**Uso de Inteligeˆncia Artificial para Predic¸a˜o de Sobreviveˆncia de Pacientes com Caˆncer de Pro´stata**

**Alcir Canella Filho**2**, Felipe Cle´ Monteiro**2**, Matheus do Nascimento Marques**2

1Sistemas de Informac¸a˜o Faculdade de Computac¸a˜o e Informa´tica Universidade Presbiteriana Mackenzie

Sa˜o Paulo – SP – Brasil

2Cieˆncia da Computac¸a˜o Faculdade de Computac¸a˜o e Informa´tica Universidade Presbiteriana Mackenzie

Sa˜o Paulo – SP – Brasil

*{*10396389,10395521,10395894*}*@mackenzista.com.br

***Abstract.*** *The International Agency for Research on Cancer (IARC-UN), to- gether with The Lancet Commission, predicts that the number of prostate can- cer cases will more than double by 2040, and that this imminent increase will imply a rapid rise in the global rate of deaths from the disease. In view of this scenario, this study was initiated to apply survival prediction models using data from patients with prostate cancer, focusing on the morphology of acinar cell carcinoma, which is the most common for this type of cancer. A public database maintained by the Fundac¸a˜o Oncocentro de Sa˜o Paulo was used, where data was found from the year 2000 to 2020 on approximately 70 thousand patients with the studied morphology, belonging to multiple age groups and different cli- nical staging groups. In this initial phase of the study, preparation, exploratory analysis and pre-processing of data were completed to enable the application of computational predictive models, with the aim of validating their accuracy to assist clinical or public decisions.*

***Resumo.*** *A Ageˆncia Internacional de Pesquisa em Caˆncer (IARC-ONU), junto a The Lancet Commission, preveˆ que o nu´mero de casos de caˆncer de pro´stata mais que dobrara´ ate´ 2040, e que este aumento iminente implicara´ uma subida ra´pida da taxa global de mortes pela enfermidade. Tendo em vista tal cena´rio, este estudo foi iniciado para aplicar modelos de predic¸a˜o de sobreviveˆncia uti- lizando dados de pacientes com caˆncer de pro´stata, focando na morfologia car- cinoma de ce´lulas acinosas, sendo esta a mais comum para tal tipo de caˆncer. Foi utilizada uma base de dados pu´blica mantida pela Fundac¸a˜o Oncocentro de Sa˜o Paulo, onde foram encontrados dados desde o ano 2000 ate´ 2020 de aproximadamente 70 mil pacientes com a morfologia estudada, pertencendo a mu´ltiplas faixas eta´rias e diferentes grupos de estadiamento cl´ınico. Nesta fase do estudo, foram conclu´ıdas a preparac¸a˜o, a ana´lise explorato´ria, o pre´-processamento dos dados e a aplicac¸a˜o de modelos preditivos computacionais, para auxiliar deciso˜es cl´ınicas ou pu´blicas.*

***Palavras-chave:*** *Caˆncer, Dados, Ana´lise, Dashboard, Predic¸a˜o, Aprendizado, Ma´quina.*

# Introduc¸a˜o

O caˆncer e´ uma das principais causas de morbidade e mortalidade em todo o

mundo[Ferlay et al. 2015], apresentando um significativo desafio para a sau´de pu´blica. De acordo com o Instituto Nacional de Caˆncer[INCA 2022], as projec¸o˜es para o Brasil no trieˆnio 2023-2025 indicam que, excluindo os casos de caˆncer de pele na˜o melanoma, havera˜o cerca de 704 mil casos novos de caˆncer. Dentre estes, sa˜o previstos aproximada- mente 71 mil casos de cancer de pro´stata.

Ha´ a expectativa de que o nu´mero de pacientes com caˆncer aumente. A Ageˆncia Internacional de Pesquisa em Caˆncer (IARC - ONU) em conjunto com a The Lancet Comission[James 2024] preveˆ que a carga global de cancer de pro´stata mais que dobrara´ ate´ 2040, aproximando-se de treˆs milho˜es de novos casos, comparado ao nu´mero de casos estimados atualmente. Segundo a The Lancet Comission, se ac¸o˜es na˜o forem tomadas, o aumento iminente de casos de caˆncer de pro´stata causaria uma subida ra´pida da taxa global de mortes pela enfermidade.

Diante de tal cena´rio, faz-se necessa´rio o estudo de me´todos de ana´lise e predic¸a˜o

que possam auxiliar tomadas de decisa˜o governamentais e cl´ınicas. E´ poss´ıvel apli-

car modelos que estimam as chances de sobreviveˆncia de um paciente utilizando as informac¸o˜es dispon´ıveis em bases de dados. Modelos de sobreviveˆncia sa˜o amplamente utilizados para auxiliar deciso˜es cl´ınicas, contando com diversos me´todos de aprendizado de ma´quina para obter predic¸o˜es de tempo ate´ o evento quando alguns dados esta˜o cen- surados [Suresh 2022]. Nesses contextos, os modelos devem ser precisos e interpreta´veis para que os utilizadores (como os me´dicos) possam confiar no modelo e compreender as previso˜es.

Existem bases de dados abertas com grandes volumes de dados sobre pacientes de caˆncer. Uma delas e´ a da Fundac¸a˜o Oncocentro de Sa˜o Paulo [FOSP 2022], que armazena informac¸o˜es sobre pacientes com caˆncer no estado de Sa˜o Paulo desde o ano de 2000, bem como detalhes do estadiamento cl´ınico, faixa eta´ria, cirurgias realizadas e sobreviveˆncia.

No artigo *Machine Learning for Predicting Survival of Colorectal Cancer Patients*[Buk Cardoso et al. 2023] sa˜o treinados treˆs diferentes modelos para predic¸a˜o de sobreviveˆncia de pessoas com caˆncer colorretal, utilizando Naive Bayes[Pedregosa et al. 2011], Random Forest[Ho 1995] e XGBoost[Chen and Guestrin 2016]. A ana´lise dos resultados mostram que os mo-

delos Random Forest e XGBoost, sendo estes dois baseados em a´rvores de decisa˜o,

obtiveram acura´cia superior quando comparados ao modelo Naive Bayes, comumente utilizado para classificac¸a˜o.

E´ estabelecido como objetivo deste estudo treinar modelos para predic¸a˜o de so-

breviveˆncia utilizando dados de caˆncer de pro´stata dispon´ıveis no site da Fundac¸a˜o Onco- centro de Sa˜o Paulo[FOSP 2022], tendo o me´todo publicado por Buk Cardoso como base para o processo. Esta˜o inclusos como objetivos os itens abaixo:

* Realizar ana´lise explorato´ria e preparar dados pu´blicos de caˆncer de pro´stata.

Tal parte foi conclu´ıda neste semestre e esta´ dispon´ıvei no *Github*. E´ poss´ıvel visualizar

enderec¸o do reposito´rio na sec¸a˜o Metodologia deste documento.

* Aplicar um modelo preditivo que integre os dados analisados para prever o im- pacto do caˆncer de pro´stata (carcinoma de ce´lulas acinosas) na longevidade dos pacientes.
* Validar o modelo preditivo para garantir sua precisa˜o e confiabilidade.
* Concluir a edic¸a˜o do artigo.

O estudo colaborara´ com a validac¸a˜o da utilizac¸a˜o tanto do me´todo quanto da

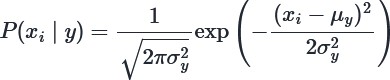
base de dados da Fundac¸a˜o Oncocentro de Sa˜o Paulo para predic¸a˜o de sobreviveˆncia de pacientes com caˆncer de pro´stata.

# Referencial Teo´rico

Segundo [Suresh 2022], modelos de sobreviveˆncia sa˜o utilizados em algum momento inicial de um paciente, tomando por exemplo a data do diagno´stico ou do in´ıcio do trata- mento, para descobrir a probabilidade de sobreviveˆncia do mesmo a determinadas janelas de tempo. Os me´dicos, enta˜o, utilizam os resultados de tais modelos para tomar deciso˜es cl´ınicas, tais como ajustar o tempo de monitoramento ou a aplicac¸a˜o de diferentes tera- pias.

No artigo *Machine Learning for Predicting Survival of Colorectal Cancer Pa- tients*, Buk Cardoso et al. demonstrou ser poss´ıvel utilizar uma das bases de dados pu´blicos disponibilizada pela Fundac¸a˜o Oncocentro de Sa˜o Paulo [FOSP 2022]para apli- car modelos e prever sobreviveˆncia de pacientes com caˆncer colorretal. Foram uti- lizados os modelos Naive Bayes [Pedregosa et al. 2011], Random Forest [Ho 1995] e XGBoost[Chen and Guestrin 2016] para classificar a sobrevida dos pacientes em 1, 3 e 5 anos, sendo estes treˆs modelos tambe´m objeto deste estudo.

O modelo Naive Bayes utiliza o teorema de Bayes, que pode ser implementado de diferentes maneiras e e´ amplamente utilizado para classificac¸a˜o, isto e´, o modelo tenta prever o ro´tulo correto de um determinado dado de entrada. Neste estudo, sera´ utilizada uma variac¸a˜o do algoritmo Naive Bayes representada pela fo´rmula abaixo, onde probabi- lidade dos recursos e´ considerada gaussiana[Pedregosa et al. 2011]:



e *µy*

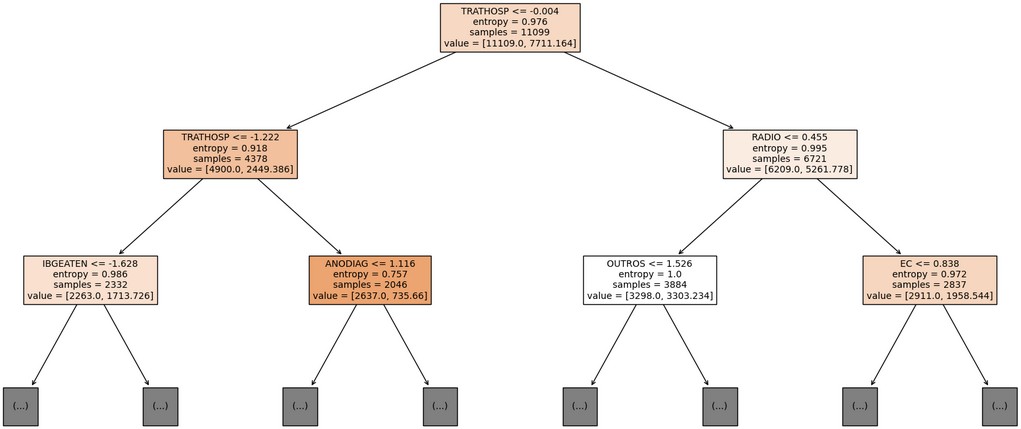
**Figura 1. Naive Bayes Gaussiano**

Os dois paraˆmetros a seguir sa˜o estimados usando a ma´xima verossimilhanc¸a: *σy*

O modelo Random Forest e´ um algoritmo de aprendizado de ma´quina que uti-

liza um conjunto (ou ”floresta”) de a´rvores de decisa˜o para melhorar a precisa˜o preditiva, [Ho 1995] sendo de ra´pido processamento. Um conjunto de a´rvores e´ um me´todo que utiliza mu´ltiplos classificadores de a´rvore de decisa˜o em diferentes subamostras de um conjunto de dados e emprega a me´dia dos resultados para aumentar a precisa˜o nas pre- viso˜es e reduzir o risco de sobreajuste.

Uma a´rvore de decisa˜o particiona recursivamente o espac¸o dos atributos para agru- par amostras que possuem ro´tulos ideˆnticos ou valores-alvo similares. Abaixo, um exem- plo do uma implementac¸a˜o de a´rvores de decisa˜o do artigo de [Buk Cardoso et al. 2023]



**Figura 2. Random Forest**

O algoritmo XGBoost e´ usado para problemas de aprendizagem supervisionada, onde sa˜o usados dados de treinamento (com mu´ltiplos recursos) para prever uma varia´vel

alvo. O XGBoost e´

um me´todo que cria um conjunto de

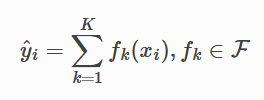
a´rvores de classificac¸a˜o e

regressa˜o (CART). Uma pontuac¸a˜o e´

atribu´ıda a cada uma das folhas da

a´rvore, di-

ferente das tradicionais a´rvores de decisa˜o que possuem somente valores de decisa˜o. [Chen and Guestrin 2016]. O modelo somara´ a previsa˜o de mu´ltiplas a´rvores juntas, como pode ser visto na figura 3



**Figura 3. Fo´ rmula de a´rvores XGBoost**

K representa o nu´mero de a´rvores, *fk* e´ uma func¸a˜o dentro do espac¸o funcional F, F e´ o conjunto de todas as A´ rvores de Classificac¸a˜o e Regressa˜o (CARTs) poss´ıveis.

O conjunto de dados onde sera˜o aplicados os modelos acima e´ oriundo de uma base pu´blica mantida pela Fundac¸a˜o Oncocentro de Sa˜o Paulo (FOSP), sendo a mesma fonte do *dataset* utilizado no artigo *Machine Learning for Predicting Survival of Colo- rectal Cancer Patients*. A FOSP foi fundada em 1974, sendo uma instituic¸a˜o vinculada a` Secretaria de Sau´de do Governo de Sa˜o Paulo com o intuito de prover assisteˆncia a` onco- logia e incentivar ensino e pesquisa, e tambe´m de estimular a detecc¸a˜o precoce do caˆncer e atividades de prevenc¸a˜o.[FOSP 2022]

A base de dados mantida e´ atualizada a cada treˆs meses com dados gerados por instituic¸o˜es coordenadas pela FOSP no estado de Sa˜o Paulo, sendo estas mantenedo-

ras do Registro Hospitalar de Caˆncer (RHC). O objetivo da Fundac¸a˜o Oncocentro ao

disponibilizar tais dados e´ pec´ıficas.[FOSP 2022]

# Metodologia

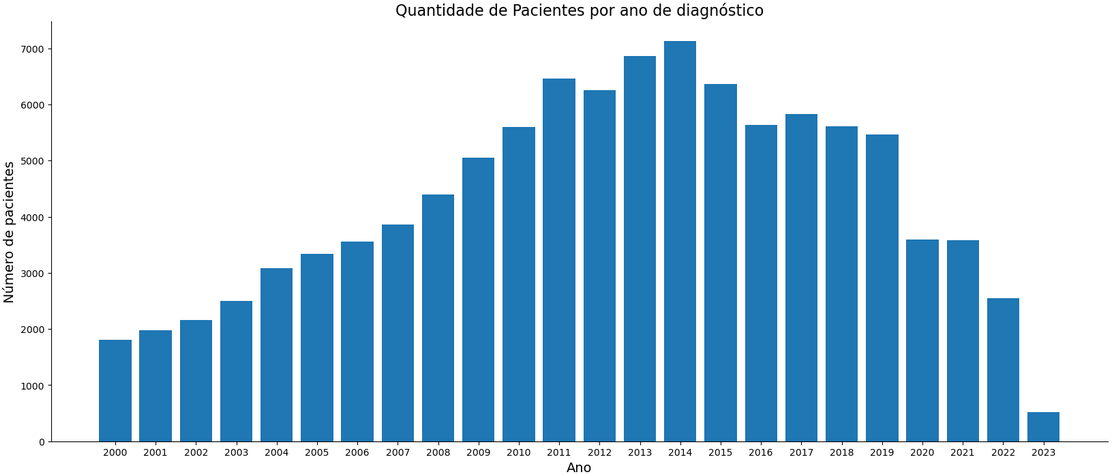
auxiliar profissionais da a´rea da sau´de a gerar ana´lises es-

Na fase de *Data Understanding* foi feita a extrac¸a˜o e revisa˜o dos dados. Foram apli- cadas as bibliotecas Pandas, Numpy , Matplotlib, Seaborn e Plotly para visualizac¸a˜o e entendimento dos dados.

Todos os pacientes esta˜o anonimizados, na˜o contendo nomes ou documentos. Os dados foram tratados de forma similar a` adotada no artigo de [Buk Cardoso et al. 2023], havendo necessidade de algumas mudanc¸as pontuais. Algumas diferenc¸as entre for- mato de campos de data ou atributos que na˜o foram mencionados no artigo esta- vam presentes no *dataset*. Tanto a base de dados quanto a descric¸a˜o dos campos esta˜o dispon´ıveis em: https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao- e-epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc

O *dataset* conta com mais de 100 mil pacientes de caˆncer de pro´stata, sendo pouco mais de 72 mil com carcinoma de ce´culas acinosas (morfologia 85503) entre 2000 e 2023. Tal morfologia e´ o objeto deste estudo, que tem foco nos pacientes registrados de 2000 a 2020.

O pico de diagnosticos foi no ano de 2014, com aproximadamente 7 mil pacientes diagnosticados, como pode ser visto na Figura 4.

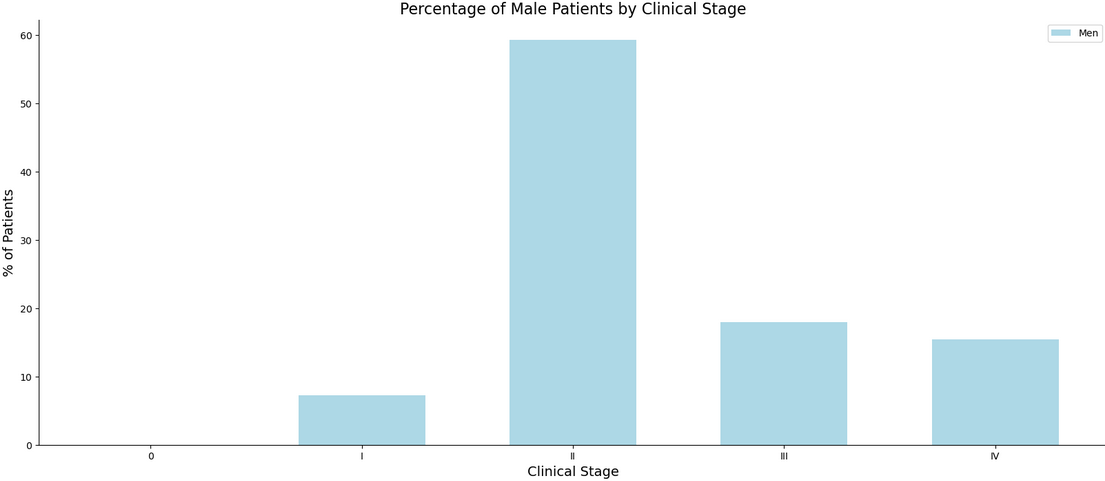


**Figura 4. Quantidade de diagnosticos por ano**

A maioria dos pacientes esta´ classificada no nível 2 do atributo estadiamento

cl´ınico (ver Figura 5). Este campo determina a extensa˜o do caˆncer pelo corpo do pa- ciente, bem como e´ utilizado para tomada de decisa˜o em relac¸a˜o a tratamentos.

Para a preparac¸a˜o dos dados, foram exclu´ıdos atributos com maior quantidade de dados faltantes no dataframe. Tambe´m foi aplicada correlac¸a˜o de Pearson para identificar atributos que na˜o sa˜o relevantes. Parte dos atributos foram descartados por na˜o possu´ırem variedade relevante de dados, na˜o sendo aproveita´veis pelos modelos.

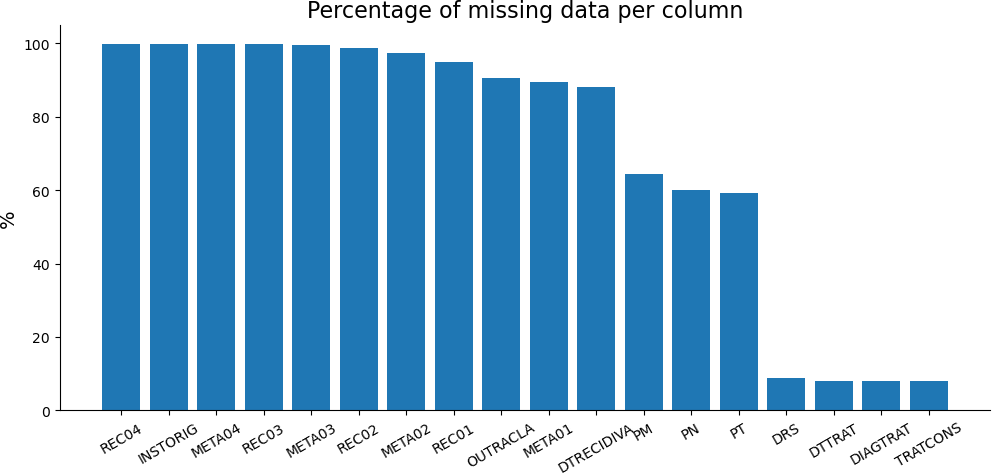


**Figura 5. Porcentagem de pacientes por grupo de estadiamento cl´ınico**

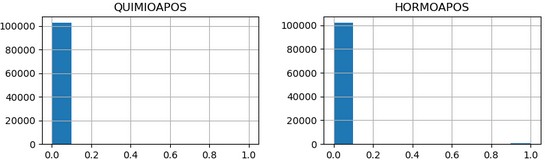
Foram deletadas as colunas abaixo por diversos motivos. Algumas possuem mui- tos dados faltantes, e na˜o seria poss´ıvel fazer um preenchimento preciso dos mesmos, ou por na˜o causarem impacto significativo (baixa ou nula variabilidade de valores). Ci- tando como exemplo a coluna MORFO, deletada pelo fato da ana´lise cobrir somente uma morfologia, ou o campo ECGRUP, que possui as mesmas informac¸o˜es contidas no campo EC. Os campos HABIT11, HABILIT1, e HABILIT2, por exemplo, na˜o continham informac¸o˜es de tratamento dos pacientes:

’UFRESID’, ’UFNASC’, ’CIDADE’, ’DESCTOPO’, ’DESC- MORFO’,’OUTRACLA’,’INSTORIG’ ’META01’, ’META02’,’META03’, ’META04’, ’REC01’, ’REC02’, ’REC03’,’REC04’, ’MORFO’, ’TOPO’, ’TOPOGRUP’, ’T’,

’N’, ’M’, ’NAOTRAT’, ’TRATAMENTO’, ’TRATFAPOS’, ’NENHUMAPOS’, ’CIRURAPOS’, ’RADIOAPOS’, ’QUIMIOAPOS’, ’HORMOAPOS’, ’IMUNO- APOS’,’OUTROAPOS’, ’RECLOCAL’, ’RECREGIO’, ’RECDIST’, ’HABI- LIT’,’INSTORIG’, ’CICI’,’CICIGRUP’,’CICISUBGRU’,’CLINICA’,’ECGRUP’, ’TRATFANTES’,’FAIXAETAR’,’PERDASEG’,’HABIT11’,’HABILIT1’,’HABILIT2’, ’CIDADEH’,’,’CIRU’,’S’,’QUIMIOANT’,’HORMOANT’,’TMOANT’,’IMUNOANT’, ’OUTROANT’,



**Figura 6. Colunas com maior quantidade de dados faltantes**



**Figura 7. Exemplos de atributos com os mesmos valores para todos os pacientes**

Os atributos abaixo foram exclu´ıdos apo´s aplicac¸a˜o da correlac¸a˜o de Pearson, tendo mostrado importaˆncia baixa ou nula:

’SEXO’,’LOCALTNM’,’IDMITOTIC’,’LATERALI’,’ERRO’, ’TMOAPOS’, ’CIRURANT’,’RADIOANT’

Foram excluídas linhas de pacientes cuja morfologia e´ diferente de carcinoma de ce´lulas acinosas, sendo esta o objeto do estudo.

Abaixo, os atributos do dataset escolhidos para o estudo:

Campo Descric¸a˜o

ESCOLARI Co´digo para escolaridade do paciente IDADE Idade do paciente

IBGE Co´digo da cidade de resideˆncia do paciente CATEATEND Categoria de atendimento ao diagno´stico

DIAGPREV Diagno´stico e tratamento anterior

BASEDIAG Co´digo da base do diagno´stico EC Estadio clinico

G Classificac¸a˜o TNM – G (Grau)

PSA Classificac¸a˜o TNM - PSA

GLEASON Classificac¸a˜o TNM - Gleason

TRATHOSP combinac¸a˜o dos tratamentos realizadosno hospital

NENHUM Tratamento recebido no hospital = nenhum

CIRURGIA Tratamento recebido no hospital = cirurgia RADIO Tratamento recebido no hospital = radioterapia

QUIMIO Tratamento recebido no hospital = quimioterapia HORMONIO Tratamento recebido no hospital = hormonioterapia TMO Tratamento recebido no hospital = tmo IMUNO Tratamento recebido no hospital = imunoterapia

OUTROS recebido no hospital = outros

NENHUMANT Nenhum tratamento recebido fora do hospital e antes da admissa˜o

ULTINFO

U´ ltima informac¸a˜o sobre o paciente

CONSDIAG Diferenc¸a em dias datas de consulta o diagno´stico

TRATCONS Diferenc¸a em dias datas de consulta e tratamento

DIAGTRAT Diferenc¸a em dias datas de tratamento e diagno´stico

ANODIAG Ano de diagno´stico

DRS Departamento Regional de Sau´de

RRAS Rede Regional de Atenc¸a˜o a` Sau´de

RECNENHUM Sem recidiva

IBGEATEN Co´digo IBGE da instituic¸a˜o

ULTICONS Diferenc¸a de dias entre u´ltima informac¸a˜o e consulta ULTIDIAG Diferenc¸a de dias entre u´ltima informac¸a˜o e diagno´stico ULTITRAT Diferenc¸a de dias entre u´ltima informac¸a˜o e tratamento

Seguindo a linha de Buk Cardoso et al. tambe´m foram exclu´ıdos pacientes cujo valor do campo ECGRUP fosse igual a *x* ou *y*. Estes valores indicam, respectivamente, casos em que o tumor prima´rio, linfonodos regionais ou meta´stases na˜o possam ser ava- liados pelo exame f´ısico ou exames complementares, ou estadiamento feito durante ou apo´s o tratamento. Logo, pacientes nestas categorias na˜o colaborariam para uma ana´lise de sobreviveˆncia precisa. Por fim, foram exclu´ıdos todos os pacientes cuja escolaridade na˜o teria sido informada. Tal abordagem difere da utilizada por Buk Cardoso et al., que preferiu utilizar aprendizado de ma´quina para estimar a escolaridade de pacientes que na˜o

a haviam informado.

Ao final da preparac¸a˜o dos dados, restaram 44094 registros de pacientes que sera˜o utilizados para treinamento dos modelos.

Para este trabalho, foram treinados e avaliados três modelos diferentes. Na exploração de dados foram criadas novas colunas que dizem respeito a sobrevivencia ou não dos pacientes em determinadas faixas de tempo, a partir do diagnóstico. São elas: vivo\_ano1, vivo\_ano3 e vivoano\_5, o dataset de exploração de dados entra em mais detalhes de cada uma delas. Neste contexto vamos trabalhar com os modelos para prever a coluna vivo\_ano5.

Antes da aplicação de cada modelo, mais etapas de pre-processamento são realizadas a partir de funções importadas do arquivo “functions.py”.Estas etapas podem incluir:

- Filtragem por intervalo de anos, no caso a coluna vivo\_ano5 que escolhemos.

- Divisão Treinamento/teste

- Codificação e Normalização dos dados de treino e teste

- Balanceamento: Usa SMOTE para balancear os dados de treinamento.

- Seleção de features,etc...

Para cada modelo, Naive Bayes, Random Forest e XGBoost realizamos o treinamento com base nos dados selecionados da coluna vivo\_ano5. Para validação utilizamos matriz de confusão, curva ROC para conjunto de treino e teste e também uma análise final das features mais importantes, em alguns casos com a biblioteca e o gráfico SHAP. Também utilizamos a biblioteca Optuna do python para busca por melhores hiperâmetros nos modelos Random Forest e XGBoost.

1. Resultados

resultados para cada modelo estão apresentados pós-otimização hiperparâmetros, no notebook temos o processo completo:

-Naive Bayes:

Matriz de confusão:

Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente

Uma imagem contendo ao ar livre, relógio, cidade, medidor

Descrição gerada automaticamente

Curva ROC: Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Analisando as curvas não houve overfitting, mas este modelo se apresentou inadequado para previsão da classe 1, desta forma, decidimos que qualquer análise além destas não possuem tanta relevância.

-Random Forest:

Matriz de confusão:

Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente

Calendário

Descrição gerada automaticamente

Curva ROC:

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Importância das features:

Gráfico, Gráfico de barras

Descrição gerada automaticamente

XGBoost:

Matriz de confusão:Gráfico, Gráfico de mapa de árvore

Descrição gerada automaticamente

Tela de celular

Descrição gerada automaticamente

Curva ROC: Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Importância das features:

Gráfico, Gráfico de barras

Descrição gerada automaticamente

1. Conclusão

Podemos concluir que o modelo Naïve Bayes se mostrou inadequado para este estudo. Quando analisamos os resultados obtidos pelos outros modelos, obtivemos valores aceitáveis nas métricas da matriz de confusão. Os valores de Accuracy atingiram o esperado com base no proposto inicialmente, entretanto com o desenvolvimento do projeto foi também possível notar que eles podem e devem ser melhorados. Como este trabalho será aprimorado ainda mais no futuro, é importante revisar a exploração e preparação dos dados a fim de tentar melhorar estas métricas.

Podemos observar que o melhor modelo obtido foi o Random Forest pois, apesar de ser possível obter uma accuracy maior com o XGBoost, o Random Forest apresenta um equilibrio maior para prever os valores de classe 1, tornando este modelo mais adequado.

Nos próximos passos também vamos aplicar os mesmos modelos realizados aqui para todas as outras colunas “vivo” criadas.

1. Links de acesso:

Todas as modificac¸o˜es feitas no dataset bem como o diciona´rio de dados, o note- book Jupyter contendo a ana´lise explorato´ria, a preparac¸a˜o dos dados e aplicação e validação dos modelos estão disponíveis no seguinte reposito´rio do GitHub. A Segunda parte do projeto com a conclusão do trabalho e os notebooks executados se encontra na pasta “Parte 2”

https://github.com/acanellafilho/ia

*Nota: Foi necessa´rio comprimir o dataset, pois o mesmo ultrapassava o limite de 25mb imposto pelo github.*

Youtube com explicação:

https://www.youtube.com/watch?v=97VI90zBh0E

# Refereˆncias Bibliogra´ficas

Buk Cardoso, L., Cunha Parro, V., Verzinhasse Peres, S., Curado, M. P., Fer- nandes, G. A., Wu¨nsch Filho, V., and Natasha Toporcov, T. (2023). Machine learning for predicting survival of colorectal cancer patients. Dispon´ıvel em: [http://dx.doi.org/10.1038/s41598-023-35649-9.](http://dx.doi.org/10.1038/s41598-023-35649-9) Acesso em:5 mai.2024.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Dis- pon´ıvel em: https://xgboost.readthedocs.io/en/stable/tutorials/model.html. Acesso em: 20 out.2024.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Par- kin, D. M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in globocan 2012. Dispon´ıvel em: https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.29210. Acesso em: 05 mai.2024.

FOSP (2022). Banco de dados do rhc. Dispon´ıvel em:https://fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e- epidemiologia/rhc-registro-hospitalar-de-cancer/banco-de-dados-do-rhc. Acesso em:5 mai.2024.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international confe- rence on document analysis and recognition*, volume 1, pages 278–282. IEEE.

INCA (2022). Estimativa 2023: Incideˆncia de caˆncer no brasil. Dispon´ıvel [em:https://www](http://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-).inca.gov[.br/publicacoes/livros/estimati](http://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-)v[a-2023-incidencia-de-cancer](http://www.inca.gov.br/publicacoes/livros/estimativa-2023-incidencia-de-cancer-)- no-brasil. Acesso em: 16 out.2024.

James, N. D. e. a. (2024). The lancet commission. Dispon´ıvel em: [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)00651-](http://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)00651-) 2/fulltext”. Acesso em: 23 out.2024.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine lear- ning in Python. ”https://scikit-learn.org/1.5/modules/naive*bayes.html*”*. Acessoem* : 27*out.*2024*.*

Suresh, K., S. C. . G. D. (2022). Survival prediction models: an introduction to discrete-time modeling. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022- 01679-6citeas. Acesso em: 27 out.2024