Universidade Estadual de Campinas
Instituto de Computação

Ciência e Visualização de Dados na Saúde – MO413

Professor André Santanchè

Análise de Neutropenia
Prática 03 – 1º Semestre de 2023

Maurício Pereira Lopes – RA 225242

---

1 – Introdução

Este relatório foi gerado no contexto da disciplina *"Ciência e Visualização de Dados na Saúde – MO413"* da pós-graduação da Unicamp como resultado de uma atividade prática.

O objetivo geral do projeto é, para pacientes com neutropenia, avaliar a associação entre características clínicas e laboratoriais com desfechos: (i) morte, (ii) número de dias de internação.

A análise deve separar grupos diagnosticados com câncer versus aqueles que não têm diagnóstico de câncer.

Para este estudo, foram preparadas tabelas do MIMIC. Só terão acesso às mesmas aqueles que ganharam autorização de acesso.

A análise deve partir de tabelas do MIMIC preparadas para este estudo, que partem de admissões de pacientes em dois contextos relacionados com neutropenia:

- admissões de pacientes diagnosticados com neutropenia;
- admissões de pacientes cujo exame laboratorial indica neutropenia.

Os dados foram extraídos da base de dados MIMIC e fornecidos em quatro tabelas:

- *neutro_admissions.csv*
- *neutro_diagnoses.csv*
- *neutro_itemid_loinc.csv*
- *neutro_labevents_min.csv*

2 – Análises

A tabela *neutro_admissions.csv* fornece dados sobre as admissões de pacientes identificados com Neutropenia. São informações sobre o paciente, datas de entrada e saída, e data da morte, se esta ocorreu.

Há pacientes com mais de uma admissão em datas diferentes e não há informações sobre as condições clínicas deles.

Tabela 1 - Visão parcial da tabela neutro_admissions.csv

	subject_id	hadm_id	admittime	dischtime	deathtime	admission_type	marital_status	race
0	10003019	25179393	2175-12-08 17:38:00	2175-12-11 17:20:00	\N	EW EMER.	MARRIED	WHITE
1	10006431	28771670	2128-03-23 17:06:00	2128-03-30 15:00:00	\N	EW EMER.	MARRIED	WHITE
2	10010231	20687038	2118-03-07 17:27:00	2118-03-09 14:05:00	\N	OBSERVATION ADMIT	SINGLE	HISPANIC/LATINO - GUATEMALAN
3	10010231	23083980	2118-01-31 17:58:00	2118-02-07 11:30:00	\N	OBSERVATION ADMIT	SINGLE	HISPANIC/LATINO - GUATEMALAN
4	10010231	25499227	2117-11-08 20:21:00	2117-12-05 19:40:00	\N	EW EMER.	SINGLE	HISPANIC/LATINO - GUATEMALAN
...	...	...	...	...	...	...	...	...
4244	19992875	22729360	2161-05-09 23:30:00	2161-05-16 13:30:00	\N	EW EMER.	DIVORCED	WHITE
4245	19992875	22872788	2161-12-23 15:27:00	2161-12-25 17:45:00	\N	URGENT	DIVORCED	WHITE
4246	19994730	25836955	2169-04-17 21:33:00	2169-04-21 14:20:00	\N	EW EMER.	MARRIED	WHITE
4247	19995127	29614352	2138-03-30 13:37:00	2138-04-05 16:31:00	\N	EW EMER.	MARRIED	BLACK/AFRICAN AMERICAN
4248	19998562	26846592	2166-04-06 20:38:00	2166-04-16 16:20:00	\N	EW EMER.	MARRIED	WHITE

4249 rows × 8 columns

A segunda tabela fornecida é a *neutro_diagnoses.csv* que fornece informações sobre vários diagnósticos feitos aos pacientes em cada admissão. Diversos diagnósticos diferentes existem para cada paciente.

Tabela 2 - Visão parcial da tabela neutro_diagnoses.csv

	subject_id	hadm_id	seq_num	icd_code	icd_version	long_title
0	10003019	25179393	1	28800	9	Neutropenia, unspecified
1	10003019	25179393	2	20190	9	Hodgkin's disease, unspecified type, unspecifi...
2	10003019	25179393	3	2853	9	Antineoplastic chemotherapy induced anemia
3	10003019	25179393	4	78061	9	Fever presenting with conditions classified el...
4	10003019	25179393	5	E9331	9	Antineoplastic and immunosuppressive drugs cau...
...	...	...	...	...	...	...
74479	19998562	26846592	9	2859	9	Anemia, unspecified
74480	19998562	26846592	10	71941	9	Pain in joint, shoulder region
74481	19998562	26846592	11	33829	9	Other chronic pain
74482	19998562	26846592	12	78052	9	Insomnia, unspecified
74483	19998562	26846592	13	71947	9	Pain in joint, ankle and foot

74484 rows × 6 columns

A terceira tabela é a *neutro_itemid_loinc.csv* que serve com um catálogo dos tipos de exames que foram feitos nos pacientes que fazem parte dos dados fornecidos nas demais tabelas.

Não há informação que associe estes exames aos diagnósticos que cada paciente recebeu.

Tabela 3 - Visão parcial da tabela neutro_itemid_loinc.csv

	itemid	label	fluid	category	valueuom	loinc	loinc_version	notes
0	50856	Acetaminophen	Blood	Chemistry	ug/mL	NaN	2.71	\N
1	50867	Amylase	Blood	Chemistry	IU/L	NaN	2.71	\N
2	50873	Anti-Nuclear Antibody	Blood	Chemistry	NaN	NaN	2.71	\N
3	50874	Anti-Nuclear Antibody, Titer	Blood	Chemistry	NaN	NaN	2.71	\N
4	50893	Calcium, Total	Blood	Chemistry	mg/dL	NaN	2.71	\N
...	...	...	...	...	...	...	...	...
750	52358	NRBC#	Other Body Fluid	Hematology	\N	NaN	2.71	\N
751	52369	\N	Other Body Fluid	Hematology	\N	NaN	2.71	\N
752	52373	Blasts#	Pleural	Hematology	\N	NaN	2.71	\N
753	52391	FL1-S	Q	Hematology	\N	NaN	2.71	\N
754	52419	RFXUCU	Urine	Hematology	\N	NaN	2.71	\N

755 rows × 8 columns

Por fim, a quarta tabela fornecida é a *neutro_labevents.csv*, que contêm informações sobre os exames de laboratório feitos com cada paciente em suas admissões. Esta tabela identifica os exames e seus resultados, os relacionando com a identificação da admissão de cada paciente. Se trata de um número bastante grande de exames, sendo que são vários exames para cada admissão e muitos são repetidos em datas diferentes para uma mesma admissão.

Tabela 4 - Visão parcial da tabela neutro_labevents.csv

	hadm_id	specimen_id	itemid	charttime	value	valuenum	valueuom
0	25179393	30599113	51200	2175-12-09 08:10:00	0	0	%
1	25179393	30599113	51250	2175-12-09 08:10:00	102	102	fL
2	25179393	75561856	51200	2175-12-10 08:00:00	0	0	%
3	25179393	75561856	51250	2175-12-10 08:00:00	105	105	fL
4	25179393	59548859	51200	2175-12-11 08:25:00	1	1	%
...	...	...	...	...	...	...	...
3626561	26846592	68072203	51249	2166-04-12 07:00:00	34.8	34.8	%
3626562	26846592	72555587	51249	2166-04-13 07:15:00	34.1	34.1	%
3626563	26846592	5125140	51249	2166-04-14 07:20:00	33.8	33.8	%
3626564	26846592	20300538	51249	2166-04-15 07:32:00	33.8	33.8	%
3626565	26846592	31556856	51249	2166-04-16 07:20:00	34.3	34.3	%

3626566 rows × 7 columns

2.1 – Análise descritiva

Os dados contam com 4249 admissões para 2691 pacientes. Destes pacientes, 1492 têm diagnóstico de câncer, 55.4%.

Destas admissões, 292 tiveram desfecho de morte, e destes, 99 tinham diagnóstico de câncer, 33.9%.

A quantidade de pacientes de cada raça é bem desbalanceada e há 33 raças diferentes, que levam em conta não somente cor de pele, mas também, nacionalidade.

Os dados têm algumas inconsistências quanto à variável “race” pois há admissões diferentes de um mesmo paciente onde este foi categorizado em raças diferentes. Isso fez com que a quantidade de valores em “race” (2735) ficasse maior que a quantidade de pacientes (2691).

Por conta do desbalanceamento, não me parece ser viável qualquer análise considerando a informação de raça, pois 65.7% dos pacientes são da raça WHITE.

Figura 1 - Frequência das raças dos pacientes

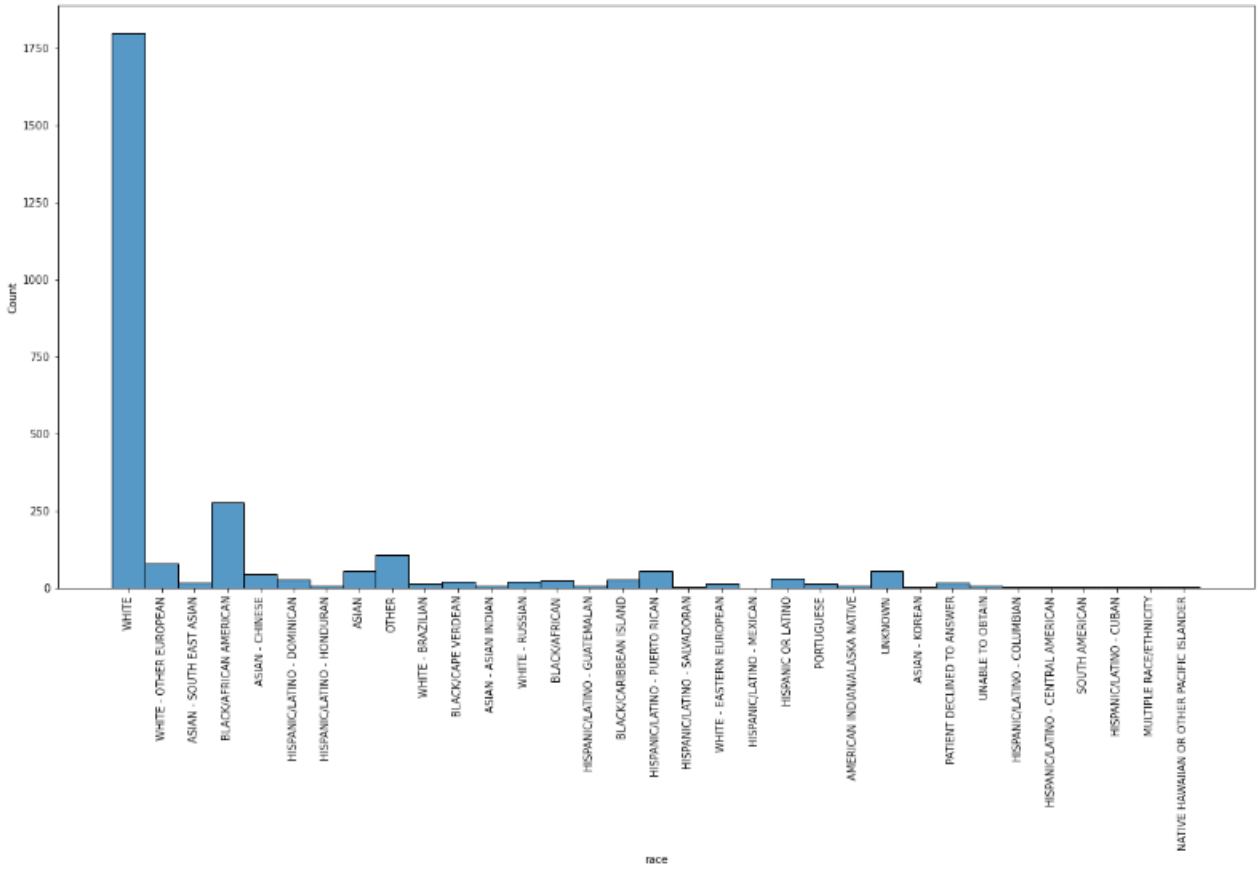


Tabela 5 - Valores das ocorrências de cada raça por paciente

WHITE	1797
BLACK/AFRICAN AMERICAN	277
OTHER	107
WHITE - OTHER EUROPEAN	79
HISPANIC/LATINO - PUERTO RICAN	56
UNKNOWN	55
ASIAN	54
ASIAN - CHINESE	46
HISPANIC OR LATINO	30
BLACK/CARIBBEAN ISLAND	27
HISPANIC/LATINO - DOMINICAN	26
BLACK/AFRICAN	23
WHITE - RUSSIAN	20
BLACK/CAPE VERDEAN	19
PATIENT DECLINED TO ANSWER	17
ASIAN - SOUTH EAST ASIAN	15
WHITE - EASTERN EUROPEAN	12
WHITE - BRAZILIAN	12
PORTUGUESE	12
UNABLE TO OBTAIN	7
ASIAN - ASIAN INDIAN	7
HISPANIC/LATINO - GUATEMALAN	7
AMERICAN INDIAN/ALASKA NATIVE	5
HISPANIC/LATINO - HONDURAN	5
HISPANIC/LATINO - CENTRAL AMERICAN	3
HISPANIC/LATINO - SALVADORAN	3
ASIAN - KOREAN	3
SOUTH AMERICAN	2
NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	2
HISPANIC/LATINO - CUBAN	2
MULTIPLE RACE/ETHNICITY	2
HISPANIC/LATINO - COLUMBIAN	2
HISPANIC/LATINO - MEXICAN	1

Name: race, dtype: int64
Total: 2735

Com relação aos dias de internação, as admissões com 3 dias de duração foram as mais frequentes.

Figura 2 - Contagem de admissões por dias de internação

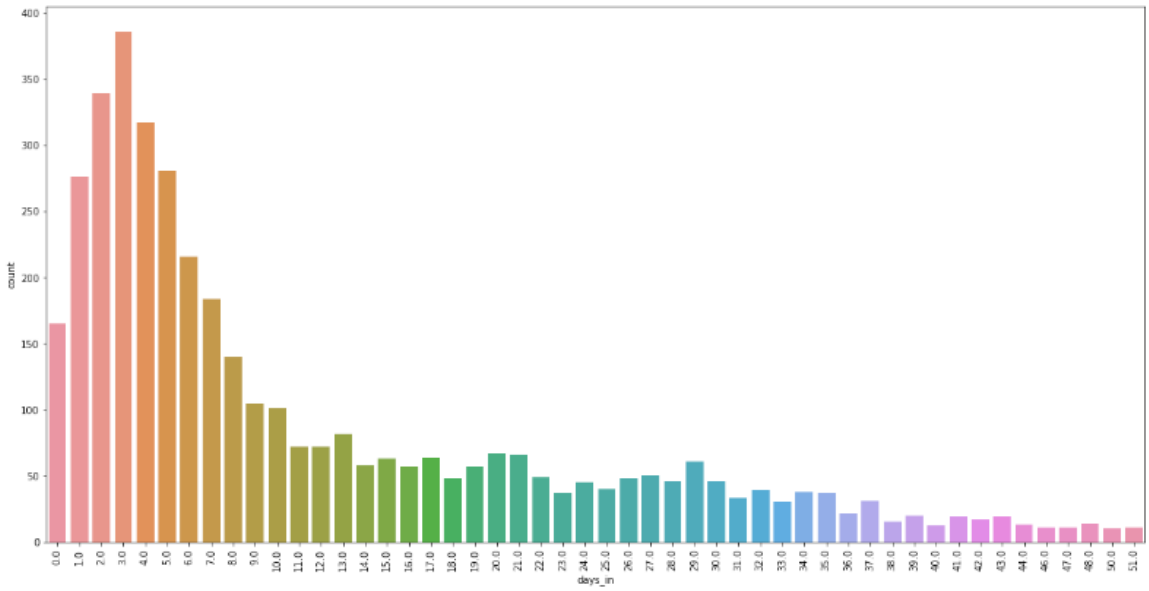


Tabela 6 - Visualização parcial das quantidades de admissões para cada quantidade de dias de internação

	days_in	count
3	3.0	386
2	2.0	340
4	4.0	317
5	5.0	281
1	1.0	276
...	...	...
94	100.0	1
93	99.0	1
89	93.0	1
85	89.0	1
122	230.0	1

123 rows × 2 columns

Já os desfechos de morte ocorreram com maior frequência em admissões com 01 e 06 dias de internação. Ambos os casos tiveram 13 ocorrências.

Figura 3 - Contagem de admissões com desfecho de morte em relação à quantidade de dias de internação

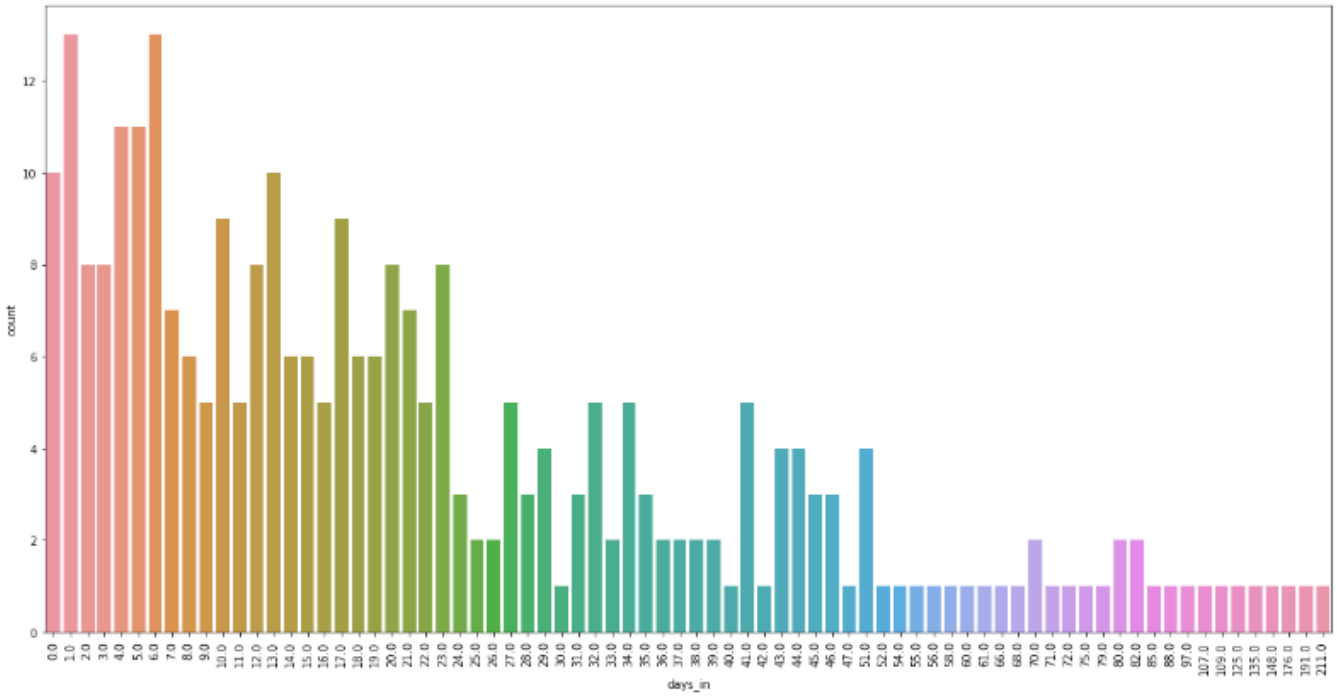


Tabela 7 - Visualização parcial das quantidades de mortes em comparação com os dias de duração de internação

	days_in	count
1	1.0	13
6	6.0	13
4	4.0	11
5	5.0	11
0	0.0	10
...	...	...
51	55.0	1
50	54.0	1
49	52.0	1
47	47.0	1
75	211.0	1

76 rows × 2 columns

Também fiz uma avaliação das quantidades de admissões levando em conta o tipo. A maior porte das admissões foi de pacientes que buscaram atendimento de emergência com 1456 admissões.

Tabela 8 - Número de admissões por tipo

	admission_type	count
5	EW EMER.	1456
6	OBSERVATION ADMIT	1158
1	DIRECT EMER.	733
3	ELECTIVE	353
8	URGENT	297
4	EU OBSERVATION	141
2	DIRECT OBSERVATION	83
7	SURGICAL SAME DAY ADMISSION	19
0	AMBULATORY OBSERVATION	9

Considerando somente as admissões com desfecho de morte, e assumindo que “EM EMER.”, “DIRECT EMER.” E “URGENT” tratam de admissões de emergência, o total de mortes para esses tipos de admissão foi de 194, que representa 66.4% do total de casos de óbito.

Tabela 9 - Número de admissões por tipo e que tiveram desfecho de morte

	admission_type	count
2	EW EMER.	110
3	OBSERVATION ADMIT	84
0	DIRECT EMER.	56
5	URGENT	28
1	ELECTIVE	13
4	SURGICAL SAME DAY ADMISSION	1

Também fiz uma avaliação de quantos foram os casos de morte para cada tipo de diagnóstico de cada paciente. Porém esta análise não permite uma conclusão sobre a associação destas informações pois um mesmo paciente é contado várias vezes, já que há várias informações de diagnóstico para um mesmo paciente.

Tabela 10 - Vinte diagnósticos mais frequentes nos casos onde o desfecho foi de morte

	icd_code	long_title	count
1708	Z66	Do not resuscitate	131
207	28800	Neutropenia, unspecified	118
1705	Z515	Encounter for palliative care	113
1221	J9601	Acute respiratory failure with hypoxia	77
1504	R5081	Fever presenting with conditions classified el...	72
849	D709	Neutropenia, unspecified	71
1418	N179	Acute kidney failure, unspecified	69
1769	Z87891	Personal history of nicotine dependence	65
566	78061	Fever presenting with conditions classified el...	63
649	99592	Severe sepsis	60
681	A419	Sepsis, unspecified organism	57
930	E872	Acidosis	56
1522	R6521	Severe sepsis with septic shock	55
398	51881	Acute respiratory failure	54
1568	T451X5A	Adverse effect of antineoplastic and immunosup...	54
281	4019	Unspecified essential hypertension	52
1050	I10	Essential (primary) hypertension	51
1652	V4986	Do not resuscitate status	51
481	5849	Acute kidney failure, unspecified	49
906	E785	Hyperlipidemia, unspecified	48

Outra análise foi feita quanto ao número de exames feitos para cada paciente onde houve desfecho de morte. Novamente há uma dificuldade na análise pelo fato de que um único caso de morte é contado várias vezes pois vários exames foram feitos para cada paciente.

Além disso, não há informação sobre o relacionamento entre os exames e os diagnósticos, o que poderia permitir simplificar e reduzir a quantidade de variáveis nas análises.

Tabela 11 - Vinte exames mais frequentes nos casos onde o desfecho foi de morte

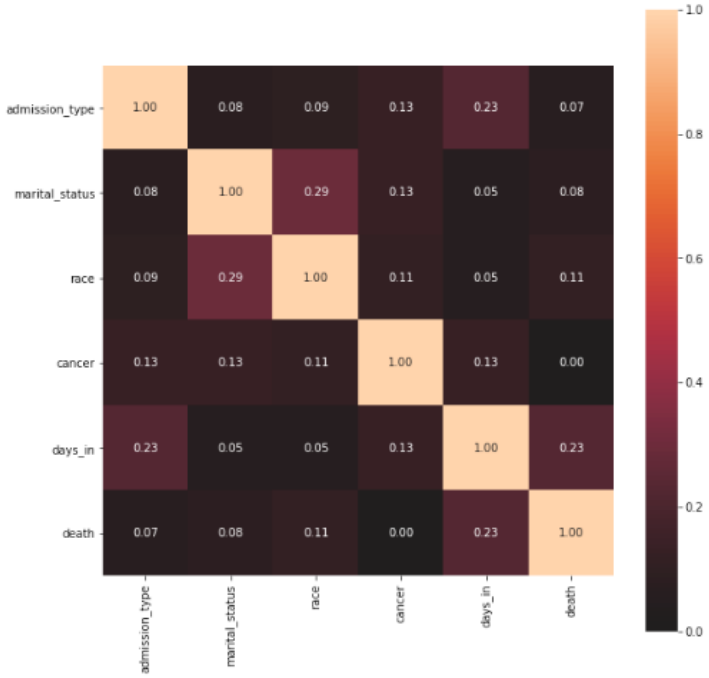
	itemid	label	count
210	51265	Platelet Count	11994
138	50971	Potassium	10672
146	50983	Sodium	10606
99	50902	Chloride	10539
86	50882	Bicarbonate	10416
107	50912	Creatinine	10405
81	50868	Anion Gap	10398
156	51006	Urea Nitrogen	10347
119	50931	Glucose	10319
200	51221	Hematocrit	10249
14	50960	Magnesium	10111
137	50970	Phosphate	10104
4	50893	Calcium, Total	10080
201	51222	Hemoglobin	10004
45	51301	White Blood Cells	9952
216	51279	Red Blood Cells	9926
206	51250	MCV	9926
205	51249	MCHC	9926
204	51248	MCH	9926
215	51277	RDW	9923

2.2 – Análise de correlações

Com a biblioteca Dython, foi feita uma análise de associação entre as variáveis para encontrar uma correlação entre elas incluindo os desfechos.

As variáveis se apresentam muito pouco correlacionadas, principalmente quando consideramos a correlação dos casos de morte com as demais variáveis, onde a maior correlação foi de 23% entre o número de dias de internação e o desfecho de morte ou não morte.

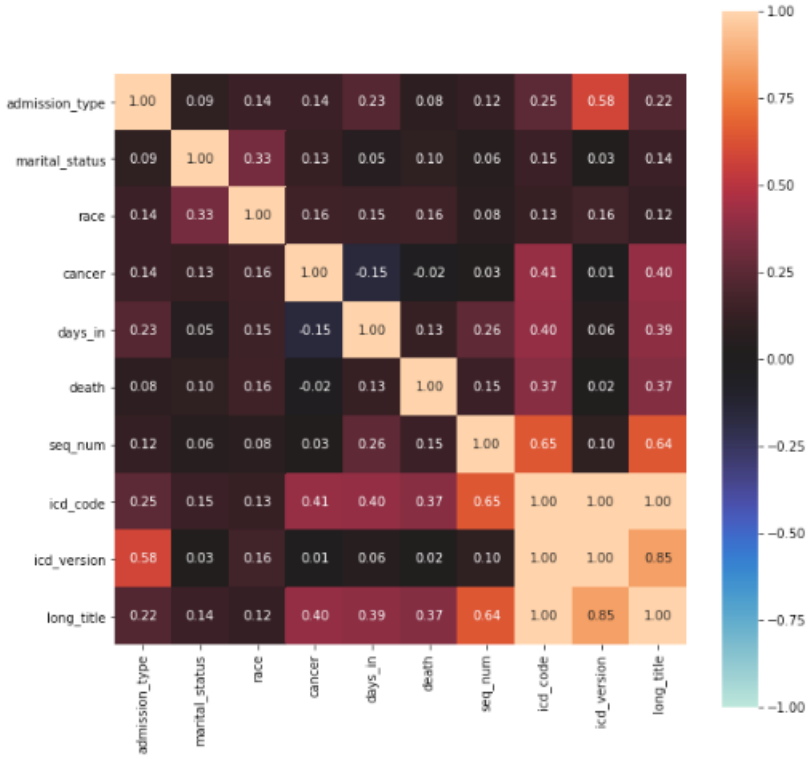
Figura 4 - Correlação entre as variáveis principais do problema



Levando-se em conta os diagnósticos que foram feitos para cada paciente em suas admissões, o resultado da matriz de correlação apresenta um valor de 37% entre os diagnósticos, representados pelas variáveis “icd_code” e “long_title”, e o desfecho de morte ou não morte.

Os diagnósticos também apresentam uma correlação maior com a variável que indica paciente com câncer e a que indica o número de dias de internação. Estes valores ficaram entre 39% e 41%.

Figura 5 - Matrix de correlação entre as variáveis de admissão e de diagnósticos



2.3 – Análise preditiva

As baixas correlações entre as variáveis dão uma indicação de que a predição do desfecho pode ser bem desafiadora. Além disso, há um enorme desbalanceamento entre as classe do problema.

Em minha primeira abordagem, como há muitas informações repetidas para pacientes, admissões e exames, busquei definir alguns critérios que me permitissem unir as tabelas em uma única que pudesse ser analisada de forma conjunta.

Optei por ter uma tabela onde os exames laboratoriais aos quais cada paciente foi submetido fossem colocados como variáveis binárias com as variáveis de admissão.

Antes disso, eu removi admissões de um mesmo paciente, deixando somente a mais recente. Eliminando assim, linhas de pacientes repetidos, mas garantindo que a admissão onde houve o desfecho mais importante, de morte ou não morte, fosse mantida.

Outro critério para a montagem desta tabela foi a de considerar a primeira ocorrência de cada exame laboratorial de cada paciente. Com isso, elimino exames repetidos para um mesmo paciente e considero os exames feitos no início da internação, refletindo as condições clínicas iniciais do paciente.

Todas as demais variáveis categóricas também foram binarizadas e valores nulos foram substituídos por zero.

Esta abordagem resultou em uma tabela com 4209 linhas e 814 colunas. Uma visão parcial da tabela é vista na *tabela 12*.

Há um grande desbalanceamento entre as classes de desfecho de morte ou não morte onde 291 amostras do total de 4209 era de desfecho de morte. Isso representa 7% das amostras.

Os modelos que utilizei foram do ScikitLearn e possuem opção de fazer balanceamento dos dados através da atribuição de pesos às amostras para que a classe minoritária tenha um aumento do peso em comparação à classe majoritária.

Tabela 12 – Visão parcial do data frame gerado com todas as variáveis dos exames e das admissões. Foram removidas as variáveis de datas e identificação do paciente e da admissão. As variáveis categóricas foram binarizadas.

	51200	51250	51251	51301	51252	50902	51254	50954	51255	51256	...	race_UNABLE TO OBTAIN	race_UNKNOWN	race_WHITE	race_WHITE - BRAZILIAN	race_WHITE - EASTERN EUROPEAN	race - EUR
0	1.0	102.0	4.0	1.9	NaN	109.0	28.0	371.0	11.0	19.0	...	0	0	1	0	0	
1	2.0	85.0	0.0	2.5	NaN	102.0	45.0	0.0	0.0	16.0	...	0	0	1	0	0	
2	1.0	93.0	0.0	0.2	NaN	106.0	6.0	100.0	0.0	2.0	...	0	0	0	0	0	
3	0.0	92.0	2.0	0.2	NaN	103.0	27.0	130.0	5.0	1.0	...	0	0	0	0	0	
4	1.0	98.0	1.0	6.4	NaN	104.0	5.0	334.0	3.0	9.0	...	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4204	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0	0	1	0	0	
4205	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0	0	1	0	0	
4206	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0	0	1	0	0	
4207	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0	0	0	0	0	
4208	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0	0	1	0	0	

4209 rows × 814 columns

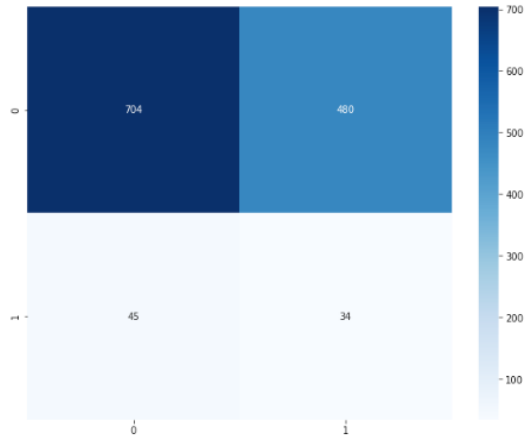
2.3.1 – Regressão logística

Foi feita uma regressão logística que não chegou a resultado satisfatório, mesmo com o uso do recurso de dar maior peso à classe minoritária.

A acurácia foi de 58.4%, mas o F1 Score ficou em 11.5%.

A matriz de confusão apresentou um número muito alto de Falsos Negativos, 480 num conjunto de teste de 1263 casos.

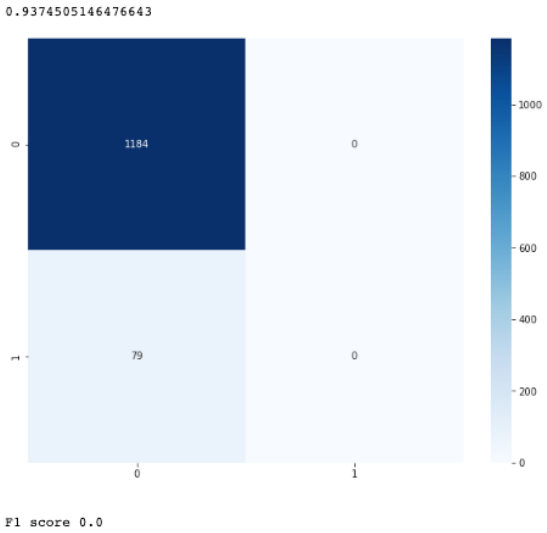
Figura 6 - Matriz de confusão da Regressão Logística para predizer o desfecho considerando o conjunto de dados de admissão e exames laboratoriais



2.3.2 – Random Forest Classifier

Fiz então uma tentativa de fazer a classificação usando Random Forest. O resultado foi ainda pior que o anterior. O modelo gerou uma acurácia de 93.7%, mas não acertou um único caso Verdadeiro Positivo. O F1 score foi zero. A matriz de confusão pode ser vista na *figura 7*.

Figura 7 - Matriz de confusão gerada pelo modelo de Árvores Aleatórias. Apesar da acurácia ter sido 0.937, o F1 score foi zero, mesmo tendo sido utilizado o recurso de balanceamento das classes através da atribuição de pesos pelo modelo.



2.3.3 – Análise preditiva somente para pacientes com câncer

Meu próximo passo foi avaliar a capacidade preditiva de desfecho de morte somente dos pacientes diagnosticados com câncer.

Se trata de um volume de 1476 pacientes que tinham diagnóstico de câncer, dos quais 98 tiveram desfecho de morte.

Novamente, há um desbalanceamento entre as classes, onde somente 7% das amostras tiveram desfecho de morte.

Os modelos que utilizei foram do ScikitLearn que possuem opção de fazer balanceamento dos dados através da atribuição de pesos às amostras para que a classe minoritária tenha um aumento do peso em comparação à classe majoritária.

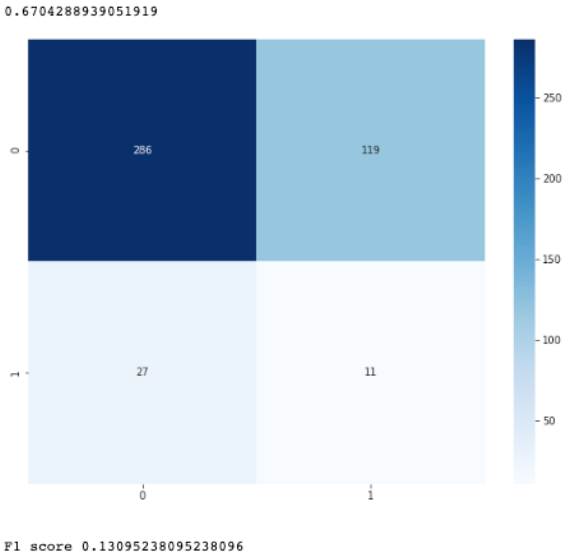
2.3.3.1 – Regressão Logística

A regressão logística não foi capaz de gerar bons resultados.

A acurácia foi de 67.0%, mas o F1 Score ficou em 13.1%.

A matriz de confusão gerada por este modelo é mostrada na *figura 8*.

Figura 8 - Matriz de confusão resultante do modelo de regressão logística aplicado somente às amostras de pacientes com câncer. Novamente, ele não foi capaz de predizer os desfechos de morte de forma aceitável.

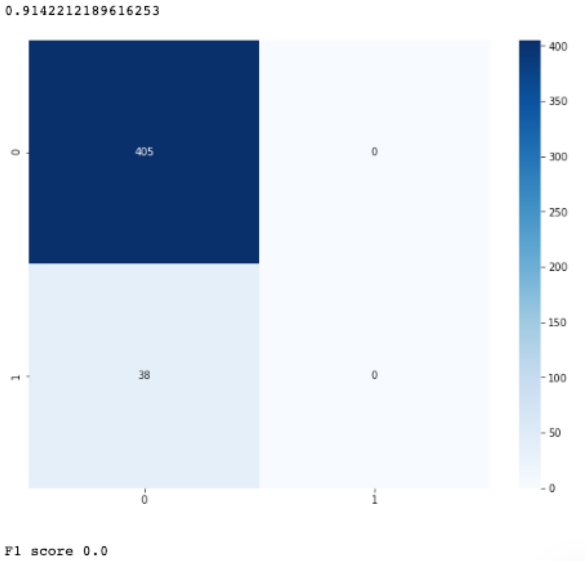


2.3.3.2 – Random Forest Classifier

Mesmo considerando somente os casos de pacientes com câncer, o Random Forest não foi capaz de fazer boas predições.

A acurácia foi de alta, 91.4%, mas o F1 score foi zero. A matriz de confusão é mostrada na *figura 9*.

Figura 9 - Resultado da predição de desfecho com Random Forest somente em pacientes com câncer



2.3.4 – Abordagem com balanceamento de classes pela geração de dados sintéticos

Como os resultados de predição foram muito ruins, e assumindo que um dos motivos deva ser por conta do alto desbalanceamento entre os dois tipos de desfecho, decidi fazer mais uma tentativa de gerar modelos, mas agora com o uso de dados gerados sinteticamente para que haja um melhor balanceamento dos dados.

No Scikit-Learn existe um método para geração de dados sintéticos chamado de ADASYN (Adaptive Synthetic). Utilizando este método decidi, agora, fazer uma separação de dados que me permitisse fazer um ciclo *treino, validação* e *teste* utilizando todos os dados que foram usados nos itens *2.3.1* e *2.3.2*, contendo pacientes com e sem câncer.

Foi feita esta divisão dos dados:

- ⇨ Treinamento e validação: 75% dos dados originais
- ⇨ Teste: 5% dos dados originais

Os dados sintéticos foram gerados em cima do pacote de treinamento e validação. Nestes dados foi feita a seguinte divisão:

- ⇨ 80% para o treinamento
- ⇨ 20% para validação do modelo

Foram treinados dois modelos, regressão logística e árvores aleatórias. O modelo que viesse a ter melhor desempenho na validação seria avaliado com os dados de teste.

O resultado obtido com regressão logística é apresentado na *figura 10* e com as árvores aleatórias, da *figura 11*.

Figura 10 - Resultado da regressão logística com os dados completos e balanceados com dados sintéticos

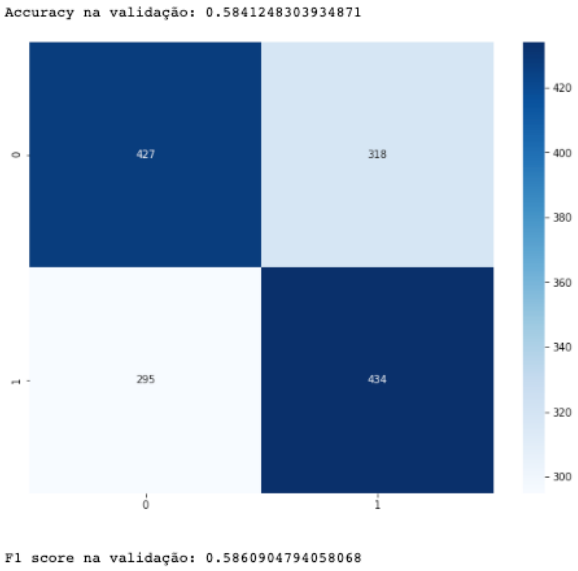
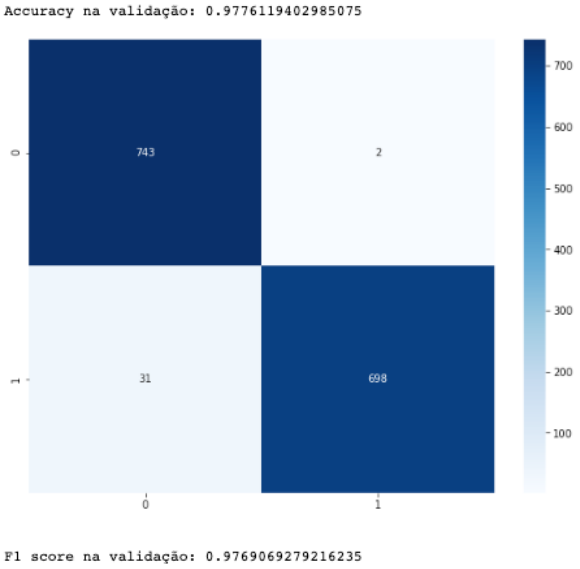


Figura 11 - Resultado da random forest com os dados completos e balanceados com dados sintéticos

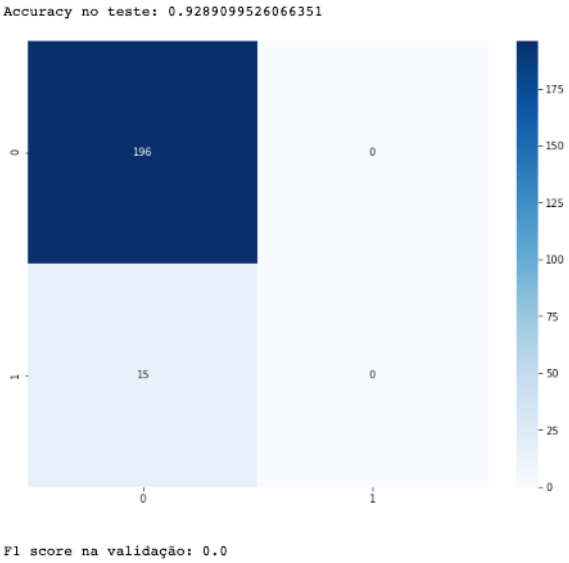


Houve grande melhoria no processo de treinamento e validação, onde a *random forest* conseguiu um resultado de 97.8% de acuraria e F1 score de 97.7%.

Sendo assim, o modelo de *random forest* foi escolhido para o teste com os dados reservados para este fim. Porém, o modelo não foi capaz de generalizar os resultados e obteve um resultado muito ruim com os dados de teste, não sendo capaz de identificar nenhum dos 15 casos de desfecho de morte. Apesar de ter obtido uma acurácia de 92.9%, o F1 score foi zero.

A matriz de confusão do modelo com os dados de teste é apresentada na *figura 12*.

Figura 12 - Matriz de confusão resultante do modelo random forest aplicado aos dados de teste separados previamente



3 - Conclusão

A análise dos dados do problema mostrou-se bem desafiadora. A variabilidade e número de valores distintos para exames e diagnósticos, além da quantidade de exames, e a repetição de exames, diagnósticos e admissões para um mesmo paciente, adicionaram complexidade ao problema.

Com as abordagens que utilizei no desenvolvimento das minhas análises não consegui identificar associações significativas entre as variáveis fornecidas e os desfechos de morte e não morte. Mesmo os casos de mortes onde havia diagnóstico de câncer, pelos critérios definidos no problema, não apresentou diferença significativa no desfecho de morte, que permitisse associar os casos de morte com o diagnóstico de câncer. A mortalidade geral ficou muito próxima da mortalidade dos pacientes com câncer, com 7.1% de mortes de pessoas sem câncer e 6.6% de mortes de pacientes com câncer.

A análise de correlações também não apontou valores significativos que permitissem a associação entre as variáveis.

Na análise preditiva, o grande desbalanceamento entre as classes de desfecho, afetou fortemente os resultados. A geração de dados sintéticos apontou uma possibilidade de melhores resultados

quando no treinamento e validação de modelos de regressão logística e de florestas aleatórias, mas o resultado não se comprovou com dados de teste separados previamente.

Uma seleção de dados clínicos gerais no dia da admissão, associados a resultados de exames específicos de neutropenia, bem como um critério mais preciso quanto ao diagnóstico de câncer, poderiam delimitar melhor o espaço de variáveis do problema e talvez adicionar informações que evidenciassem a associação entre as variáveis. Esse poderia ser um caminho a ser seguido para o desenvolvimento de modelos preditivos que conseguissem resultados melhores de predição.