UNIVERSIDADE FEDERAL DE PERNAMBUCO

Centro de Informática
UFPE

Universidade Federal de Pernambuco

Centro de Informática - CIn

Predição de Readmissão Hospitalar para Pacientes com Diabetes
Data Understanding

Grupo:

Matheus Henrique
Gilberto Medeiros
Vinícius Barbosa

Walmir Bispo

Professor:

Leandro Maciel Almeida

Abril de 2023

Conteúdo

1 Introdução

Em um projeto de Data Science, a Análise Exploratória de Dados (EDA) é a abordagem utilizada para analisar um conjuntos de dados com intuito de resumir suas principais características, geralmente com métodos visuais. A EDA é usada para ver o que os dados podem nos dizer antes da tarefa de modelagem. Não é fácil olhar para uma coluna de números ou para uma planilha inteira e determinar características importantes dos dados. Pode ser bastante desafiador obter insights observando apenas valores puramente numéricos, por isso técnicas de análise exploratória de dados foram concebidas como uma ajuda nestas situações.

A EDA realmente revela a verdade sobre o conteúdo sem fazer nenhuma suposição intuitiva. Devido a isto, cientistas de dados usam esse processo para realmente entender que tipo de modelagem e hipóteses são as mais adequadas ao problema proposto. Os principais componentes da análise exploratória de dados incluem resumo de dados, análise estatística e visualização de dados. Neste projeto utilizou-se Python e suas ferramentas especializadas para análise exploratória, como os pacotes pandas, scipy, matplotlib e plotly.

2 Data Understanding

2.1 Visão geral dos dados

O Diabetes 130-US Hospitals dataset é uma coleção de dados médicos contendo informações de mais de 100.000 pacientes com diabetes que foram internados em 130 hospitais nos Estados Unidos entre 1999 e 2008. O objetivo deste conjunto de dados é ajudar os pesquisadores a entender os fatores que levam a reinternações hospitalares de pacientes com diabetes. O conjunto de dados é composto por cerca de 50 colunas, incluindo informações sobre a idade, gênero, etnia, tipo de diabetes, resultado de exames laboratoriais, prescrições de medicamentos e muito mais.

De forma simples, os modelos têm um desempenho tão bom quanto os dados que alimentamos neles. Então, em primeiro lugar, vamos carregar o dataset e examinar algumas das colunas.

	encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	...	citoglipton	insulin	glyburide-metformin	glipizide-metformin	glimepiride-pioglitazone
0	2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	...	No	No	No	No	No
1	149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3	...	No	Up	No	No	No
2	64410	86047875	AfricanAmerican	Female	[20-30)	?	1	1	7	2	...	No	No	No	No	No
3	500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2	...	No	Up	No	No	No
4	16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1	...	No	Steady	No	No	No

5 rows × 50 columns

Figura 1: Primeiras cinco linhas do conjunto de dados Diabetes 130-US Hospitals

Olhando rapidamente pelas colunas, podemos ver que há algumas colunas de identificação, algumas colunas numéricas e algumas colunas categóricas. Essas colunas estão descritas com mais detalhes nas tabelas 1 e 2. As features "admission_type_id", "discharge_disposition_id"e "admission_souce_id"(estão marcadas com o caractere "*"na tabela 1) possuem uma descrição adicional, a qual está presente na tabela 2. Ambas foram construídas baseadas no artigo [1].

2.2 Dados Faltantes

O segundo passo consiste em lidar com os valores ausentes. No caso específico deste conjunto de dados, os valores ausentes são catalogados com o caractere “?”, que não é um formato padrão de codificação de valores ausentes, portanto é necessário corrigir este detalhe antes da análise.

A figura 2 exibe a matriz de nulidade (ferramenta gráfica que permite visualizar e identificar a existência de padrões em valores ausentes nos dados). De cara percebemos que a variável *weight* é composta quase em sua totalidade por valores nulos, então optou-se por removê-lo. Como regra geral, as variáveis com 50% ou mais valores ausentes devem ser descartadas da análise. A variável *medical_specialty* possui 49% das observações faltando. Em termos de proporção, toda a coluna deve ser descartada. No entanto, com base na recomendação de pesquisas anteriores, essa variável é fundamental na previsão de reinternação. Portanto, os valores ausentes deverão ser codificados em uma nova categoria chamada "Missing". Vale salientar também que o status socioeconômico do paciente é um fator crítico na previsão de reinternações, portanto, variáveis como *payer_code* devem ser preservadas no conjunto de dados.

Para o restante das variáveis com taxa de falta baixa a média, a imputação deverá ser realizada para manter o máximo de dados possível para modelagem.

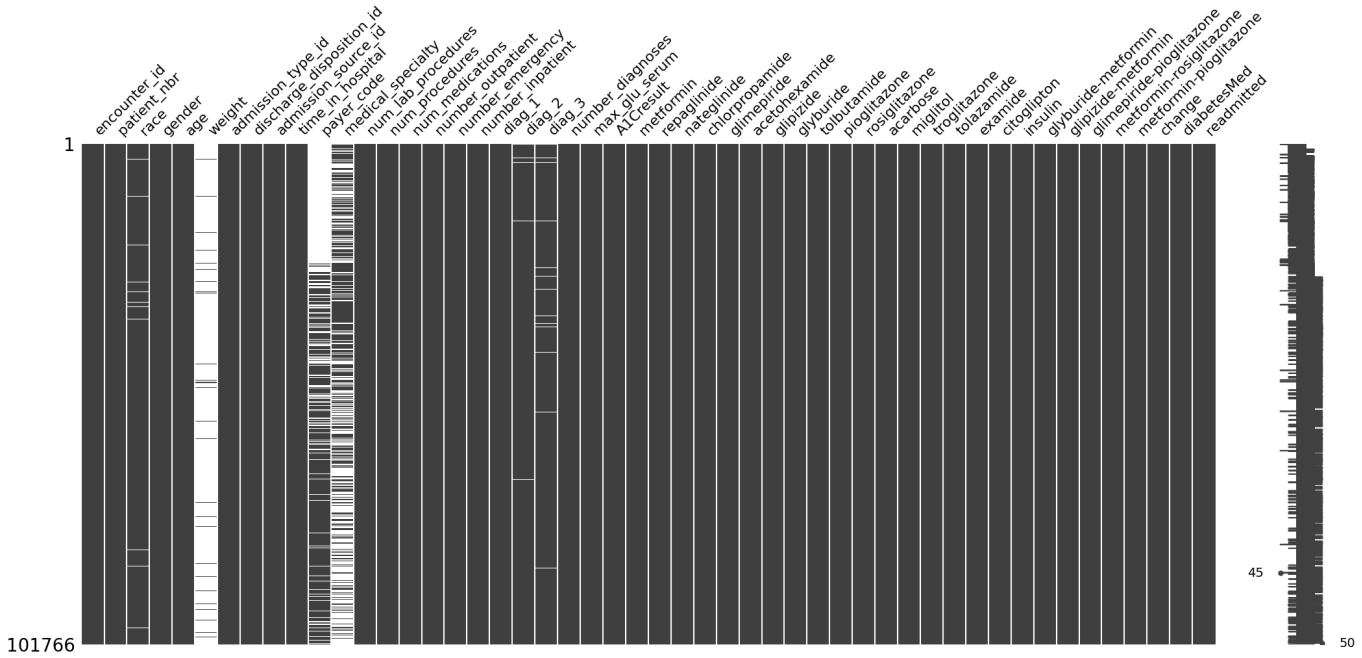

Figura 2: Matriz de dados faltantes

2.3 Análise de Features

A partir de uma análise mais aprofundada da tabela 1, percebe-se que há uma mistura de dados categóricos e numéricos. Algumas coisas merecem serem destacadas:

- *encounter_id* e *patient_nbr*: São apenas identificadores e não são variáveis úteis.
- *age* e *weight*: São categóricos neste conjunto de dados.
- *patient_nbr*: De acordo com alguns pesquisadores, esse conjunto de dados apresenta algumas inconsistências. Por exemplo, vários pacientes tiveram múltiplas internações e não devem ser tratados como algo independente, pois isso influenciaria outras observações. A fim de garantir um identificador inerentemente único, é sugerido manter apenas o primeiro encontro quando um paciente tiver vários registros.
- *admissão_type_id*, *discharge_disposition_id*, *admission_source_id*: São numéricos, todavia são IDs (consulte tabela 2). Eles devem ser tratados como categóricos.

Note: *discharge_disposition_id* refere-se à localização ou estado da pessoa após a internação. Se olharmos para a tabela 2, notaremos que 11,13,14,19,20,21 estão relacionados a morte ou hospício. Obviamente, pacientes que morreram durante a internação não têm probabilidade de serem reinternados e, portanto, devem ser excluídos da análise. Pelo mesmo motivo, também serão omitidos doentes com alta para hospício.

- *examide* e *citoglipton*: Possuem apenas 1 valor (figura 8, 7), então não usaremos essas variáveis.
- *medical_speciality*, *diag_1* , *diag_2*, *diag_3*: São features categóricas com altíssima cardinalidade (figura 6, 3, 4, 5). Estas variáveis compreendem uma enorme quantidade de valores únicos tornando extremamente difícil seu uso em modelos de machine learning. Portanto, iremos considerar criar algum tipo de agregação para redução de dimensionalidade.
- *24 features for medication*: O dataset inclui 24 variáveis relacionadas com a medicação, cada uma associada a 4 classes (“No”, “Steady”, “Up” e “Down”). Tais categorias visam avaliar se ocorreu alguma mudança de medicamento durante a internação do paciente. Diversos pesquisadores destacaram mudanças de medicação como um fator influente para reinternações. Um número total maior de medicamentos também pode ser um indicador da severidade da condição do paciente. Portanto, aqui tem-se margem para posterior elaboração de novas features.
- *Readmitted*: Reduzir classe de saída para binário: O principal objetivo do projeto é prever se um paciente será readmitido ou não nos próximos 30 dias após a alta. Portanto, estudo é limitado à diferenciar duas classes «30"e "NO". No entanto, como exibido na figura 9 o conjunto de dados compreende 3 classes, incluindo reinternações abaixo de 30 dias (11,2%), acima de 30 dias (34,9%) e não-internações (53,9%). Descartar reinternações após 30 dias resultaria na perda de um terço de todas as observações, logo, as reinternações ocorridas após 30 dias serão consideradas como não-reinternações.

  Note: A variável target após esta pequena agregação certamente ficará desbalanceada (pouca incidência da classe «30"). Isso deve ser tratado com técnicas de sampling, pois pode alterar a generalização do modelo.

2.4 Figuras Adicionais

	value	count	frequency (%)
0	428	6862	6.742920
1	414	6581	6.466796
2	786	4016	3.946308
3	410	3614	3.551284
4	486	3508	3.447124
...	...	...	...
711	817	1	0.000983
712	61	1	0.000983
713	148	1	0.000983
714	870	1	0.000983
715	V51	1	0.000983

716 rows × 3 columns

Figura 3: Frequência da variável diag_1 (alta cardinalidade)

	value	count	frequency (%)
0	276	6752	6.634829
1	428	6662	6.546391
2	250	6071	5.965647
3	427	5036	4.948608
4	401	3736	3.671167
...	...	...	...
743	232	1	0.000983
744	908	1	0.000983
745	52	1	0.000983
746	E817	1	0.000983
747	927	1	0.000983

748 rows × 3 columns

Figura 4: Frequência da variável diag_2 (alta cardinalidade)

	value	count	frequency (%)
0	250	11555	11.354480
1	401	8289	8.145157
2	276	5175	5.085195
3	428	4577	4.497573
4	427	3955	3.886367
...	...	...	...
784	657	1	0.000983
785	684	1	0.000983
786	603	1	0.000983
787	E826	1	0.000983
788	971	1	0.000983

789 rows × 3 columns

Figura 5: Frequência da variável diag_3 (alta cardinalidade)

	value	count	frequency (%)
0	InternalMedicine	14635	14.381031
1	Emergency/Trauma	7565	7.433720
2	Family/GeneralPractice	7440	7.310890
3	Cardiology	5352	5.259124
4	Surgery-General	3099	3.045221
...	...	...	...
67	Perinatology	1	0.000983
68	Neurophysiology	1	0.000983
69	Psychiatry-Addictive	1	0.000983
70	Pediatrics-InfectiousDiseases	1	0.000983
71	Surgery-PlasticwithinHeadandNeck	1	0.000983

Figura 6: Frequência da variável medical_specialty (alta cardinalidade)

	value	count	frequency (%)
0	No	101766	100.0

Figura 7: Frequência da variável examide (valor constante)

	value	count	frequency (%)
0	No	101766	100.0

Figura 8: Frequência da variável citoglipton (valor constante)

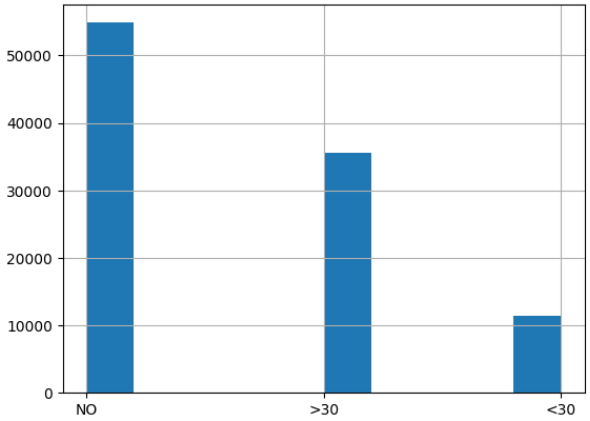

Figura 9: Distribuição da variável target

2.5 Tabelas

Tabela 1: Nomes e descrições das features do dataset Diabetes US-130 Hospitals

Nome da Feature	**Descrição**
encounter_id	Identificador único da internação hospitalar
patient_nbr	Identificador único do paciente
race	Raça do paciente (Caucasian, Asian, African American, Hispanic, and other)
gender	Gênero do paciente (male, female, and unknown/invalid)
age	Idade do paciente (Agrupado em intervalos de 10 anos: [0, 10), [10, 20) ...))
weight	Peso do paciente
admission_type_id*	Identificador numérico do tipo de internação
discharge_disposition_id*	Identificação do tipo de alta hospitalar
admission_source_id*	Identificação da fonte da internação
time_in_hospital	Tempo entre internação e alta hospitalar (em dias)
payer_code	Código do pagador do paciente. Por exemplo: Blue Cross/Blue Shield, Medicare, and self-pay Medical, etc.
medical_specialty	Especialidade do médico responsável ao paciente. Por exemplo: cardiologia, medicina interna, clínico geral, cirurgião, etc
num_lab_procedures	Número de testes laboratoriais realizados durante a internação
num_procedures	Número total de procedimentos médicos (exceto os laboratoriais) realizados durante a internação
num_medications	Número total de medicamentos prescritos durante a internação
number_outpatient	Número de visitas do paciente nos últimos 12 meses
number_emergency	Número de visitas de emergência do paciente nos últimos 12 meses
number_inpatient	Número de consultas hospitalares do paciente nos últimos 12 meses
diag_1	Código do diagnóstico principal (codificado como os três primeiros dígitos do ICD9)
diag_2	Código do segundo diagnóstico
diag_3	Código do diagnóstico secundário adicional
number_diagnoses	Número total de diagnósticos registrados
max_glu_serum	Resultado do teste de glicose mais recente. Indica o intervalo do resultado ou se o teste não foi feito. Valores: “>200,” “>300,” “normal,” e “none” se não medido
A1Cresult	Resultado do teste de hemoglobina A1C mais recente. Indica o intervalo do resultado ou se o teste não foi feito. Valores: “>8”, “>7”, “normal” se o resultado for inferior a 7% e “nenhum” se não medido.
Diabetes medications	Indica se foi prescrito algum medicamento para diabéticos. Valores: “yes” e “no”
24 features for medications	Para os nomes genéricos: metformina, sitagliptina, insulina, etc, a característica indica se o medicamento foi prescrito ou houve alteração na posologia. Valores: “up” se a dosagem foi aumentada durante o encontro, “down” se a dosagem foi diminuída, “steady” se a dosagem não mudou e “no” se o medicamento não foi prescrito
change	Indica se houve alteração nos medicamentos para diabéticos (seja posologia ou nome genérico). Valores: “change” e “no change”
Readmitted	Dias para reinternação do paciente. Valores: “<30” se o paciente foi reinternado em menos de 30 dias, “>30” se o paciente foi reinternado em mais de 30 dias e “No” para nenhum registro de reinternação.

Tabela 2: Descriçoes de valores IDs

ID	**Descrição de Valores**
admission_type_id	1: Emergência, 2: Urgente 3: Eletivo, 4: Recém-Nascido, 5: Não Disponível, 6: NULL, 7: Centro de Trauma, 8: Não Mapeado
discharge_disposition_id	1: Alta para casa, 2: Alta/transferida para outro hospital de curta permanência, 3: Alta/transferida para SNF, 4: Alta/transferida para ICF, 5: Alta/transferida para outro tipo de instituição de internação, 6: Alta/transferida para outro tipo de instituição de internação, 6: Alta/transferida para outro tipo de instituição de internação transferido para casa com serviço de saúde domiciliar, 7: Esquerda AMA, 8: Alta/transferido para casa sob cuidado do provedor de IV domiciliar, 9: Admitido como paciente internado neste hospital, 10: Recém-nascido liberado para outro hospital para pós-tratamento neonatal, 11: Expirado, 12: Ainda paciente ou com previsão de retornar para serviços ambulatoriais, 13: Hospício/domicílio, 14: Hospício/instalação médica, 15: Alta/transferida dentro desta instituição para leito oscilante aprovado pelo Medicare, 16: Alta/transferida/encaminhada para outra instituição para serviços ambulatoriais, 17: Alta/transferida/encaminhada para esta instituição para serviços ambulatoriais, 18: NULL, 19: "Expirou em casa. Somente Medicaid, hospício.", 20: "Expirou em uma instalação médica. Somente Medicaid, hospício. ", 21: "Expirado, local desconhecido. Somente Medicaid, hospício.", 22: Alta/transferida para outra unidade de reabilitação, incluindo unidades de reabilitação de um hospital, 23: Alta/-transferida para um hospital de cuidados prolongados, 24: Alta/-transferida para uma instalação de enfermagem certificada pelo Medicaid, mas não certificada pelo Medicare, 25: Não mapeado, 26: Desconhecido/inválido, 30: Alta/transferida para outro tipo de instituição de saúde não definida em outro lugar, 27: Alta/-transferida para um serviço de saúde federal estabelecimento, 28: Alta/transferida/encaminhada para um hospital psiquiátrico de unidade psiquiátrica distinta de um hospital, 29: Alta/transferida para um Hospital de Acesso Crítico (CAH)
admission_source_id	1: Encaminhamento médico, 2: Encaminhamento clínico, 3: Encaminhamento HMO, 4: Transferência de um hospital, 5: Transferência de uma unidade de enfermagem especializada (SNF), 6: Transferência de outra unidade de saúde, 7: Sala de emergência, 8: Tribunal/aplicação da lei, 9: não disponível, 10: transferência de hospital de acesso crítico, 11: parto normal, 12: parto prematuro, 13: bebê doente, 14: parto extramural, 15: não disponível, 17: nulo, 18: transferência De outra agência de saúde domiciliar, 19: Readmissão para a mesma agência de saúde domiciliar, 20: Não mapeado, 21: Desconhecido/inválido, 22: Transferência de hospital interno/mesmo resultado em um pedido de setembro, 23: Nascido neste hospital, 24: Nascido fora deste hospital, 25: Transferência do Centro de Cirurgia Ambulatorial, 26: Transferência do Hospice

Referências

[1] B. Strack, J. P. DeShazo, C. Gennings, J. L. Olmo, S. Ventura, K. J. Cios, and J. N. Clore, "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70, 000 clinical database patient records," *BioMed Research International*, vol. 2014, pp. 1–11, 2014.

[2] "UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set — archive.ics.uci.edu." `https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008`. [Accessed 05-Apr-2023].

[3] "Statistics About Diabetes | ADA — diabetes.org." `https://diabetes.org/about-us/statistics/about-diabetes`. [Accessed 05-Apr-2023].