto folha em questão e juntar seus exemplos com os do seu nó pai, eliminando este nó pai que possuía um atributo considerado irrelevante, ou seja, havia um ruído e nós o eliminamos [2]. Fazemos a poda até onde consideramos relevante, isto é, podamos quantos nós teste irrelevantes pudermos [2]. A relevância de um atributo pode ser medida por um teste de signficância, que pode utilizar o método da Hipótese Nula ou o Ganho de informação [2].

III. Trabalhos Relacionados

Para desenvolver este trabalho nos baseamos em alguns artigos já publicados e encontrados através de buscas no Google Scholar, utilizando as strings “Empresa júnior” e “Machine Learning”.

A. *Artigos*

- O paper “Planejamento Estratégico da Rede: Resultados do movimento empresa júnior a partir de estímulo para crescimento”, escrito por Rutzen e Paula Benedetti utiliza duas bases de dados da FEJESP para medir lacunas como metas a serem atingidas e lacunas como metas já atingidas. Os resultados mostram os estados com maior e menor percentual de crescimento, bem como os grupos de empresas juniores.

IV. Objetivo

O principal objetivo deste projeto é analisar a probabilidade de uma empresa júnior ou instância alcançar suas metas, utilizando dados disponíveis na base da FEJESP (Federação das Empresas Juniores do Estado de São Paulo). Por meio da aplicação de técnicas avançadas de análise de dados e modelagem estatística, o projeto visa identificar padrões, tendências e fatores que influenciam o desempenho das empresas juniores no Estado de São Paulo e fornecer insights para gestores e tomadores de decisão melhorarem a performance de suas empresas.

V. Metodologia Experimental

Como foi dito, o modelo estatístico escolhido foi o regression tree para prever a probabilidade de uma empresa junior ou instância alcançar suas metas. A tecnologia utilizada será o interpretador Python de qualquer versão e a infraestrutura será o Google Colab.

A. *Principais etapas da metodologia*:

- Conhecimento de Domínio : Realizar uma investigação detalhada dos dados fornecidos pela FEJESP para compreender a estrutura e as características dos dados, identificando possíveis desafios e oportunidades.
- Pré-Processamento: Esta fase colabora com a seleção dos atributos mais relevantes para serem incluídos no modelos, tratamento de valores ausentes em atributos, codificação de variáveis e limpeza de valores sujos (não legíveis pelo modelo).
- Extração de Padrões: Utilizar to modelo estatístico árvore de regressão com alguma técnica de validação cruzada como K-Fold ou Holdout para avaliar a performance do modelo e estimar a capacidade de generalização para diferentes conjuntos de dados.
- Pós-Processamento: Avaliar a performance do modelo desenvolvido utilizando métricas como MSE (mean square error), coeficiente de determinação e gráficos de dispersão (mede as previsões e os valores reais, lembrando que estamos utilizando um modelo de aprendizado supervisionado).
- Utilização do Conhecimento: Identificar e analisar os principais fatores que influenciam a probabilidade de sucesso das empresas juniores, fornecendo insights valiosos para gestores e tomadores de decisão.

B. Principais *bibliotecas*:

- Pandas: para manipular dataframes.
- Numpy: para manipular matrizes.
- Matplotlib e Seaborn: para visualizar dados.
- Scikit-Learn: para pré-processamento, implementação do modelo de regression tree e calcular métricas de avaliação.

VI. O Que Será Entregue No Final?

Será entregue um modelo de aprendizado supervisionado regression tree, capaz de performar em bases de dados de diferentes anos da FEJESP para auxiliar na tomada de decisão dos líderes das empresas júniores a fim de que atinjam suas metas, bem como também será entregue uma dashbaord interativa para mostrar a probabilidade de cada empresa júnior alcançar suas metas.

A FEJESP como federação de coordenação e apoio às empresas juniores do estado de São Paulo será capaz de prover estímulos às diferentes empresas júniores através dos seus núcleos, bem como as próprias empresas juniores serão capazes de analisar seu próprio desempenho naquele período e a se preparar antecipadamente para atingir os seus resultados, tudo isso orientado a dados.

Referencias

[1] S. O. Rezende, “Sistemas Inteligentes Fundamentos e Aplicações”, 2003.

[2] S. Russel, P. Norvig, “Artificial Intelligence: A Modern Approach,” 3rd ed, pp. 529–551, 2009.