Aprendizado Não Supervisionado - Clustering dos alunos da Engenharia de Computação

Hugo de Andrade Leuchs, Willian Rodrigo Huber

Sistemas Inteligentes

Engenharia de Computação, UTFPR Toledo/PR

1 de dezembro de 2021

1 Base de dados

Com o intuito da elaboração do presente trabalho foi utilizado uma base de dados dos alunos de Engenharia de Computação da Universidade Tecnológica Federal do Paraná do campus Toledo. Antes de iniciar qualquer tipo de pré processamento é preciso se atentar ao ponto relativo ao conjunto de caracteres reconhecidos, isto posto, para evitar possíveis problemas os acentos foram removidos do arquivo **alunos_engcomp.csv**. Outro ponto a ser comentado é o fato dos elementos do cabeçalho terem tido o primeiro caractere alterado de maiúsculo para minusculo. Os primeiros campos da base descrita podem ser vistos na tabela a seguir.

Tabela 1: Base de dados - Sem pré-processamento

sexo	coeficiente	periodo	escola	enem
M	0.725	10	publica	613.00
M	0.0738	1	publica	594.07
M	0.4205	2	publica	599.80
M	0.0000	1	publica	529.83
M	0.3770	2	particular	530.57

Com a base de dados apresentada, pode-se verificar uma definição acerca de cada uma das colunas vistas na Tabela 1.

Tabela 2: Base de dados - Descrição dos campos

Coluna	Descrição
sexo	Sexo do aluno: M (masculino), F (feminino)
coeficiente	Coeficiente de rendimento academico do aluno
periodo	período em que o aluno se encontra
escola	Tipo de escola que o aluno frequentou: Pública ou Privada
enem	Nota que o aluno obteve no ENEM

A figura Tabela 1 apresenta os campos de interesse (sexo, coeficiente, periodo, escola e enem) de cada um dos alunos, pode-se atentar ainda ao fato de existirem dados não numéricos, que podem interferir no andamento do projeto, para tal sera realizado um pré-processamento para torna-los no formato adequado, o resultado dessa etapa pode ser visto a seguir.

Tabela 3: Base de dados - Com pré-processamento

sexo	coeficiente	periodo	escola	enem
0	0.725	10	0	613.00
0	0.0738	1	0	594.07
0	0.4205	2	0	599.80
0	0.0000	1	0	529.83
0	0.3770	2	1	530.57

2 Implementação e Resultados

Após a realização do pré-processamento dos dados foi realizada a distinção dos dados de entrada (coeficiente, periodo, escola, enem) e saída (sexo) para posterior divisão dos dados em dois grupos (teste e treinamento) de forma aleatória. Após essas etapas foi aplicado um algoritmo de árvore de decisão utilizando a biblioteca **tree** do **sklearn**. A representação gráfica da mesma pode ser vista abaixo.

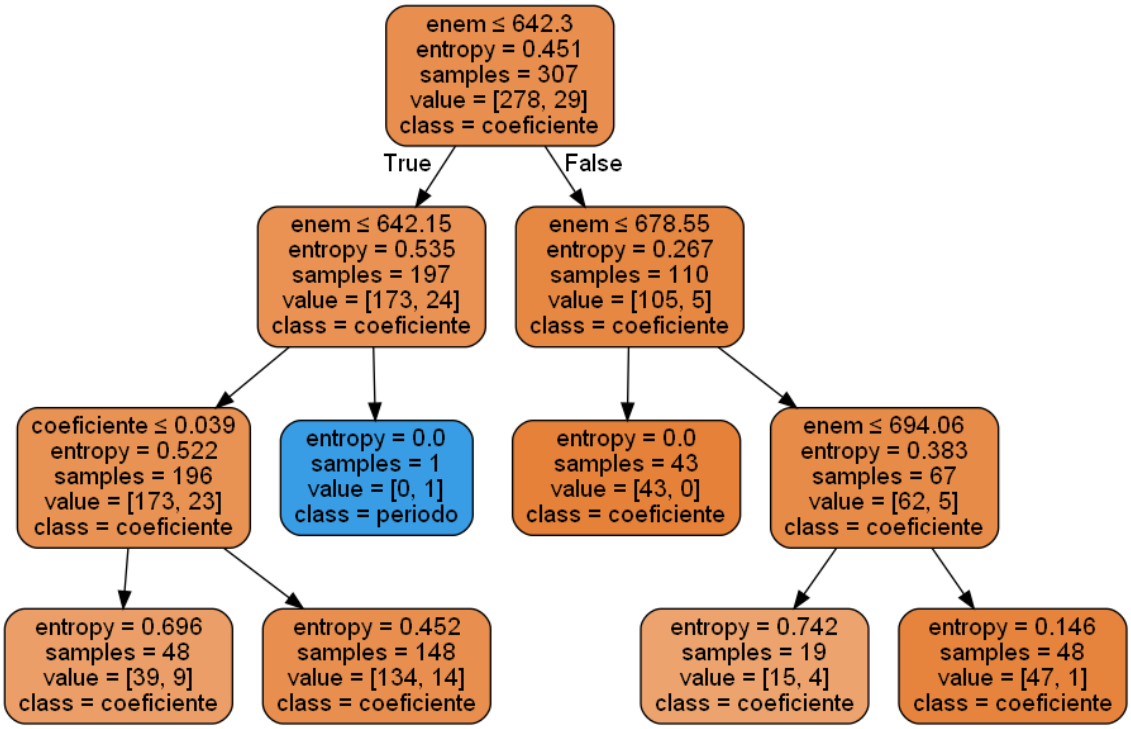

Figura 1: Resultados obtidos (**max_depth** = 3)

Pode se observar na Figura 2 cada um dos nós representa uma decisão tomada pelo algoritmo baseado nos dados amostrado pela base apresentada previamente, processo este repetido até se atingir um nó folha. No contexto de Machine Learning, o nó raiz é um atributo da base de dados, enquanto o nó folha é a resposta gerada para essa informação.

Para a realização do processo de decisão em cada um dos nós é realizado um conjuntos cálculos acerca de características singulares dos dados, dentre esses métodos pode-se citar o ganho de informação e a entropia. Essas duas variáveis dizem respeito à desorganização e falta de uniformidade nos dados. Quanto mais alta a entropia, mais caóticos e misturados estão os dados. Quanto menor a entropia, mais uniforme e homogênea está a base.

Para definir os posicionamentos, é preciso calcular a entropia das classes de saída e o ganho de informação dos atributos da base de dados. Quem tiver maior ganho de informação entre os atributos é o nó-raiz. Para calcular a esquerda e a direita, deve-se realizar novos cálculos de entropia e ganho com o conjunto de dados que atende à condição que leva à esquerda ou à direita.

Como os dados agrupados em cada um dos nós são aproximadamente homogêneos, tem-se um alto ganho de informação e conhecimento acerca do comportamento dos dados. Durante a implementação do algoritmo da árvore de decisão existem alguns atributos que alteram a maneira com a estruturação da árvore é realizada, uma propriedade relevante é **max_depth** que define a profundidade mínima da árvore, ou seja, variando seu valor o algoritmo passa a considerar um conjunto de características

maior na hora de realizar a etapa de treinamento, consequentemente no cálculo da entropia e ganho de informação. A Figura 2 possuí um **max_depth** igual a 3, aumentando o mesmo para 5 obteve-se o resultado da figura abaixo.

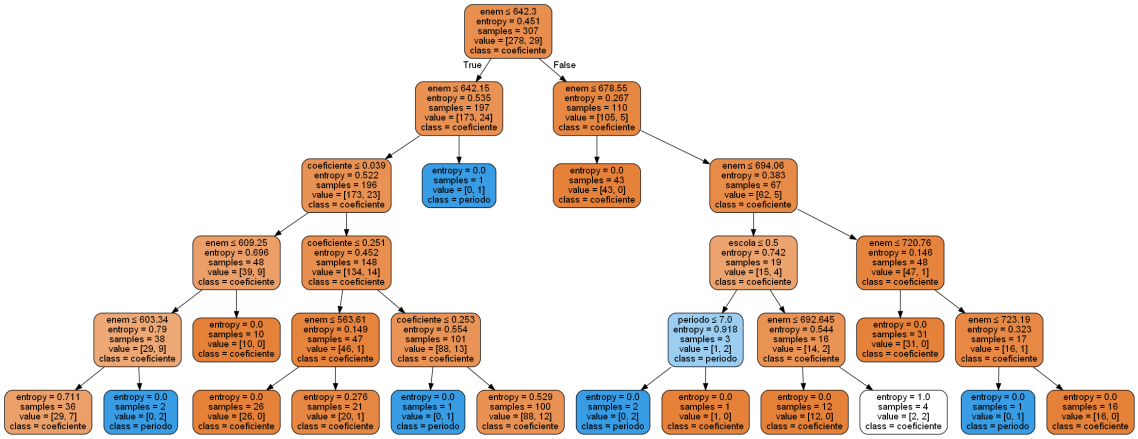

Figura 2: Resultados obtidos (**max_depth** = 5)

Uma curiosidade acerca do modelo, é o fato do aumento do **max_depth** diminuir a acurácia do mesmo, com um valor de **max_depth** = 3 a assertividade foi de aproximadamente 88% enquanto com um **max_depth** = 5 seu valor diminuiu para aproximadamente 82%.

3 Discussão do resultados

O cerne para a classificação de cada nó foi o sexo dos alunos, foi optado pela escolha desta característica, principalmente pelo fato da mesma ser binária (Homem ou Mulher), e pelo seu impacto direto pelo ponto de vista social, principalmente no âmbito de áreas STEM (Science, Technology, Engineering e Math) onde as mulheres contam com uma taxa de participação baixa quando comparadas aos homens, alguns aspectos podem ser levados em consideração para a perspectiva descrita, tais como:

- Estereótipos de gênero: A uma clara prepotência na atribuição de campos das áreas STEM a homens, além de que ocorre de forma intrínseca a depreciação da participação da mulher nas áreas descritas.
- Culturas dominadas por homens: Como a grande maioria parcela de membros das áreas STEM são compostas por homens não ocorre um apoio, muito menos incentivo a inserção de mulheres na área.
- Menos modelos de papéis: Não existem uma grande variedade de mulheres que sirvam de exemplo, e atraiam dessa forma as mesmas a essa área.

- Ansiedade matemática: De forma indireta como há um número maior de professoras a uma certa pressão em cima das meninas para que as mesmas alcancem resultados mais elaborados e satisfatórios que os homens.

Tais fatores podem de forma direta ou indireta influenciar o número de mulheres atuantes em áreas STEM.