Sistemas Inteligentes para a Bioinformática

Mestrado em Bioinformática

Docente Miguel Rocha

29 de janeiro de 2021

Cátia Mendes (pg30435), Joana Ribeiro (pg38272), Maria João Lopes (pg40965), Miguel Castro (pg40970)

Análise de *datasets* de expressão relacionados com COVID-19

1. Contexto e Motivação

A apresentação clínica da infeção por COVID-19 é muito diversa, podendo variar de um estado assintomático para uma condição letal. Dados recentes indicam que a severidade da doença depende, para além de fatores virais, de fatores associados ao hospedeiro (Zhang et al., 2020a, 2020b, 2020c). Diferentes assinaturas genéticas (Ellinghaus et al., 2020), fisiológicas (Gattinoni et al., 2020), patológicas (Fox et al., 2020) e clínicas (Richardson et al., 2020) parecem conseguir diferenciar doentes com diagnóstico de COVID-19 e doentes sem esta condição.

2. Objetivos

O objetivo deste trabalho consiste em tentar compreender melhor as respostas individuais, a nível molecular, que possam explicar estas diferenças clínicas observadas. Adicionalmente, pretende-se criar modelos de *machine learning* e *deep learning* que permitam prever o estado e a severidade da infeção COVID-19 com base em dados ómicos, fisiológicos e clínicos. Para tal, recorreu-se a um *dataset* composto por diversos dados clínicos pertencentes a 128 doentes, com e sem diagnóstico de COVID-19. Adicionalmente, foram utilizados dados de *RNA-sequencing* e dados laboratoriais provenientes de amostras de sangue recolhidas destes doentes (Overmyer et al., 2020).

Pretende-se, neste trabalho, responder concretamente às seguintes questões:

- Determinar a existência de associação entre biomarcadores ou perfil molecular com diagnóstico de COVID-19 e severidade da apresentação clínica. Desta forma pretende-se compreender que fatores estão associados a uma maior severidade da apresentação clínica desta doença;
- Perceber que vias metabólicas se encontram associados à resposta do hospedeiro ao SARSCoV-2 e à severidade da doença no sentido de elucidar potenciais *targets* terapêuticos;
- Prever a severidade da doença e *outcome* dos pacientes com diagnóstico de COVID-19 com base em dados ómicos, clínicos e laboratoriais disponíveis por forma a priorizar doentes e seus tratamentos.

3. *Dataset*

O *dataset* utilizado encontra-se disponível na base de dados *GEO* com o número de acesso GSE157103. Consiste em dados clínicos recolhidos de 102 doentes com diagnóstico COVID-19 e 26 doentes com dificuldades respiratórias sem diagnóstico COVID-19. Fazem parte deste conjunto de dados, medições provenientes de análises laboratoriais e genómicas realizadas a amostras de sangue recolhidas destes doentes (Overmyer et al., 2020).

Dos dados clínicos disponíveis fazem parte as seguintes variáveis:

- necessidade de internamento em unidade de cuidados intensivos (IUC)
- 4 scores de severidade da apresentação clínica da doença: *hospital-free days at day 45* (HFD-45); *acute psysiologic assessment and chronic health evaluation* (APACHE-II); *sequential organ failure assessment* (SOFA); *Charlson comorbidity index*
- Número de dias ligado a ventilação mecânica;

Medições provenientes de análises laboratoriais:

- *C-reactive protein* (CRP)
- *D-dimer*
- ferritina
- lactato
- procalcitonina (PCT)
- fibrinogénio

Os *scores* de severidade foram utilizados para integrar diversos fatores. Por exemplo o HFD-45 integra a duração da estadia no hospital e a mortalidade, e é atribuído um *score* com valor zero a doentes que permaneceram internados mais de 45 dias ou morreram durante o internamento, e valores de *score* mais elevados a doentes com internamentos hospitalares de menor duração e severidade da doença menor. Já o APACHE-II utiliza um *score* baseado nos valores iniciais (primeiras 24 horas nos IUC) de 12 medidas fisiológicas de rotina, idade, e estado de saúde prévio (Ferreira et al., 2001; Knaus et al., 1985). O SOFA correlaciona a falha ou disfunção dos órgãos durante a estadia total nos IUC e a mortalidade (Ferreira et al., 2001). Por fim, o Charlson classifica a comorbidade para prever a mortalidade a curto e a longo prazo, havendo diferentes *scores* para diferentes condições presentes entre diversos diagnósticos secundários (D'Hoore et al., 1996).

Como mencionado, também foram utilizados diversos biomarcadores nas diferentes análises laboratoriais. A CRP é um biomarcador inflamatório utilizado para avaliar o risco cardiovascular (Genest, J., 2010). O *D-dimer* também tem um papel na área da prevenção cardiovascular, permitindo diagnosticar doenças como *venous thromboembolism, disseminated intravascular coagulation, coronary and carotid artery atherosclerosis and aortic diseases* (Soomro et al., 2016). A ferritina é utilizada como biomarcador em casos em que se suspeita de deficiência em ferro, mas também se manifesta no caso de doenças inflamatórias (Kappert et al., 2020). O lactato é frequentemente utilizado para avaliar a severidade da doença como prognóstico (Okorie et al., 2011). A PCT é um biomarcador de doenças infeciosas, visto que é produzido em resposta à invasão de bactérias patogénicas, fungos e alguns parasitas (Lee, H., 2013). Por fim, o fibrinogénio está associado ao risco de desenvolver *chronic obstructive pulmonary disease*, severidade da doença, progressão e mortalidade (Duvoix et al., 2013).

O acesso a estes dados foi conseguido através do package em *pyhton GEOparse* [1]. Os dados de expressão já se encontram normalizados (*transcripts per kilobase million, TPM*) e são de *RNA-sequencing* (*Illumina NovaSeq6000*) de amostras de sangue destes doentes. A matriz de contagem de genes por amostra foi descarregada diretamente do site GEO [2].

4. Métodos

4.1 – Análise Exploratória e Análise estatística dos dados

O *dataset* utilizado encontra-se disponível na base de dados *GEO* com o número de acesso GSE157103. Consiste em dados clínicos recolhidos de 102 doentes com diagnóstico COVID-19 e 26 doentes com dificuldades respiratórias sem diagnóstico COVID-19.

Inicialmente procedeu-se a uma análise exploratória dos metadados existentes no *dataset* para estudar melhor as variáveis existentes nele. Desta forma, o estudo focou-se nas seguintes variáveis: *Disease state, HFD-45, ICU, mechanical ventilation, Severity Scores* e, por fim, níveis de biomarcadores moleculares nas amostras recolhidas dos pacientes. Foram feitas algumas análises estatísticas, recorrendo-se a métodos de inferência, para complementar as tendências verificadas no decorrer da exploração dos dados.

4.2 – Análise não Supervisionada

De modo a compreender melhor a estrutura global dos dados recorreu-se a métodos não supervisionados de estatística multivariada. Numa fase inicial, realizou-se uma análise das componentes principais com o intuito de reduzir a dimensionalidade dos dados e identificar conjuntos de variáveis não correlacionadas entre si, que explicam a maior parte da variabilidade dos dados. Além disso, procedeu-se ao agrupamento dos dados com base no seu grau de semelhança, para tal foram utilizadas duas abordagens, o *clustering* hierárquico e o *k-means.*

4.3 – Expressão Diferencial

Com o objetivo de estudar diferenças significativas nos transcritos associados a doentes COVID-19 e não COVID-19, utilizou -se o package *rpy2* (este combina a linguagem *R* e *python)*. Primeiramente as *read counts* foram lidas a partir de um *CSV file* para um dicionário. Criou-se uma *count matrix* que serviu como *input* no package *edgeR*, utilizado para análise de expressão diferencial.

4.3 – Machine Learning

Tendo em conta os resultados obtidos selecionamos três variáveis para o estudo de previsão. De modo a prever e a classificar os dados do estudo, foram desenvolvidos e avaliados modelos de *machine learning* para os seguintes tipos de variáveis:

- Variáveis qualitativas - Estado de Saúde (binária)
- Índices de severidade - Apache II (numérica e *multiclass*)
- Biomarcadores - Ferritina (numérica)

4.3.1 – *Disease State*

Com o rápido incremento do número de casos a serem testados diariamente para o Coronavírus, é impossível a sua execução devido aos fatores tempo e custo. Nos últimos anos, *machine learning* tornou-se muito confiável na área médica. O uso de *machine learning* para prever COVID-19 em pacientes reduzirá o tempo de espera para os resultados dos testes médicos e modulará os profissionais de saúde para que lhes forneçam tratamento médico adequado.

O objetivo principal desta análise é desenvolver um modelo de *machine learning* que possa prever se um paciente apresenta COVID-19. Para criar os modelos, dividimos os dados ómicos fornecidos num conjunto de dados para treinar e testar, onde o primeiro corresponde a 1/5 dos dados completos.

Também selecionamos 30% das *features* do conjunto de dados para reduzir a dimensionalidade e, possivelmente, melhorar os modelos testados.

Os modelos selecionados nesta análise foram:

- *Logistic Regression*
- *K-Nearest Neighbors (KNN)*
- *Support Vector Machine (SVM)*
- *Naive Bayes*

- *Random Forest*
- *Decision Tree*

Para cada modelo foi realizada a validação cruzada e obteve-se as métricas resultantes da previsão do conjunto de dados do teste.

Além disso, realizou-se o *Receiver Operating Characteristics* (ROC), *Precision-Recall* e respetiva área sob a curva (AUC).

Esses resultados foram representados graficamente na curva ROC e na curva *Precision-Recall* de cada modelo.

Nos casos necessários realizou-se a otimização de híper parâmetros. A otimização ou ajuste de híper parâmetros permite escolher um conjunto de híper parâmetros ideais para um algoritmo. Um híper parâmetro é um parâmetro cujo valor é usado para controlar o modelo.

Por fim realizou-se o *ensemble* dos 3 melhores modelos.

4.3.2 – APACHE II – Numérico

O principal objetivo desta análise é desenvolver um modelo de *machine learning* com a capacidade de prever a severidade de COVID-19 no início da admissão dos pacientes, de modo a auxiliar os profissionais de saúde a priorizar os pacientes para os tratamentos. Para criar os modelos, dividimos os dados ómicos fornecidos num conjunto de dados para treinar e testar, onde o primeiro corresponde a 1/5 dos dados completos.

Também selecionamos 30% das *features* do conjunto de dados para reduzir a dimensionalidade e, possivelmente, melhorar os modelos testados.

Os modelos selecionados nesta análise foram:

- *Linear Regression*
- *Neural Network Regression*
- *Support Vector Regression*
- *Random Forest Regression*
- *Decision Tree Regression*

Para cada modelo foi realizada a validação cruzada e obteve-se as métricas resultantes da previsão do conjunto de dados do teste.

Além disso, realizou-se R^2 (*coefficient of determination*), *explained Variance Scores, Median Absolute Error* (MAE) e o *Median Squared Error* (MSE).

Para cada modelo implementou-se o *over-sampling* de modo a melhorar os resultados obtidos.

4.3.3 – APACHE II – *Multiclass*

De modo a obter melhores resultados, convertemos a variável APACHE II numa variável *multiclass*. Em que, G_1 inclui valores menores que 9, os valores G_2 estão entre 10-19, os valores G_3 entre 20-29 e os valores G_4 são os maiores que 30.

Os modelos selecionados nesta análise foram:

- *Logistic Regression*
- *K-Nearest Neighbors (KNN)*
- *Support Vector Machine (SVM)*
- *Naive Bayes*
- *Random Forest*
- *Decision Tree*

Para cada modelo foi realizada a validação cruzada e obteve-se as métricas resultantes da previsão do conjunto de dados do teste.

4.3.4 – Ferritina – Numérico

O principal objetivo desta análise é desenvolver um modelo de *machine learning* com a capacidade de prever a severidade de COVID-19 no início da admissão dos pacientes de modo a auxiliar os profissionais de saúde a priorizar os pacientes para os tratamentos. Para criar os modelos, dividimos os dados ómicos fornecidos num conjunto de dados para treinar e testar, onde o primeiro corresponde a 1/5 dos dados completos.

Também selecionamos 30% das *features* do conjunto de dados para reduzir a dimensionalidade e, possivelmente, melhorar os modelos testados.

Os modelos selecionados nesta análise foram:

- *Linear Regression*
- *Neural Network Regression*
- *Support Vector Regression*
- *Random Forest Regression*
- *Decision Tree Regression*

Para cada modelo foi realizada a validação cruzada e obteve-se as métricas resultantes da previsão do conjunto de dados do teste. Além disso, realizou-se R^2, MAE e o MSE.

Para cada modelo implementou-se o *over-sampling* de modo a melhorar os resultados obtidos.

4.4 – *Deep Learning*

Desenvolveram-se modelos de *deep learning* na tentativa de obter uma precisão mais elevada e mais eficiente, em comparação com os modelos desenvolvidos de *machine learning*

Numa primeira fase, os dados foram divididos em conjuntos de dados treino e teste de acordo com a formulação estabelecida em *machine learning*.

Para o *endpoint disease state* do doente, o *Multi-Layer Perceptron* (MLP) e *Deep Neural Network* (DNN) foram os modelos selecionados para esta análise.

Para o primeiro, foi realizada a validação cruzada e obtiveram-se as métricas resultantes da previsão do conjunto de dados do teste. Para o DNN realizou-se a otimização de híper parâmetros.

No caso dos *endpoints* numéricos APACHE II e Ferritina, pretendeu-se desenvolver um modelo que conseguisse prever a severidade de COVID-19, em que o modelo selecionado foi o DNN. Obtiveram-se as métricas resultantes da previsão do conjunto de dados de teste. Adicionalmente, foram construídos gráficos de *Training Loss and Validation Loss* para ajudar a identificar a dinâmica dos modelos.

5. Resultados

Os resultados encontram-se disponíveis e comentados no *GitHub* cujo link está referenciado [3].

6. Discussão e conclusões

6.1 – Análise exploratória e análise estatística dos dados

Tendo em conta os resultados obtidos na exploração dos dados verificou-se uma predominância dos pacientes de sexo masculino em relação aos do sexo feminino; aproximadamente metade dos pacientes foram submetidos nos ICU, 72% dos pacientes são COVID-19 positivos e que os homens são mais suscetíveis para a necessidade de ventilação mecânica. Verifica-se também que há maior número de pacientes do sexo masculino com COVID-19, indicando que o sexo masculino poderá ser mais suscetível a ser contagiado pelo vírus.

De seguida, procedeu-se a análise à influência do estado COVID nos internamentos em cuidados intensivos. Em que não se verificaram grandes diferenças entre estes grupos, o que pode ser justificado pela pequena dimensão do *dataset*.

Contudo, existe uma correlação entre a necessidade de ICU, com a necessidade de ventilação mecânica nos pacientes. A grande maioria que esteve internado em ICU necessitou de ventilação mecânica.

Relativamente à variável quantitativa idade, observou-se que mais de 50% dos pacientes tem uma idade igual ou superior a 70 anos. Não se verificaram diferenças estatisticamente significativas nas idades entre diferentes grupos (por sexo e por estado COVID ou não COVID).

Quanto aos *scores* de severidade, estes também revelaram ter resultados pertinentes. No caso do *score* HFD-45, o valor médio é menor nos casos em que o diagnóstico COVID-19 é positivo. Este resultado é consistente com a definição do *score*, uma vez que valores de *score* mais baixos correspondem a situações mais graves ou letais.

Através da análise estatística dos *scores* foi possível concluir que no grupo de doentes em UCI, as diferenças não são significativas. Mas, quando comparando doentes com diagnóstico COVID-19 positivo e negativo, o mesmo não se observa para o *score Charlson*, onde já se encontra uma diferença significativa entre os 2 grupos.

Em relação aos biomarcadores medidos nas análises ao sangue dos pacientes foi possível observar diferenças nos valores médios e medianas para os diferentes estados de saúde. Por exemplo, ao observar o biomarcador para a degradação do fibrinogénio, o *D-dimer*, este está consistentemente elevado no caso de doentes COVID. A análise estatística aos biomarcadores permite concluir que a ferritina, CRP, lactato e fibrinogénio exibem diferentes medianas comparando casos COVID positivo e negativo, aquando testados estatisticamente.

De seguida, procedeu-se à comparação dos biomarcadores entre os 2 grupos de doentes, tendo em conta a necessidade de cuidados intensivos. A análise estatística permitiu concluir que a ferritina, o fibrinogénio e o lactato apresentam diferenças estatísticas significativas nos doentes em UCI para os grupos COVID e não COVID.

6.2 – Análise não supervisionada

De seguida, procedeu-se à análise não supervionada com inclusão dos dados de *RNA seq*. Primeiramente foram realizadas as análises aos componentes principais das variáveis presentes nos metadados. No caso das variáveis qualitativas, o *disease state*, o internamento em ICU e ventilação mecânica foi possível discriminar facilmente uma ligeira separação dos diferentes atributos testados.

Relativamente aos *scores* de severidade realizou-se a categorização dos valores. A categoria *A* corresponde aos valores menores de *score* e *B*, *C* e *D* aos maiores valores, respetivamente. Neste caso nas variáveis APACHEII e SOFA foi possível discriminar mais facilmente uma ligeira separação das categorias.

O mesmo raciocínio foi aplicado nos biomarcadores. Neste caso apenas nas variáveis *D-dimer* e lactato se deteta uma separação nas diferentes categorias.

De seguida foi realizado o *clustering* hierárquico para as variáveis qualitativas. Relativamente ao *disease state* é visível alguns clusters exclusivamente com COVID-19 e não COVID. No caso dos ICU também é visível uma separação em *clusters* constituídos quase na totalidade por um só atributo *yes* ou *no*. O mesmo se verifica na ventilação mecânica.

Para além do *clustering* hierárquico, determinou-se o *k-means* para as mesmas variáveis. E no caso das variáveis *disease state* e ventilação mecânica o *cluster* 1 é constituído maioritariamente por um dos perfis.

6.3 – Expressão diferencial

Quanto à análise de DE não foi possível concluir sobre o número de genes significativamente associados com o estado de saúde COVID-19 nem a sua identidade. Um objetivo resultante desta análise seria a realização de uma análise de

enriquecimento *Gene Ontology* (GO), de modo a identificar funções ou vias metabólicas que estariam associados a esses genes.

Contudo, ainda se recorreu a um script com o objetivo de utilizar *Z-scores* para avaliar os genes diferencialmente expressos, mas igualmente sem sucesso.

6.4 – *Machine Learning*

A análise não supervisionada permitiu verificar quais as variáveis em análise, cujas categorias podem ser discriminadas mais facilmente e direcionou o raciocínio na exploração de dados futura. Tendo em conta os resultados obtidos selecionamos 3 variáveis para o estudo de previsão.

De modo a prever e a classificar os dados do estudo, foram desenvolvidos e avaliados modelos de *machine learning* para os seguintes tipos de variáveis:

No caso das variáveis qualitativas: *disease state* (binária).

No caso dos índices de severidade: Apache II (numérica e multiclass).

No caso dos biomarcadores: Ferritina (numérica).

6.4.1 - Disease State

Com o rápido incremento do número de casos a serem testados diariamente para o Coronavírus, é impossível a sua execução devido aos fatores tempo e custo. Nos últimos anos, *machine learning* tornou-se muito confiável na área médica. O uso de *machine learning* para prever COVID-19 em pacientes reduzirá o tempo de espera para os resultados dos testes médicos e modulará os profissionais de saúde para que lhes forneçam tratamento médico adequado.

O objetivo principal desta análise é desenvolver um modelo de *machine learning* que possa prever se um paciente apresenta COVID-19.

A análise de previsões individuais dos distintos modelos permitiu identificar que a maioria prevê corretamente o estado de saúde dos pacientes.

No caso do *F1 score,* os valores variam entre 0.895 e 1.000, o que demonstra a precisão e *recall* do teste.

A precisão é o número de resultados positivos identificados corretamente dividido pelo número de todos os resultados positivos, incluindo aqueles não identificados corretamente. O *recall* é o número de resultados positivos identificados corretamente dividido pelo número de todas as amostras que deveriam ter sido identificadas como positivas. Desta forma, o uso de uma combinação dos dois é útil ao analisar observações desequilibradas.

Relativo ao ROC, *Precision-Recall* e AUC, foram realizadas as curvas ROC e a *Precision-Recall* de cada modelo.

No caso do ROC AUC, os valores variam entre 0.846 e 1.000 e os valores da *Precision-Recall* AUC entre 0.935 e 1.000. Quanto mais elevados os valores da AUC, melhor é o modelo, o que implica que os modelos apresentados realizaram na sua maioria previsões corretas.

No caso da curva de precisão-*recall*, que mostra a precisão/*recall* dos modelos de classificação em comparação com um *no skill*: também se verificou que os modelos Regressão Logística e SVM apresentam a maior área sob a curva de precisão-*recall*, significando um modelo melhor.

No caso do modelo SVM também implementamos a otimização de híper parâmetros para melhorar nossos resultados. No entanto, nenhuma melhoria foi observada.

Diante disso, determinamos que o modelo de Regressão Logística possui o maior valor de AUC em ROC e *Precission-Recall* e também possui o maior valor de F1, quando comparado aos outros modelos. Além disso, as métricas deste modelo são todas de 1.00, então foi possível prever corretamente todo o estado da doença dos pacientes.

Portanto, acredita-se que o modelo de regressão logística é o melhor modelo para o objetivo enunciado.

Por fim realizou-se o *ensemble* dos 3 melhores modelos: Regressão Logística, SVM e Árvore de Decisão. O modelo *Ensemble* apresenta os mesmos resultados que o modelo de Regressão Logística.

6.4.2 - APACHE II - Numérico

O principal objetivo desta análise é desenvolver um modelo de *machine learning* com a capacidade de prever a severidade de COVID-19 no início da admissão dos pacientes de modo a auxiliar os profissionais de saúde a priorizar os pacientes para os tratamentos.

Para cada modelo foi realizada a validação cruzada e obteve-se as métricas resultantes da previsão do conjunto de dados do teste. Além disso, realizou-se R^2, MAE e o MSE.

Nos modelos LR, SVR e DTR obteve-se scores R^2 negativos, o que significa que os modelos não seguem a tendência dos dados. Nos demais modelos os valores de R^2 variam entre 0.129 e 0.537. O coeficiente de determinação determina o quão bem o modelo se ajusta aos dados. Neste caso o melhor modelo é o *Random Forest Regression*. Para cada modelo implementou-se o *over-sampling* de modo a melhorar os resultados obtidos. No entanto, nenhuma melhoria foi observada.

Os *Explained Variance Scores* medem o quanto os modelos têm em consideração a variância do conjunto de dados. No caso de LR e DTR os valores são negativos, portanto os modelos são inespecíficos. No NNR e no SVR também são muito baixos, com

baixa especificação. O modelo com o melhor resultado é *Random Forest Regression* com um valor de 0.568.

O MAE é uma quantidade usada para medir o quão próximas as previsões estão dos resultados reais. O erro absoluto médio é uma média de todos os erros absolutos. O erro médio absoluto é uma medida comum de erro de estimativa na análise de séries temporais. O erro quadrático médio de um estimador mede a média dos quadrados dos erros, o que significa a diferença entre o estimador e o estimado. O MSE é uma medida da qualidade de um estimador, é sempre positivo e os valores mais próximos de zero são melhores.

Os valores de MAE e MSE para os modelos são muito altos, o que significa que os erros de previsão são elevados, logo as previsões não são relevantes.

Tendo em conta os resultados obtidos o modelo com os melhores resultados é o modelo *Random Forest Regression*, mas este e os outros modelos de regressão não se ajustam com sucesso aos dados. Uma possível razão para os resultados podem ser os poucos exemplos de treino, *overfitting*, desequilíbrio de dados, etc.

6.4.3 – APACHE II – *Multiclass*

De modo a obter melhores resultados, convertemos a variável APACHE II em uma variável *multiclass*. Contudo, verificou-se que o modelo apresentado não permitiu a melhoria das métricas, logo a previsão do modelo não melhorou.

6.4.4 – Ferritina

A ferritina é um mediador chave da desregulação imunológica, especialmente sob hiperferritinemia extrema, por meio de efeitos imunossupressores e pró-inflamatórios diretos, contribuindo para a cascata de citocinas. Foi relatado que resultados fatais por COVID-19 são acompanhados pela síndrome da cascata de citocinas, portanto, foi sugerido que a severidade da doença é dependente da expressão de ferritina.

Muitos pacientes com diabetes apresentam níveis elevados de ferritina sérica e sabe-se que eles têm maior probabilidade de apresentar complicações graves com COVID-19. Com base nisso, revisamos brevemente as evidências que sustentam a hipótese de que os níveis de ferritina podem ser um fator crucial que influencia a gravidade da COVID-19.

O principal objetivo desta análise é desenvolver um modelo de *machine learning* com a capacidade de prever a severidade de COVID-19 no início da admissão dos pacientes, de modo a auxiliar os profissionais de saúde a priorizar os pacientes para os tratamentos.

Nos modelos LR, SVR e DTR obteve-se scores R^2 negativos, o que significa que os modelos não seguem a tendência dos dados. Nos restantes modelos os valores de R^2 variam entre 0.129-0.493. Neste caso o melhor modelo é o *Random Forest Regression*. Para cada modelo implementou-se o *over-sampling* de modo a melhorar os resultados obtidos. No entanto, nenhuma melhoria foi observada.

Os *Explained Variance Scores* medem o quanto os modelos levam em consideração a variância do conjunto de dados. No caso de LR e DTR os valores são negativos, portanto os modelos são inespecíficos. No NNR e no SVR também são muito baixos, com baixa especificação. O modelo com o melhor resultado é *Random Forest Regression* com um valor de 0.172.

Os valores de MAE e MSE para os modelos são muito altos, o que significa que os erros de previsão são elevados, logo as previsões não são relevantes.

Tendo em conta os resultados obtidos o modelo com os melhores resultados é o modelo *Random Forest Regression*, mas este e os outros modelos de regressão não se ajustam com sucesso aos nossos dados. Uma possível razão para os resultados podem ser os poucos exemplos de treino, *overfitting*, desequilíbrio de dados, etc.

6.5 – *Deep Learning*

6.5.1 – *Disease State*

Relativamente à variável *disease state* recorreu-se ao *MLPClassifier* para classificar o estado de saúde dos doentes em COVID-19 e não COVID-19 (*output* binário), com base nos dados de *RNA-seq*. Este modelo apresentou um *mean score* de 61%, ou seja, é um modelo capaz de classificar corretamente 61% dos doentes num dos seus estados de saúde.

Quando utilizado o modelo DNN, foi obtido um valor de *accuracy* de 72% para a mesma classificação categórica. Tendo em conta, o gráfico *train* and *validation loss*, verifica-se uma tendência de *overfitting*: a partir de *ephoc=2*, os valores de *validation loss* tendem a oscilar, enquanto que os do treino tendem a descer.

Para tentar encontrar um equilíbrio entre *bias* e variância, no sentido de diminuir o *overfitting* aos dados de treino, realizou-se a otimização dos híper parâmetros do modelo. Verificou-se uma redução ligeira com a otimização dos híper parâmetros do modelo para 68% relativamente ao *baseline model*.

6.5.2 – APACHE II – Numérico

No caso da variável APACHE II realizou-se o modelo DNN, em que se obteve um MAE de 6 ao prever os dados de teste, o que significa que o modelo, em média, falha a previsão do *score* APACHE II em 6 pontos.

A magnitude elevada deste erro, para este *endpoint*, faz com que as previsões feitas pelo modelo não sejam relevantes o suficiente para a sua utilização.

6.5.3 – Ferritina

Finalmente, foi construído um modelo para a variável Ferritina. O modelo selecionado foi o DNN, com um resultado de MAE de 799 ao prever os dados de teste, o que significa que o modelo, em média, falha a previsão do valor deste biomarcador em 799ng/ml.

A magnitude elevada deste erro, para este *endpoit*, faz com que as previsões feitas pelo modelo não sejam relevantes o suficiente para a sua utilização.

Referências

D'Hoore, W., Bouckaert, A., & Tilquin, C. (1996). Practical considerations on the use of the Charlson comorbidity index with administrative data bases. Journal of clinical epidemiology, 49(12), 1429-1433.

Duvoix, A., Dickens, J., Haq, I., Mannino, D., Miller, B., Tal-Singer, R., & Lomas, D. A. (2013). Blood fibrinogen as a biomarker of chronic obstructive pulmonary disease. Thorax, 68(7), 670-676.

Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., Ferna´ ndez, J., Prati, D., Baselli, G., Asselta, R., et al. (2020). Genomewide association study of severe Covid-19 with respiratory failure. N. Engl. J. Med.

Ferreira, F. L., Bota, D. P., Bross, A., Mélot, C., & Vincent, J. L. (2001). Serial evaluation of the SOFA score to predict outcome in critically ill patients. Jama, 286(14), 1754-1758.

Fox, S.E., Akmatbekov, A., Harbert, J.L., Li, G., Quincy Brown, J.Q., and Vander Heide, R.S. (2020). Pulmonary and cardiac pathology in African American patients with COVID-19: an autopsy series from New Orleans. Lancet Respir. Med. 8, 681–686.

Gattinoni, L., Coppola, S., Cressoni, M., Busana, M., Rossi, S., and Chiumello, D. (2020). COVID-19 does not lead to a 'typical' acute respiratory distress syn- drome. Am. J. Respir. Crit. Care Med. 201, 1299–1300.

Genest, J. (2010). C-reactive protein: risk factor, biomarker and/or therapeutic target?. Canadian Journal of Cardiology, 26, 41A-44A.

Kappert, K., Jahić, A., & Tauber, R. (2020). Assessment of serum ferritin as a biomarker in COVID-19: bystander or participant? Insights by comparison with other infectious and non-infectious diseases. Biomarkers, 1-10.

Knaus, W. A., Draper, E. A., Wagner, D. P., & Zimmerman, J. E. (1985). APACHE II: a severity of disease classification system. Critical care medicine, 13(10), 818-829.

Lee, H. (2013). Procalcitonin as a biomarker of infectious diseases. The Korean journal of internal medicine, 28(3), 285.

Okorie, O. N., & Dellinger, P. (2011). Lactate: biomarker and potential therapeutic target. Critical care clinics, 27(2), 299.

Overmyer, K. A., Shishkova, E., Miller, I. J., Balnis, J., Bernstein, M. N., Peters-Clarke, T. M., ... & Jaitovich, A. (2020). Large-scale multi-omic analysis of COVID-19 severity. Cell systems.

Richardson, S., Hirsch, J.S., Narasimhan, M., Crawford, J.M., McGinn, T., Davidson, K.W., the Northwell COVID-19 Research Consortium, Barnaby, D.P., Becker, L.B., Chelico, J.D., et al. (2020). Presenting characteristics, co-morbidities, and outcomes Among 5700 patients hospitalized with COVID-19 in the New York City area. JAMA 323, 2052–2059.

Soomro, A. Y., Guerchicoff, A., Nichols, D. J., Suleman, J., & Dangas, G. D. (2016). The current role and future prospects of D-dimer biomarker. European Heart Journal-Cardiovascular Pharmacotherapy, 2(3), 175-184.

Zhang, X., Tan, Y., Ling, Y., Lu, G., Liu, F., Yi, Z., Jia, X., Wu, M., Shi, B., Xu, S., et al. (2020a). Viral and host factors related to the clinical outcome of COVID- 19. Nature 583, 437–440.

Zhang, X.J., Qin, J.J., Cheng, X., Shen, L., Zhao, Y.C., Yuan, Y., Lei, F., Chen, M.M., Yang, H., Bai, L., et al. (2020b). In-hospital use of statins is associated with a reduced risk of mortality among individuals with COVID-19. Cell Metab. 32, 176–187.e4.

Zhang, Y., Xiao, M., Zhang, S., Xia, P., Cao, W., Jiang, W., Chen, H., Ding, X., Zhao, H., Zhang, H., et al. (2020c). Coagulopathy and antiphospholipid anti- bodies in patients with Covid-19. N. Engl. J. Med. 382, e38.

[1] https://pypi.org/project/GEOparse/

[2] ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE157nnn/GSE157103/suppl/GSE157103%5Fgenes%2Etpm%2Etsv%2Egz

[3] https://github.com/rjoana1/SIB_omics