Universidade Presbiteriana Mackenzie

FILA DE ESPERA PARA TRANSPLANTE RENAL NO BRASIL

Eduardo Nogueira Mota

Enzo Vemado

Jenifer Rinara

Vagner Milani

Trabalho apresentado como critério de avaliação da disciplina

PROJETO APLICADO II (Turma 03A).

Professor : Anderson Adaime de Borba

São Paulo

2024

Universidade Presbiteriana Mackenzie

Sumário

1. MÉTODO ANALÍTICO

1.1. PREPARAÇÃO E EXPLORAÇÃO DOS DADOS

Carregamento dos Dados: Inicialmente, os dados foram carregados a partir de um arquivo CSV contendo diversas informações demográficas e médicas dos pacientes na fila de espera por um transplante de rim.

Limpeza e Tratamento de Dados: Realizamos a limpeza dos dados, tratando valores ausentes e removendo registros onde os pacientes foram removidos da lista de espera, para garantir a precisão das previsões. As variáveis categóricas foram transformadas em variáveis numéricas através de codificação dummy, permitindo a utilização em modelos estatísticos.

```
colunas_categoricas = ['gestation', 'prior_transplant', 'subregion', 'cPRA_cat', 'DR_00', 'B_

df_dummies = pd.get_dummies(df[colunas_categoricas], drop_first=True)
df_dummies.head()
```

✓ 0.0s

	gestation_Sim	prior_transplant_Sim	subregion_HCFMUSP	subregion_UNICAMP	subregion_UNIFESP
0	False	False	False	False	True
1	False	False	False	True	False
2	False	False	False	False	True
3	False	True	False	False	True
4	False	False	False	False	True

Imagem 1 - Criação do dataset dummies.

```python
from sklearn.preprocessing import LabelEncoder

# criar uma copia do dataframe para manipulacoes
df = data.copy()

# codificar variaveis categoricas
label_encoders = {}
categorical_columns = ['sex', 'race', 'Blood_type', 'underline_disease', 'age_cat']

for col in categorical_columns:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le

# verificar se ha valores ausentes nas colunas relevantes e preenche-los se necessario
missing_values = df.isnull().sum()

# preenchendo valores ausentes na coluna 'time_on_Dialysis' com a media, ja que e uma variavel crucial
df['time_on_Dialysis'].fillna(df['time_on_Dialysis'].mean(), inplace=True)

# mostrar a nova estrutura do dataframe
df.head(), missing_values[missing_values > 0]
```

Imagem 2 - Limpeza e Tratamento.

1.2. SELEÇÃO E REDUÇÃO DE VARIÁVEIS

Seleção Inicial de Variáveis: Selecionamos variáveis iniciais baseadas na relevância clínica e demográfica, como idade, sexo, raça, tempo em diálise, tipo sanguíneo, entre outras.

Análise de Redução de Dimensionalidade com LASSO: Utilizamos o modelo de Regressão Logística com regularização Lasso para identificar e reter as variáveis mais significativas. O Lasso ajudou a reduzir a complexidade do modelo, penalizando e potencialmente reduzindo a zero os coeficientes de variáveis menos informativas.

	Feature	Coefficient
31	Time_Tx	911.787700
29	Time_death	19.045280
30	X36MthsTx	-18.157001
21	calculated_frequency_DR.f	-7.028829
20	calculated_frequency_DR.f2	6.354734
2	age_cat	-5.174321
27	calculated_frequency_A.f	-4.184418
26	calculated_frequency_A.f2	3.808732
19	calculated_frequency_DR.f1	3.378271
9	number_transfusion	-3.190059
13	HLA_A1	-2.580698
10	number_gestation	-2.320186
3	time_on_Dialysis	-2.237654
14	HLA_A2	2.218630
17	HLA_DR1	-2.121510

Imagem 3 - Resultado do modelo LASSO indicando o coeficiente de algumas das features

1.3. MODELAGEM ESTATÍSTICA

Construção do Modelo de Regressão Logística: Com as variáveis selecionadas, construímos um modelo de Regressão Logística. Este modelo é particularmente adequado para prever resultados binários, neste caso, se o tempo de espera é maior ou menor que a mediana.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# selecionar apenas as features com maior coefieciente
selected_features = [ 'Time_Tx','Time_death', 'X36MthsTx', 'calculated_frequency_DR.f', 'calculated_frequency_D

x_numericas = x_all_transformada[selected_features]


# get dummies para colunas categoricas
x_colunas = pd.concat([x_numericas, df_dummies], axis=1)
x_colunas = pd.concat([x_colunas, y_all], axis=1)

x_colunas= x_colunas[x_colunas['removed_list_Sim']==True]
y_all= x_colunas['time']
y_selected = (y_all > y_all.median()).astype(int)  # variavel binaria baseada na mediana do tempo de espera
x_colunas = x_colunas.drop('time', axis=1)
```

Imagem 4 - Selecionando as colunas e criando a variável de predição

Treinamento e Teste do Modelo: Dividimos os dados em conjuntos de treinamento e teste para validar a eficácia do modelo. O modelo foi treinado com o conjunto de treinamento e posteriormente testado para verificar sua precisão e robustez com dados não vistos anteriormente.

```
# treino e teste
X_train_col, X_test_col, y_train_sel, y_test_sel = train_test_split(x_colunas, y_selected, test_size=0.3, random_state=0)

# escalar os dados
X_train_col_scaled = scaler.fit_transform(X_train_col)
X_test_col_scaled = scaler.transform(X_test_col)

#modelo com todas as variaveis
log_reg_enh = LogisticRegression(max_iter=1000)
log_reg_enh.fit(X_train_col_scaled, y_train_sel)


y_pred_enh = log_reg_enh.predict(X_test_col_scaled)
```

Imagem 5 - Preparando o modelo, divisão entre treino e teste.

2. CÁLCULO DAS MÉTRICAS

```
# metricas de desempenho
accuracy_enh = accuracy_score(y_test_sel, y_pred_enh)
precision_sel = precision_score(y_test_sel, y_pred_enh)
recall_enh = recall_score(y_test_sel, y_pred_enh)
f1_enh = f1_score(y_test_sel, y_pred_enh)

accuracy_enh, precision_sel, recall_enh, f1_enh
```

✓ 0.0s

```
(0.9827849204619743, 0.9763676148796498, 0.9889184397163121, 0.982602950891874)
```

Imagem 6 - Resultado das métricas do modelo

Após o treinamento, aplicamos o modelo ao conjunto de teste para prever se os pacientes teriam um tempo de espera acima da mediana. As seguintes métricas de desempenho foram calculadas para avaliar a precisão do modelo:

ACURÁCIA (ACCURACY): Esta métrica fornece uma visão geral da proporção de previsões corretas feitas pelo modelo em relação ao total de previsões. Uma acurácia de 98.27% indica que o modelo é extremamente preciso em classificar os pacientes corretamente quanto ao seu tempo de espera na fila.

PRECISÃO (PRECISION): A precisão é a proporção de identificações positivas corretas (pacientes com longo tempo de espera identificados pelo modelo) em relação ao total de identificações positivas feitas pelo modelo. Uma precisão de 97.63% mostra que quase todas as previsões de longo tempo de espera feitas pelo modelo estavam corretas.

RECALL (SENSIBILIDADE): O recall mede a proporção de positivos reais (pacientes com longo tempo de espera) que foram corretamente identificados. Um

recall de 98.89% indica que o modelo conseguiu identificar quase todos os pacientes que realmente enfrentam longos tempos de espera.

F1-SCORE: O F1-Score combina precisão e recall em uma única métrica, proporcionando um equilíbrio entre ambas. Um F1-Score de 98.26% é indicativo de que o modelo é equilibrado e eficaz, com alta precisão e capacidade de recall.

3. DESCRIÇÃO DOS RESULTADOS PRELIMINARES

3.1. ANÁLISE ESTATÍSTICA E IDENTIFICAÇÃO DE PADRÕES

Visualização de Importância das Características: Os gráficos a seguir mostram as características mais importantes identificadas pelo modelo. Variáveis como o tempo de tratamento antes do transplante (Time_Tx) e a presença de condições médicas específicas foram altamente preditivas para identificar longos tempos de espera.

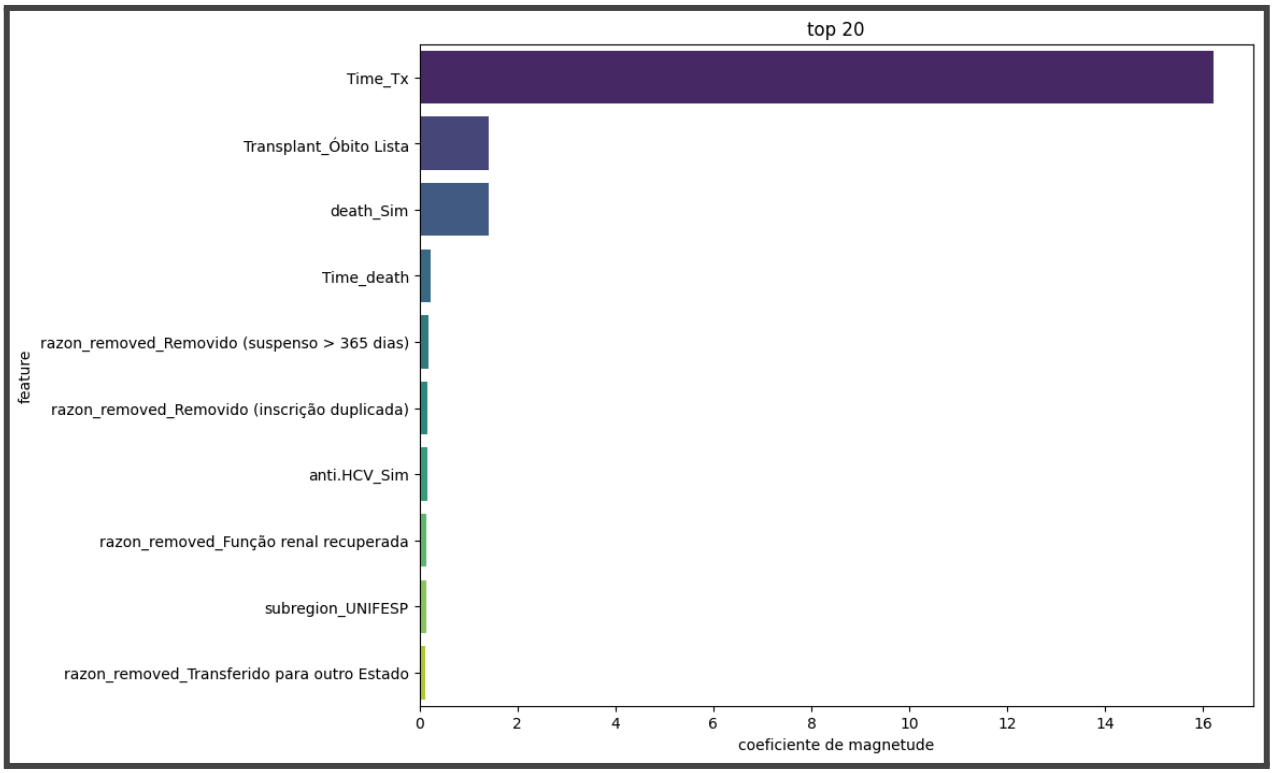

Imagem 7 - Características mais importantes identificadas pelo modelo

Distribuição dos Pacientes com Longos Tempos de Espera: Os gráficos de distribuição para cada variável importante ajudam a visualizar como os pacientes com longos tempos de espera diferem dos demais. Por exemplo, pacientes mais velhos ou com certas comorbidades tendem a esperar mais tempo.

Os resultados preliminares indicam que o modelo é extremamente eficaz em identificar pacientes com maior probabilidade de enfrentar longos períodos de espera. Com esses insights, podemos propor a criação de uma plataforma ou serviço para gestores de saúde que fornece análises preditivas sobre a fila de espera para transplantes de rins. Esta plataforma poderia ajudar a alocar recursos de forma mais eficiente, priorizar pacientes com base em urgência médica e demografia, e potencialmente colaborar com políticas públicas para melhorar os sistemas de saúde em relação aos transplantes de órgãos.

3.2. IMPACTO POTENCIAL

Eficiência Operacional: Melhoria na alocação de recursos e na priorização de pacientes, reduzindo o tempo médio de espera e potencialmente salvando mais vidas.

Políticas de Saúde: Informação detalhada e baseada em dados que pode influenciar políticas públicas e práticas recomendadas para a gestão de listas de espera.

4. STORYTELLING

4.1. INTRODUÇÃO

Contexto: Todos os anos, milhares de pacientes aguardam um transplante de rim no Brasil, enfrentando longos períodos de espera que podem afetar significativamente a qualidade e expectativa de vida.

Problema: A distribuição desigual do tempo de espera entre diferentes grupos demográficos levanta questões sobre eficiência e justiça no processo de alocação de órgãos.

4.2. ANÁLISE DE DADOS

Dados utilizados: Informações demográficas e médicas de pacientes na fila de espera para transplante de rim.

Método analítico: Utilização da Regressão Logística para prever quem enfrentará longos tempos de espera, com base em variáveis como idade, sexo, raça, e tempo em diálise.

Tratamento de dados: Exclusão de dados de pacientes removidos da lista para garantir a precisão das previsões.

4.3. DESCOBERTAS CHAVE

Insights importantes: Identificação de variáveis-chave que influenciam o tempo de espera, como o tipo sanguíneo e comorbidades como diabetes.

Visualização de dados: Gráficos mostrando a distribuição do tempo de espera e a importância relativa das diferentes variáveis.

Ajuste fino: Refinamento do modelo através da análise de cluster para entender melhor os grupos de pacientes.

4.4. SOLUÇÃO PROPOSTA

Plataforma analítica: Desenvolvimento de uma ferramenta (Dashboard) que utiliza o modelo para oferecer previsões sobre o tempo de espera, ajudando hospitais a priorizar pacientes.

Benefícios esperados: Melhoria na eficiência da alocação de órgãos, redução nas disparidades entre diferentes grupos e aumento na transparência do processo.

4.5. CHAMADA À AÇÃO

Para profissionais de saúde: Encorajamento para adotar a ferramenta e adaptar práticas baseadas em uma compreensão aprofundada dos dados.

Para gestores executivos: Convite para considerar dados analíticos como parte integral das estratégias de melhoria do sistema de saúde.