**Applied Statistics – Turma 9DTSR**

Atividade prática aula 07

**Habilidades desenvolvidas: Modelo de Regressão Linear Múltipla. Modelos preditivos usando modelos de regressão logística.**

**Parte 1:**

**Modelo do Bike Sharing:**

**Compare os resultados dos dois modelos: estação do ano como variável preditora qualitativa versus variável preditora quantitativa. Qual a sua conclusão?**



**R: Ao tratar a variável "estação" como categórica, o Rˆ2 da regressão aumenta significativamente (0,34 contra 0,16) em comparação com quando ela é considerada numérica. Isso ocorre porque, ao representar "estação" como qualitativa, utilizamos variáveis dummies, permitindo que o modelo capture diferenças específicas entre as categorias. Já ao tratá-la como numérica, o modelo assume uma relação linear entre os valores atribuídos a cada estação, o que pode não refletir corretamente a realidade dos dados.**

**Na codificação dummy, cada categoria recebe um coeficiente próprio, exceto a primeira, que é removida para evitar colinearidade perfeita. Essa categoria omitida serve como referência, e os coeficientes das demais indicam seu efeito relativo em comparação a ela. Esse método melhora a capacidade explicativa do modelo, resultando em um ajuste mais preciso e um Rˆ2 mais elevado.**

**Parte 2: Regressão Logística**

Utilize os resultados do modelo de inadimplência do cartão de crédito para completar os exercícios a seguir.

1. Use o resultado da Regressão Logística Múltipla para calcular a probabilidade de inadimplência do cartão de crédito:



|  |  |  |
| --- | --- | --- |
| ID | Características | Probabilidade de Inadimplência |
| 1 | Cliente com 2 anos de relacionamento com a Instituição Financeira,, renda mensal de R$4000, reside na região I, com 20 dias de atraso no pagamento, valor da fatura R$2000 e percentual de gastos com alimentação no cartão de 35% | Y= 1.5857+0.7172-1.1061+0 +2.2605 -0.  6676-2.2470  Y= 0.5427  0.63 |
| 2 | Cliente com 1 ano de relacionamento com a Instituição Financeira,, renda mensal de R$1000, reside na região IV, com 20 dias de atraso no pagamento, valor da fatura R$500 e percentual de gastos com alimentação no cartão de 5% | Y = 1.5857 – 0.4814 -1.6490 + 2.2605 + 0.0669  Y = 1,7827  Prob = 0.8564 |
| 3 | Cliente com 9 anos de relacionamento com a Instituição Financeira,, renda mensal de R$7000, reside na região II, com 2 dias de atraso no pagamento, valor da fatura R$4000 e percentual de gastos com alimentação no cartão de 35% |  |

1. Utilize os resultados da matriz de confusão do modelo da Regressão Logística na amostra de treinamento. Preencha o quadro a seguir e responda:

|  |  |  |  |
| --- | --- | --- | --- |
| Resposta real | Resposta do modelo Predito | | Total |
| 0 | 1 |
| 0 |  |  |  |
| 1 |  |  |  |
| Total |  |  |  |

1. Qual o percentual de acerto da resposta 0 (adimplente) ? Apresente o cálculo.
2. Qual o percentual de acerto da resposta 1 (inadimplente) ? Apresente o cálculo.
3. Qual o percentual de acerto médio do modelo? Apresente o cálculo.
4. Utilize os resultados da matriz de confusão do modelo da Regressão Logística na amostra de validação. Preencha o quadro a seguir e responda:

|  |  |  |  |
| --- | --- | --- | --- |
| Resposta real | Resposta do modelo Predito | | Total |
| 0 | 1 |
| 0 |  |  |  |
| 1 |  |  |  |
| Total |  |  |  |

1. Qual o percentual de acerto da resposta 0 (adimplente) ? Apresente o cálculo.
2. Qual o percentual de acerto da resposta 1 (inadimplente) ? Apresente o cálculo.
3. Qual o percentual de acerto médio do modelo? Apresente o cálculo.
4. O modelo de regressão logística ajustado tem poder de generalização? Por quê?

**Parte 3:**

3) Utilizando a base descrita e disponibilizada de vinhos desta região de Portugal com as variáveis de características (composição) dos vinhos.

Contexto dos dados:

**Source:**

Paulo Cortez, University of Minho, Guimarães, Portugal, <http://www3.dsi.uminho.pt/pcortez>   
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009

**Data Set Information:**

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: [[Web Link]](http://www.vinhoverde.pt/en/) or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).   
  
**Relevant Papers:**

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties.   
In Decision Support Systems, Elsevier, 47(4):547-553, 2009.   
  
Available at: [[Web Link]](http://dx.doi.org/10.1016/j.dss.2009.05.016)

|  |  |  |
| --- | --- | --- |
| **Característica** | **Descrição** | |
| * **fixedacidity**  *(Acidez fixa)* | | a maioria dos ácidos presentes no vinho ou fixos ou não voláteis (não evaporaram prontamente) |
| * **volatileacidity** *(Acidez volátil)* | | a quantidade de ácido acético no vinho, que em níveis muito altos pode levar a um gosto desagradável de vinagre. |
| * **Citricacid** (Ácido cítrico) | | encontrado em pequenas quantidades, o ácido cítrico pode adicionar “frescura” e sabor aos vinhos. |
| * **Residualsugar**   *(Açúcar residual)* | | a quantidade de açúcar restante depois que o processo de fermentação (em contato com a levedura, o açúcar da uva vai se transformando em álcool, ou seja, a fermentação vai ocorrendo) |
| * **Chlorides**  *(*Cloretos) | | a quantidade de sal no vinho. |
| * **freesulfurdioxide**   *(Dióxido de enxofre livre)* | | a forma livre de SO2 existe em equilíbrio entre o SO2 molecular (como gás dissolvido) e o íon bissulfito; impede o crescimento microbiano e a oxidação do vinho. Garante condições melhores para os processos de vinificação da bebida, elimina bactérias e leveduras frágeis e indesejáveis, o que permite que apenas as melhores prossigam com o processo fermentativo. Além do mais, melhora o aroma e afina a cor da bebida |
| * **totalsulfurdioxide**   *(Dióxido de enxofre total)* | | quantidade de formas livres e encadernadas de S02; em baixas concentrações, o SO2 é quase indetectável no vinho, mas nas concentrações de SO2 livre acima de 50 ppm, o SO2 se torna evidente no nariz e no sabor do vinho. |
| * **Density**  (*Densidade)* | | a densidade é próxima à da água, dependendo do percentual de álcool e teor de açúcar. |
| * **pH** (*PH)* | | descreve como o vinho é acídico ou básico numa escala de 0 (muito ácido) a 14 (muito básico); a maioria dos vinhos tem entre 3-4 na escala de pH. |
| * **Sulphates** *(Sulfatos)* | | um aditivo de vinho que pode contribuir para os níveis de gás de dióxido de enxofre (S02), que age como um antimicrobiano e antioxidante. |
| * **alcohol** *(Álcool)* | | o teor alcoólico percentual do vinho. |
| * **quality** *(Qualidade)* | | variável de saída (com base em dados sensoriais) que poderiam ser de 0 a 10 sendo zero muito ruim e 10 muito excelente. |

Exemplo da Base (10 linhas)



**Etapa 1**:

Com objetivo de estimar a nota de qualidade dos vinhos (variável “Quality”) em função das características físico-químicas dos vinhos;

1. Qual é a técnica adequada para essa resolver esse objetivo:

|  |  |
| --- | --- |
| X | Técnicas Supervisionadas |
|  | Técnicas Não Supervisionadas |

1. Há variáveis dependentes (D) e independentes (I)? Caso sim, preencha o quadro a seguir.

|  |  |
| --- | --- |
| Variáveis Dependentes: | quality |
| Variáveis Independentes: | Todas, exceto id\_vinho e quality |

1. Qual (is) a tarefa(s) de Mining é (são) adequada (s) para solução desse objetivo:

|  |  |
| --- | --- |
| X | Predição |
|  | Classificação |
|  | Regras de Associação |
|  | Agrupamento |
|  | Sumarização |
|  | Detecção de Anomalias |

1. Você pode indicar alguns algoritmos possíveis para essa tarefa?

|  |  |
| --- | --- |
| X | Regressão Linear |
|  | Árvore de Decisão |
|  | Regras de Associação |
|  | Séries Temporais |
|  | Análise de Clusters |
|  | Componentes Principais |
| X | Redes Neurais |
| X | Random Forest, |
|  | Outras: Árvore de Regressão, SVM |

1. Quais cuidados que são indicados a fazer antes da etapa de aplicar os algoritmos/técnicas?

**R. Realizar uma análise exploratória dos dados (distribuição e missing); padronização/normalização das variáveis quantitativas; correlação de Pearson, encode das variáveis; variáveis dummies; feature selection; mudar o formato da variável numérica (quali) para texto;**

**Etapa 2:**

Com o objetivo de classificar os vinhos dessa região em vinhos em “Bom” ou “Ruim” em função de suas características físico-químicas;

1. Qual é a técnica adequada para essa resolver esse objetivo:

|  |  |
| --- | --- |
| X | Técnicas Supervisionadas |
|  | Técnicas Não Supervisionadas |

1. Há variáveis dependentes (D) e independentes (I)?

|  |  |
| --- | --- |
| Variáveis Dependentes: | Bom (quality > 7) e Ruim (quality <= 7) |
| Variáveis Independentes: | Todas, exceto id\_vinho e quality |

1. Qual (is) a tarefa(s) de Mining é (são) adequada (s) para solução desse objetivo:

|  |  |
| --- | --- |
|  | Predição |
| X | Classificação |
|  | Regras de Associação |
|  | Agrupamento |
|  | Sumarização |
|  | Detecção de Anomalias |

1. Você pode indicar alguns algoritmos possíveis para essa tarefa?

|  |  |
| --- | --- |
|  | Regressão Linear |
| X | Árvore de Decisão |
|  | Regras de Associação |
|  | Séries Temporais |
|  | Análise de Clusters |
|  | Componentes Principais |
| X | Redes Neurais |
| x | Random Forest, |
| x | Outras: Regressão Logística, Knn, SVM |

1. Quais cuidados que são indicados a fazer antes da etapa de aplicar os algoritmos/técnicas?

**R. Criar a variável resposta do modelo; realizar uma análise exploratória dos dados (distribuição e missing); padronização/normalização das variáveis quantitativas; correlação de Pearson, encode das variáveis; variáveis dummies; feature selection; mudar o formato da variável numérica (quali) para texto;**

Entrega do exercício no formato word.

Data de entrega: 20/02/2025

Regina Bernal

10/02/2025