# Nanodegree Engenheiro de Machine Learning

## Proposta de projeto final

Rodrigo Marques Pessoa
13 de setembro de 2018

### Histórico do assunto

As técnicas avançadas de análise de dados e aprendizagem de máquina trazem um ganho exponencial: permitem olhar adiante e auxiliar a medicina preventiva. Com base em dados de pacientes, é possível usar a análise de dados de uma maneira mais inteligente para sugerir cuidados preventivos em determinadas épocas do ano, ou até mesmo promover um estilo de vida mais saudável.

O repertório de informações que os cientistas precisam podem ser encontrados em softwares e equipamentos médicos que emitem dados.

Podemos minerar os registros de dados que os hospitais emitem sobre seus pacientes por meio de um CRM. A partir das informações selecionadas, é possível criar campanhas e ações preventivas.

Com algorítimos matemáticos cada vez mais precisos e baseado nesse histórico de dados, a máquina consegue alertar sobre possíveis doenças, identificar grupos de tendências, entre outros fatores.

### Descrição do problema

A falta de leito em hospitais é um problema constante encontrado em várias cidades brasileiras. Gerenciar essas vagas é cada vez mais essencial e ter essa informação o quanto antes pode ajudar a estabelecer prioridades de atendimento.

A medida que consultas são realizadas nas diversas unidades de saúde da cidade, essas informações poderiam ser enviadas a um servidor central que mediante a estes dados pode aplicar um modelo preditivo e contabilizar em tempo real a necessidade de utilização de leitos em hospitais, ou tomar ações preventivas, por exemplo.

## Conjuntos de dados e entradas

*Para tentar prever a necessidade de internação utilizaremos os dados das unidades de saúde com os dados dos pacientes e dados dos diagnósticos realizados pelos médicos. O modelo preditivo utilizará os dados públicos da prefeitura de Curitiba que está disponível para download através do seguinte endereço eletrônico: http://www.curitiba.pr.gov.br/dadosabertos/consulta/?grupo=1.*

*Segue abaixo um dicionário de dados com as variáveis disponíveis e as que foram escolhidas para serem preditoras do modelo. As variáveis em verde foram escolhidas pois podem apresentar alguma correlação com a variável classe(em azul), como por exemplo dados socioeconômicos dos pacientes (pacientes com menor poder aquisitivo ou que não tem infraestrutura como coleta de lixo, água tratada poderia necessitar de internações constantes). Já as variáveis em vermelhos foram descartadas pois não apresentam nenhuma correlação para o estudo deste projeto:*

***Dicionário de Dados***

| Nome do Campo | Descrição |
|---|---|
| Data do Atendimento | Data de Realização do Atendimento |
| Data de Nascimento | Data de Nascimento do Paciente |
| Sexo | Sexo do Paciente |
| Código do Tipo de Unidade | Código do Tipo de Unidade de Atendimento |
| Tipo de Unidade | Tipo de Unidade de Atendimento |
| Código da Unidade | Código da Unidade de Atendimento |
| Descrição da Unidade | Descrição da Unidade de Atendimento |
| Código do Procedimento | Código do Procedimento Realizado |
| Descrição do Procedimento | Descrição do Procedimento Realizado |
| Código do CBO | Código da Ocupação do Profissional |
| Descrição do CBO | Descrição da Ocupação do Profissional |
| Código do CID | Código do Diagnóstico |
| Descrição do CID | Descrição do Diagnóstico |
| Solicitação de Exames | Indica se ocorreu solicitação de Exames |
| Qtde Prescrita Farmácia Curitibana | Qtde de medicamentos prescritos na Farmácia |
| Qtde Dispensada Farmácia Curitibana | Qtde de medicamentos dispensados na Farmácia |
| Qtde de Medicamento Não Padronizado | Qtde de Medicamento Não Padronizado |
| Encaminhamento para Atendimento Especialista | Indica se houve encaminhamento para Atendimento de Especialista |
| Á□rea de Atuação | Á□rea de Atuação |
| Desencadeou Internamento | Indica se desencadeou Internamento |
| Data do Internamento | Data do Internamento do paciente |
| Estabelecimento Solicitante | Estabelecimento que solicitou o internamento |
| Estabelecimento Destino | Estabelecimento que houve a internação |
| CID do Internamento | Código do diagnóstico do internamento |
| Tratamento no Domicílio | Tipo de Tratamento de Água no domicílio |
| Abastecimento | Tipo de Abastecimento de Água no domicílio |

| | |
|---|---|
| Energia Elétrica | Indica se há energia elétrica no domicílio |
| Tipo de Habitação | Tipo de habitação no domicílio |
| Destino Lixo | Destino do lixo no domicílio |
| Fezes/Urina | Destino das fezes/urina no domicílio |
| Cômodos | Qtde de Cômodos no domicílio |
| Em Caso de Doença | Serviços procurados em caso de doença |
| Grupo Comunitário | Grupo Comunitário em que o paciente participa |
| Meio de Comunicação | Meios de Comunicação utilizados no domicílio |
| Meio de Transporte | Meios de Transporte utilizados no domicílio |
| Município | Município do paciente |
| Bairro | Bairro do paciente |

## Descrição da solução

*A solução encontrada é informatizar as unidades de saúde com acesso a internet para que os dados das consultas sejam enviados a um servidor central que executará o modelo preditivo nos dados históricos, e em tempo real, nos dados de produção para ter uma previsão da necessidade de internação. Essa informação poderá ser exibida em um dashboard numa central de monitoramento da prefeitura que poderá tomar as ações necessárias para administrar da melhor forma possível os leitos.*

*Vamos utilizar os dados disponíveis para realizar a previsão. Estes dados não foram coletados com uma visão de utilização para predição e, portanto, podem não ser os mais indicados, mas como nosso intuito e apenas educacional isso não será um fator crítico. No entanto em um ambiente real poderíamos realizar um trabalho junto a prefeitura para obtermos outros dados que não estão disponíveis atualmente, como por exemplo, os dados dos resultados de exames a fim de melhorarmos a precisão das previsões.*

## Modelo de referência (benchmark)

*Acredito que os dados atualmente são agrupados manualmente e enviados ao departamento de saúde somente após alguns dias ou talvez meses e a partir disso é que o gestor vai ter uma visão histórica da quantidade de internações foram necessárias e tomar alguma ação para tentar melhorar a disponibilidade dos leitos.*

*Nossa proposta é justamente que ele tenha essa necessidade baseada em dados históricos que alimentarão um modelo preditivo e assim gerar em tempo real a necessidade de leitos e poder tomar as ações no exato momento da necessidade, e não posteriormente como deve ocorrer hoje.*

*Como não temos um modelo real utilizarei o modelo DummyClassifier do sklearn como modelo de referência para verificarmos e compararmos se o modelo escolhido realmente apresenta uma performance melhor que uma metologia aleatória.*

*DummyClassifier é um classificador que faz previsões usando regras simples. Esse classificador é útil como uma linha de base simples para comparar com outros classificadores (reais). Não deve ser usado para problemas reais conforme indicado na documentação do método: http://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html.*

## Métricas de avaliação

*Espera-se que o modelo seja capaz de predizer com um índice alto de precisão as consultas que estão sendo realizada nas diversas unidades de saúde da cidade se necessitam ou não de internação.*

*Utilizaremos o indicador F-score com beta de 2 pois precisamos de alto recall, ou seja, realmente identificar as consultas que precisam de internação, pois se identificarmos com baixo recall vamos precisar de mais leitos do que estamos realmente prevendo. Assim, uma precisão baixa não é uma situação preocupante pois estamos prevendo leitos a mais do que realmente vamos precisar.*

## Design do projeto

*Como o os dados apresentam as informações dos últimos 3 meses de atendimento das unidades de saúde, contendo quase 1.000.000 de registros utilizaremos uma amostragem para realizar o modelo preditivo.*

*Trabalharei com uma amostra de 664 registros, que foi calculado levando em consideração 5% de erro amostral e 99% de nível de confiança.*

*Como os dados não foram coletados com o intuito de utilização para o modelo de predição existe uma grande quantidade de dados ausentes e um trabalho de limpeza e padronização será necessário*

*Outra questão é fazer o balanceamento das classes já que a proporção de necessidade de internações comparada as da não necessidade são muito desproporcionais.*

*Baseado nos tipos de dados que temos atualmente realizaremos vários testes para escolha do modelo preditivo com o melhor índice de recall, que será indicado pela métrica definida no tópico anterior.*

*Por fim, vamos tentar melhorar a performance testando alguns hiperparametros com o método gridsearch();*