Filipe Santos – 55142
José Mendes - 55954

Aprendizagem Automática Avançada – Proposta de Projeto

Ensemble Model for Covid-19 Risk Prediction

Para o nosso projeto resolvemos que queríamos estudar um assunto pertinente, relacionado com acontecimentos atuais, além de incorporar a matéria lecionada na cadeira. Decidimos então que o nosso projeto deveria estar relacionado com a pandemia de COVID-19. Encontrámos um artigo no qual os autores usaram 4 técnicas, *Logical regression*, SVM, *Gradient Boosted Decison tree* e *Neural networks*, para construir e validar um modelo *ensemble* de previsão de risco de mortalidade para o COVID-19 ou MRPMC (*Mortality risk prediction model for COVID-19*) e resolvemos usá-lo como referência para o projeto. Visto que o dataset usado neste artigo não é publico, para proteger a privacidade dos doentes, tivemos de recorrer a um dataset alternativo.

Começando pelo artigo referido, o dataset usado é composto por 34 variáveis – 16 contínuas e 18 categóricas. Das variáveis categóricas fazem parte alguns dados do doente (sexo, idade) e informação sobre presença de comorbilidades (hipertensão, diabetes, doença respiratória crónica, por exemplo). As variáveis contínuas são compostas por resultados de medições e análises clínicas, entre elas temperatura, taxa respiratória e taxa de oxigenação.

Os dados dizem respeito a 1068 doentes do Hospital de Tongji, na China. Estes dados foram divididos em dataset de treino e validação na proporção 50:50. Os modelos foram construídos e testados em 2 novos datasets, de 2 hospitais diferentes. A ideia inicial seria utilizar os dados do artigo para aplicar outros tipos de modelos ou otimizar os modelos utilizados. No entanto, como referido anteriormente, estes dados não são abertos ao público. Sendo assim, o âmbito do projeto foi alterado e o plano passou a ser aplicar os modelos e técnicas referidos no artigo a um novo conjunto de dados.

Encontrámos no Kaggle um dataset com informação acerca de doentes Covid no México. Esse dataset levou-nos ao site da Direção Geral de Epidemiologia do Governo Mexicano. Este site disponibiliza diariamente dados de doentes Covid tratados em ambiente hospitalar a nível nacional, tanto em ambulatório como em internamento. Para estabelecer melhor um paralelo entre o artigo e o nosso projeto, optámos por considerar apenas dados de doentes que foram internados. Do dataset inicial com 5896765 registos sobraram 684631. O dataset original contém várias variáveis que, pensamos, não têm influência no desfecho da doença (dados relacionados com a morada do doente, por exemplo). Ficámos então com um dataset com 21 variáveis, onde o *target* é a variável binária **Morte** -1 se doente faleceu, 0 em caso contrário. As classes estão divididas na proporção 65/35, com prevalência da classe 0. As restantes são 2 variáveis contínuas e 18 categóricas, podendo ser divididas em 4 classes:

- **Demográficas:** Sexo, Idade (contínua);
- **Intervenção Médica:** Internado ou não, entubado ou não, admitido nos cuidados intensivos.
- **Covid:** número dias entre início de sintomas e internamento (contínua), teste positivo ou não, contacto com outro doente Covid, se desenvolveu pneumonia.
- **Condições pré-existentes:** diabetes, DPOC, asma, imunodeprimido, hipertensão, gravidez, doente cardiovascular, obesidade, fumador, doença renal crónica, outra doença.

Estamos, portanto, perante um dataset de composição diferente do utilizado no artigo, onde a divisão entre variáveis categóricas e contínuas era quase 50/50. As variáveis categóricas estão, de modo geral, definidas de acordo com a seguinte tabela:

CHAVE	DESCRIÇÃO
1	SIM
2	NÃO
97	NÃO SE APLICA
98	NÃO SE SABE
99	NÃO ESPECIFICADO (NULL)

Os valores **97**, **98** e **99** podem ser considerados como valores nulos e deverão ser tratados como tal. No entanto, ainda não temos uma estratégia definida para isso.

Posto isto, a nossa proposta de projeto envolve a aplicação da estratégia utilizada pelos autores do artigo a este novo conjunto de dados: criar 6 modelos *baseline* - *LogisticRegression*, SVM, *Neural Network*, *KNN*, *Random Forest* e *Gradient Boosted Decision Tree* - avaliar a sua performance medindo a AUC e comparar resultados. Os modelos que apresentem melhor performance serão selecionados e incluídos num modelo *ensemble* – *AdaBoost* e/ou *XGBoost*. Para tal iremos recorrer, em Python, às bibliotecas *pandas* e *numpy* para transformação dos dados, e às bibliotecas *sci*-kit *learn* e *xgboost* para construção dos modelos. O resultado final do projeto será um modelo que permitirá prever o risco de morte associado a um doente Covid internado, para o qual se conhecem os valores das variáveis consideradas. Este modelo teria como aplicação real, por exemplo, a triagem de doentes para os diferentes níveis de prestação de cuidados a nível hospitalar.