TÉCNICAS COLOCADAS EM PRÁTICA

- **TÉCNICAS BÁSICAS**

1. **Imputação**: Preenchimento de valores faltantes em um conjunto de dados é feito por meio da técnica de imputação.

2. Tratamento de outliers: A técnica de tratamento de valores extremos que podem afetar uma análise é conhecida como tratamento de outliers.

3. Binarização: A conversão de variáveis numéricas em binárias, com base em um limite pré-definido, é denominada binarização.

4. Transformação de log: Técnica utilizada para tornar a distribuição de uma variável assimétrica mais próxima da normal.

5. One-Hot Encoding: Técnica de transformação de variáveis categóricas em binárias, para que possam ser utilizadas em modelos de aprendizado de máquina.

6. Operações de Agrupamento: O agrupamento de dados com base em um ou mais critérios, visando a uma melhor compreensão das informações, é realizado por meio de operações de agrupamento.

7. Split (divisão) de Atributos: A técnica de divisão de uma variável em várias outras menores, com base em algum critério, é chamada de split ou divisão de atributos.

8. Scaling (Padronização): Padronização é a técnica utilizada para ajustar a escala de uma variável, de modo que seus valores fiquem dentro de um determinado intervalo.

9. Extração de Data: A extração de informações de data e hora, como dia da semana, mês, ano e hora do dia, é realizada por meio da técnica de extração de data.

10. Categorização: Técnica de agrupar dados em categorias, com base em um critério pré-definido, para análise de tendências ou padrões é chamada de categorização.

- **OUTRAS TÉCNICAS:**

1. Seleção Univariada de Atributos: A técnica de Seleção Univariada de Atributos analisa cada atributo individualmente e escolhe aqueles com maior correlação com a variável de interesse.

2. Principle Components Analysis (PCA): Técnica de redução de dimensionalidade que transforma um conjunto de variáveis em componentes principais não correlacionados, preservando a maior quantidade possível de informação.

3. Independent Component Analysis (ICA): É uma técnica de separação de fontes que identifica componentes independentes de sinais mistos, sem conhecimento prévio sobre as fontes.

4. Linear Discriminant Analysis (LDA): Busca encontrar uma combinação linear de atributos que maximize a separação entre as classes de um conjunto de dados, reduzindo a dimensionalidade.

5. Locally Linear Embedding (LLE): Técnica que preserva a geometria local dos dados, buscando um conjunto de pontos de referência para representar os dados de forma mais compacta.

6. t-distributed Stochastic Neighbor Embedding (t-SNE): Técnica de visualização de dados que preserva relações de proximidade entre pontos, transformando um conjunto de atributos em um espaço de baixa dimensionalidade.

7. Autoencoders: São redes neurais que representam dados em um espaço de baixa dimensionalidade, minimizando a diferença entre dados originais e reconstruídos. São usados para redução de dimensionalidade e detecção de anomalias, entre outras aplicações.

SOBRE O PROJETO

O objetivo desse projeto é apresentar insights sobre as reinternações hospitalares, que ocorrem quando um paciente recebe alta e é internado novamente em um curto espaço de tempo, são dispendiosas e indicam falhas no sistema de saúde.

Estatísticas importantes
Nos Estados Unidos, só o tratamento de pacientes diabéticos readmitidos ultrapassa os 300 milhões de dólares por ano.

A identificação precoce de pacientes com alto risco de readmissão permite aos profissionais de saúde conduzirem investigações adicionais e possivelmente evitarem futuras reinternações. Isso não apenas aprimora a qualidade do atendimento, mas também reduz as despesas médicas relacionadas às reinternações.

O diabetes é a sétima principal causa de morte no mundo (dados de 2016, fonte ao final do texto) e afeta aproximadamente 23,6 milhões de pessoas só nos EUA. Milhões de pessoas são diagnosticadas com diabetes em todo o mundo a cada ano. Segundo a Associação Americana de Diabetes, os custos de atendimento a pacientes diabéticos e pré-diabéticos nos Estados Unidos são os maiores do mundo.

Essa epidemia global afeta mais de 350 milhões de pessoas, com 3 milhões de pessoas morrendo a cada ano devido a complicações relacionadas ao diabetes, principalmente cardiovasculares ou nefropáticas.

A reinternação hospitalar é uma das principais preocupações no tratamento do diabetes, já que custa milhões de dólares no tratamento de pacientes diabéticos que precisam ser reinternados após receberem alta. A necessidade de reinternação indica que houve cuidados inadequados no momento da primeira internação. A taxa de reinternação tornou-se uma métrica importante para avaliar a qualidade geral de um hospital.

No Brasil
De acordo com uma pesquisa publicada em 2018 no periódico científico Revista de Saúde Pública, a taxa de readmissões hospitalares em um hospital universitário de São Paulo foi de 19,2%. Esse estudo sugere que as principais causas de readmissão foram complicações relacionadas a procedimentos cirúrgicos, doenças cardiovasculares e infecções.

A importância sobre o alto nível de readmissão hospitalar

Identificar pacientes diabéticos com alto risco de readmissão é uma tarefa crucial para melhorar a qualidade geral do cuidado de saúde. A necessidade de readmissão é um indicador de que o paciente não recebeu os cuidados adequados durante a primeira

internação. Por essa razão, a taxa de readmissão tornou-se uma métrica importante para avaliar a qualidade do serviço oferecido pelos hospitais.

Como principal analista de dados de uma organização de saúde, seu trabalho é identificar esses pacientes de alto risco por meio de registros médicos eletrônicos, utilizando informações como resultados de exames, níveis de insulina, diagnósticos de outras doenças, entre outros dados relevantes. Durante o processo, serão necessárias diversas técnicas de engenharia de atributos, as quais serão justificadas e detalhadas ao longo das aulas. Ao final do projeto, os resultados da análise serão apresentados em diversos gráficos para que se possa compreender melhor o impacto das estratégias adotadas.

IMPORTANTE: *o projeto está dividido em 2 PARTES, sendo a primeira parte destinada ao tratamento dos dados e a segunda parte, aos possíveis tomadores de decisão, sendo mais visual.*

SOBRE OS DADOS UTILIZADOS

Para este projeto, utilizaremos o conjunto de dados "Diabetes 130-US hospitals for years 1999-2008", disponível no UCI Machine Learning Repository.

DATASET (link seguro)

O conjunto de dados contém informações sobre atendimento clínico em 130 hospitais dos EUA e redes integradas, com 100.000 observações e 50 recursos, incluindo registros eletrônicos com resultados de exames dos pacientes e dados sobre cada hospital.

Os autores da coleta dos dados destacaram a importância do controle da hiperglicemia em pacientes hospitalizados, uma vez que isso pode influenciar significativamente nos resultados e na mortalidade do paciente.

A análise do banco de dados clínicos revelou que apenas 18,4% das medições da HbA1c foram realizadas durante o atendimento hospitalar. A regressão logística multivariável foi utilizada para analisar a relação entre a medida da HbA1c e a readmissão precoce, ajustando-se para variáveis demográficas, gravidade da doença e tipo de admissão.

Os resultados sugerem que a relação entre a probabilidade de readmissão e a medição da HbA1c depende do diagnóstico primário. A atenção ao diabetes refletida na determinação da HbA1c pode melhorar os resultados dos pacientes e reduzir o custo dos cuidados hospitalares.

Em resumo, a análise dos dados do conjunto pode fornecer orientações úteis para melhorar a segurança do paciente durante a hospitalização e a qualidade geral do atendimento.

DICIONÁRIO DE DADOS

O Dicionário de Dados é uma descrição detalhada de cada variável presente no nosso conjunto de dados. Aqui está o dicionário do dataset utilizado no Projeto 8:

0- **encounter_id** - identificador único de um encontro do pesquisador com o paciente
1- **patient_nbr** - identificador exclusivo de um paciente
2- **race** - valores: Caucasian, Asian, African American, Hispanic e other
3- **gender** - valores: male, female, and unknown/invalid
4- **age** - agrupados em intervalos de 10 anos: (0-10), (10-20), ...
5- **weight** - peso em libras
6- **admission_type_id** - identificador inteiro correspondente a 9 valores distintos, por exemplo, "emergência, urgência, eletiva, recém-nascido e não disponível"
7- **discharge_disposition_id** - identificador inteiro correspondente a 29 valores distintos, por exemplo, "enviado para casa, expirou e não está disponível"
8- **admission_source_id** - identificador inteiro correspondente a 21 valores distintos, por exemplo, "encaminhamento médico, e transferência de um hospital"
9- **time_in_hospital** - número inteiro de dias entre a admissão e a alta
10- **payer_code** - identificador inteiro correspondente a 23 valores distintos.por exemplo: Blue Cross / Blue Shield, Medicare e auto-pagamento"
11- **medical_specialty** - identificador inteiro de uma especialidade do médico admitidor, correspondente a 84 valores distintos, por exemplo, cardiologia, medicina interna, família / clínica geral e cirurgião"
12- **num_lab_procedures** - número de testes de laboratório realizados durante a consulta
13- **num_procedures** - número de procedimentos (exceto testes de laboratório) realizados durante a consulta
14- **num_medications** - número de medicamentos genéricos distintos administrados durante a consulta
15- **number_outpatient** - número de consultas ambulatoriais do paciente no ano anterior a consulta
16- **number_emergency** - número de visitas de emergência do paciente no ano anterior a consulta
17- **number_inpatient** - número de visitas hospitalares do paciente no ano anterior a consulta
18- **diag_1** - diagnóstico primário (codificado como três primeiros dígitos da CID9); 848 valores distintos

19- **diag_2** - diagnóstico secundário (codificado como três primeiros dígitos da CID9); 923 valores distintos
20- **diag_3** - diagnóstico secundário adicional (codificado como três primeiros dígitos da CID9); 954 valores distintos
21- **number_diagnoses** - número de diagnósticos inseridos no sistema
22- **max_glu_serum** - teste sérico de glicose que indica a faixa do resultado ou se o teste não foi realizado.

- Valores: > 200, > 300, normal e nenhum, se não for medido

23- **A1Cresult** - teste A1c que indica o intervalo do resultado ou se o teste não foi realizado.

- Valores:
  > 8 (se o resultado for maior que 8%),
  > 7 (se o resultado for maior que 7%, porém menor que 8%),
  normal (se o resultado for inferior a 7%) e nenhum, se não for medido

Aqui apresentamos os recursos de 24 a 46 para os nomes dos medicamentos genéricos: *metformina; repaglinida; nateglinida; clorpropamida; glimepirida; acetohexamida; glipizida; gliburida; tolbutamida; pioglitazona; rosiglitazona; acarbose; miglitol; troglitazona; insulazforamida; examide; sitaglagliptida; sitazagliptida; glipizida-metformina; glimepirida-pioglitazona; metformina-rosiglitazona e metformina-pioglitazona.*

Os recursos acima indicam se o medicamento foi **prescrito** ou se houve uma **alteração na dosagem** *(valores: "up" se a dose foi aumentada durante a consulta)*

47- **change** - indica se houve alteração nos medicamentos para diabéticos (dosagem ou nome genérico). Valores: "change" e "no change".
48- **diabetesMed** - indica se houve algum medicamento diabético prescrito. Valores: "sim" e "não".

49- **readmitted** - readmitido, "Dias para readmissão hospitalar.

- Valores:
  <30 (se o paciente foi readmitido em menos de 30 dias)
  > 30 (se o paciente foi readmitido em mais de 30 dias)
  No, para nenhum registro de readmissão.

Nota: Códigos ICD-9 ou CID-9 (International Classification of Diseases ou Código Internacional de Doenças).

AUTOR DO CÓDIGO:
Jonatas A. Liberato