Universidade Federal de Santa Maria
Cursos: Ciência da Computação e Sistemas de Informação
Alunos: Diogo Rocha e Francisco das Chagas
Prof: Dr. Joaquim Assunção
Disciplina: Mineração de Dados

Trabalho 2: Análise de Dados e Resultados do Código

Introdução

Este projeto foi desenvolvido para analisar um conjunto de dados sobre ingressantes e formandos, com foco na comparação entre a quantidade de formandos do sexo masculino e feminino. O código implementa duas abordagens principais: **Árvore de Decisão** e **Apriori**.

Estrutura do Código

1. Importação de Bibliotecas

O código começa com a importação de bibliotecas essenciais para análise de dados e aprendizado de máquina:

```
import pandas as pd import matplotlib.pyplot as plt import
seaborn as sns from mlxtend.frequent_patterns import apriori,
association_rules from sklearn.preprocessing import
StandardScaler, LabelEncoder from sklearn.decomposition import
PCA from sklearn.model_selection import train_test_split from
sklearn.ensemble importRandomForestClassifier
from sklearn.metrics import
classification_report,confusion_matrix
```

2. Funções de Processamento e Análise

a. Função `carregar_dados`

Esta função carrega dados de um arquivo Excel ou CSV e os retorna como um DataFrame do Pandas.

b. Função `verificar_e_transformar_dados`

Transforma os dados em uma matriz de cestas para a análise com o algoritmo Apriori. Essa função também exibe um exemplo dos dados.

c. Função `aplicar_apriori`

Aplica o algoritmo Apriori para identificar conjuntos frequentes de itens em uma matriz de cestas.

d. Função `gerar_regras`

Gera regras de associação a partir dos conjuntos frequentes encontrados, usando métricas como confiança e lift.

e. Função `gerar_grafico_comparacao_sexo`

Cria um gráfico de barras comparando a quantidade de formandos por sexo.

f. Funções para a Árvore de Decisão

- ⓾ `explorar_dados`: Exibe um resumo estatístico dos dados e um mapa de calor indicando valores ausentes.
- ⓾ `limpar_dados`: Remove duplicatas e preenche valores ausentes com a mediana das colunas numéricas.
- ⓾ `transformar_dados`: Normaliza as variáveis numéricas e transforma variáveis categóricas com codificação de rótulos.
- ⓾ `reduzir_dimensao`: Aplica PCA para reduzir a dimensionalidade dos dados.
- ⓾ `treinar_modelo`: Treina um modelo de classificação RandomForest e exibe um relatório de classificação e uma matriz de confusão.

3. Função `main`

A função principal orquestra a execução do código, solicitando ao usuário a escolha entre a análise com a árvore de decisão ou o algoritmo Apriori.

Resultados

Análise dos Dados

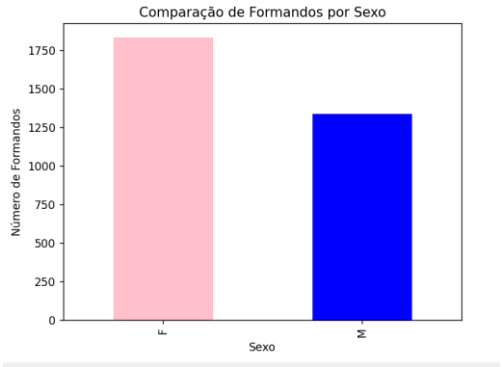
Os dados utilizados foram os dos Ingressantes e Formandos por centro e sexo. Existem 9 centros que são os seguintes: Centro de Artes e Letras (CAL), Centro de Ciências Naturais e Exatas (CCNE), Centro de Ciências Rurais (CCR), Centro de Ciências da Saúde (CCS), Centro de Ciências Sociais e Humanas (CCSH), Centro de Educação (CE), Centro de Educação Física e Desportos (CEFD) e Centro de Tecnologia (CT). Além disso, os dados possuem informações como o nome do curso e o ano, além da quantidade de ingressos e formados divididos entre sexo feminino e masculino.

Após carregar e tratar os dados, foi utilizado o método apriori para mostrar o gráfico de formandos por sexo em cada centro. Pode-se ver essa distribuição a seguir:

No CAL, nota-se cerca de 3.590 formandos, sendo 2.474 do sexo feminino, enquanto os do sexo masculino são apenas 1.116. Dessa forma, cerca de 68% dos formandos desse centro são do sexo feminino, enquanto os do sexo masculino representam aproximadamente 31% do total.

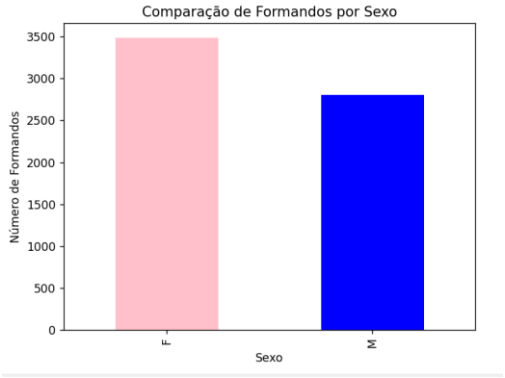
Dados do CAL

No CCNE, nota-se um total de 3.159 formandos, cerca de 1.832 do sexo feminino, e 1.337 do sexo masculino. Cerca de 58% dos formandos são do sexo feminno, e aproximadamente 42 % são do sexo masculino.



Dados do CCNE

No CCR, nota-se um total de 6.091 formandos, cerca de 3.486 do sexo feminino, e 2.805 do sexo masculino. Dessa forma, cerca de 53% dos formandos desse centro são do sexo feminino, e aproximadamente 47 % são do sexo masculino.

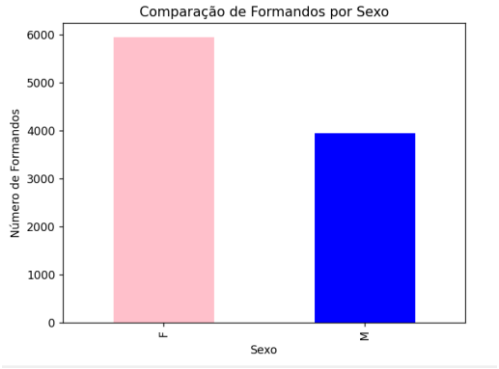


Dados do CCR

No CCS, nota-se um total de 6.940 formandos, cerca de 5.090 do sexo feminino, e 1.850 do sexo masculino. Dessa forma, cerca de 73% dos formandos desse centro são do sexo feminino, e aproximadamente 27 % são do sexo masculino.

Dados do CCS

No CCSH, nota-se um total de 9.870 formandos, cerca de 5.930 do sexo feminino, e 3.940 do sexo masculino. Dessa forma, cerca de 60% dos formandos desse centro são do sexo feminino., e aproximadamente 40 % são do sexo masculino.

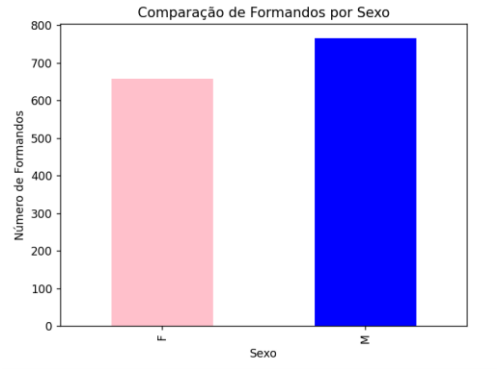


Dados do CCSH

No CE, nota-se um total de 6.810 formandos, cerca de 5.770 do sexo feminino, e 1.040 do sexo masculino. Dessa forma, cerca de 84% dos formandos desse centro são do sexo feminino., e aproximadamente 15% são do sexo masculino.

Dados do CE

No CEFD, nota-se um cenário diferente dos demais centros até agora, ele possui uma quantidade maior de formandos do sexo masculino. O total de formandos do centro é cerce de 1.421, cerca de 766 do sexo masculino, e 655 do sexo feminino. Dessa forma, cerca de 53% dos formandos desse centro são do sexo masculino., e aproximadamente 47% são do sexo feminino.



Dados do CEFD

No CT, assim como no CEFD, o número de formandos do sexo masculino é maior. O total de formandos do centro é cerce de 6.094, cerca de 3.984 do sexo masculino, e 2.110 do sexo feminino. Dessa forma, cerca de 65% dos formandos desse centro são do sexo masculino., e aproximadamente 35% são do sexo feminino.

Dados do CT

O número total de formandos do sexo feminino é 27.347, o que representa cerca de 61% do total de formandos, enquanto o do sexo masculino é 16.838, que representa aproximadamente 39% do total de formandos.

Regras encontradas

Após carregar os dados, as colunas COD_CURSO e NIVEL_CURSO foram descartadas do dataframe, pois o código do curso não tinha utilidade, visto que já tinha o nome do curso e o nível era a maioria de gradução.

No CAL, a melhor regra encontrada, ordenada por confiança, utilizando um suporte mínimo de 0.3 foi que o sexo masculino implicava no ano de 2014.

No CCNE, a melhor regra encontrada, ordenada por confiança, utilizando um suporte mínimo de 0.3 foi que o número de formados igual a 0 implicava no curso de Ciências Biológicas.

No CCR, a melhor regra encontrada, ordenada por confiança, utilizando um suporte mínimo de 0.3 foi que o curso de Agronomia implicava no sexo feminino.

No CCS, a melhor regra encontrada, ordenada por confiança, utilizando um suporte mínimo de 0.3 foi que o curso de Enfermagem implicava no sexo feminino.

No CCSH, a melhor regra encontrada, ordenada por confiança, utilizando um suporte mínimo de 0.3 foi que o ano de 2014 implicava no sexo feminino.

No CE, a melhor regra encontrada, ordenada por confiança, utilizando um suporte mínimo de 0.3 foi o sexo masculino implicava no ano de 2014.

No CEFD, a melhor regra encontrada, ordenada por confiança, utilizando um suporte mínimo de 0.3 foi o ano de 2014 implicava no curso de Dança.

No CT, a melhor regra encontrada, ordenada por confiança, utilizando um suporte mínimo de 0.3 foi o curso de Arquitetura e Urbanismo implicava no sexo feminino

Gráficos Gerados

- **Gráfico de Comparação de Formandos por Sexo**: Mostra claramente a diferença entre a quantidade de formandos do sexo feminino e do sexo masculino.
- **Matriz de Confusão**: Apresenta a acurácia do modelo de árvore de decisão e destaca a distribuição de previsões entre os diferentes sexos.
- **Gráfico de Distribuição de Previsões por Sexo**: Uma visualização adicional que detalha a distribuição das previsões feitas pelo modelo.

Conclusão

O código desenvolvido conseguiu extrair informações valiosas sobre a diferença de formação entre os sexos. A análise revelou uma diferença significativa, onde os formandos do sexo feminino são mais numerosos que os do sexo masculino. Esses insights podem ajudar na tomada de decisões acadêmicas e políticas relacionadas à inclusão e apoio ao desenvolvimento de todos os grupos.

Referências

- **Bibliotecas Python**: pandas, matplotlib, seaborn, scikit-learn, mlxtend
- **Fontes de Dados**: Dados de ingressantes e formandos em formato Excel e CSV