Predição do volume de atendimentos de saúde na cidade de Curitiba utilizando dados abertos

Mayara Regina Lorenzi[1], Cristiano da Cunha Ribas[1], Luiz Gomes Jr.[1]

[1]Deparamento Acadêmico de Informática – Universidade Tecnológica Federal do Paraná (UTFPR) – Curitiba, PR - Brasil

may.lorenzi@gmail.com, cristianorbs@gmail.com,
gomesjr@dainf.ct.utfpr.edu.br

Abstract. *This paper explores the prediction of total demand of nursing service on public hospitals and walkin clinics of the city of Curitiba. The analysis is based on data mining techniques applied to open data provided by the municipality and weather data by Weather Underground. We present here an exploratory analysis and the implementation of a preliminary model of linear regression to estimate the variation of the demand for healthcare service related to respiratory ailments.*

Resumo. *Este trabalho explora a predição de volume de atendimentos de doenças respiratórias na rede pública da cidade de Curitiba. A análise é baseada em métodos de mineração de dados aplicados em dados abertos de atendimento fornecidos pelo município e dados climáticos fornecidos pelo portal Weather Underground. Apresentamos aqui uma análise exploratória dos dados e a implementação de um modelo preliminar de regressão linear para estimativa de variação no volume de atendimentos referentes a doenças respiratórias.*

1. Introdução

As variações metereológicas e climáticas podem impactar diretamente na saúde da população. Para as doenças do trato respiratório, especula-se que a condição climática do local pode interferir no volume de pessoas afetadas. Em períodos de chuva ou com mudanças bruscas de temperatura, por exemplo, pode ocorrer maior propensão a doenças virais, como a gripe [Viveiros, 2014].

Além disso, a gestão pública de saúde não possui insumos necessários para o planejamento a médio e longo prazo de escala de medicos e enfermeiros, compra de itens medicamentosos e de uso clínico.

Isso se deve ao fato de que o volume de pacientes que necessitam de consultas e de internamentos depende de diversos fatores, como período do ano, condições meteorológicas, qualidade de vida da população, entre outros.

Nos anos 2000, foi lançado o conceito de Cidades Inteligentes e está em crescente debate politico e acadêmico. O tema é abordado para se referir a cidades que fazem uso da tecnologia para aprimorar o processo de planejamento, a fim de melhorar a sustentabilidade do local e encontrar soluções para a sociedade e para o Estado.

Nesse contexto, a Tecnologia da Informação, através de aprendizado de máquina e análise exploratória de dados, visa maximizar a obtenção de informações ocultas em um grande volume de dados, descobrir variáveis importantes nas tendências e detector comportamentos anômalos.

Esse projeto tem como objetivo criar, a partir de dados reais fornecidos pelos órgãos responsáveis, artefatos computacionais para apoio à previsão de volume de pacientes com sintomas de doenças respiratórias que necessitarão de atendimento médico na cidade de Curitiba, a fim de resolver parte do problema de escala de médicos e enfermeiros e a compra de medicamentos.

Foram utilizados dados relacionados aos atendimentos médicos e de enfermagem disponibilizados pela Secretaria de Saúde da Prefeitura de Curitiba. Aplicando técnicas de análise exploratória e algoritmos de aprendizado de máquina nos dados fornecidos, juntamente com dados de informações climáticas, pôde-se obter variáveis importantes e identificar tendências de comportamento.

2. Trabalhos Correlatos

Nos anos 2000, um estudo realizado pelo Departamento de Saúde da Prefeitura de Curitiba mostrou que as doenças do aparelho respiratório constituíram o motivo mais frequente de consulta (19,6%) e quarta causa de morte em todas as faixas etárias da população do município. Na faixa etária pediátrica a proporção se torna mais extrema, representando 50% das consultas ambulatoriais e aproximadamente 25% dos internamentos em menores de 14 anos [Prefeitura de Curitiba, 2010].

Pesquisas realizadas na Filadélfia [Hollemand, 1996] estudaram as associações entre variáveis climáticas e volume de consultas agendadas para o período em clínicas de emergência e pronto atendimento. A análise levou em consideração, além dos fatores mencionados, a estação do ano, o dia da semana e as vésperas de feriado.

Foi identificado um alto volume de pacientes durante os meses de inverno, com exceção de dezembro. Esse fato pode ser devido à estação ser facilitadora para aumentar as doenças pulmonares, dada a exposição ao frio. E, se tratando de agendamento de consultas, o número pode ser reduzido em dezembro devido aos feriados e férias.

Um estudo realizado pela Universidade Federal de Santa Maria [Gonçalves, 2010] analisou a variação da morbidade de doenças respiratórias em função da variação da temperatura entre os meses de abril e maio em São Paulo. A pesquisa demostrou que pode haver relação entre a temperatura mínima e doenças respiratórias na população infantil. Há um pico de morbidade por doenças respiratórias das vias superiores no mês de maio, devido ao problema de termo-regulação em indivíduos adaptados ao clima mais ameno. Esta tendência de aumento da diferença pode aumentar a ida aos hospitais no mês de maio, gerando impacto em hospitais e em políticas públicas.

3. Descrição dos dados utilizados

A. Dados de atendimentos hospitalares

Os dados de atendimento hospitalar foram disponibilizados pela Prefeitura de Curitiba através da plataforma de dados abertos da Universidade Federal do Paraná[1].

A base de dados é oriunda do sistema "E-saúde", que mantém as informações de atendimentos de enfermagem e médicos prestados pela Secretaria Municipal de Saúde de Curitiba em sua rede de atenção, que é composta por Unidades Básicas de Saúde, Unidades de Pronto Atendimento e Centros de Especialidades Médicas e Odontológicas.

A base inclui diversas informações sobre os atendimentos e pacientes. As utilizadas para esse projeto são:

• Código do CID-10, que é o código do diagnóstico do paciente. A CID (Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde) é publicada pela Organização Mundial de Saúde (OMS) e fornece códigos relativos à classificação de doenças utilizados globalmente para estatística de morbilidade e mortalidade.

• Data de nascimento e sexo do paciente, utilizados nesse projeto para aprofundamento dos resultados.

• Desencadeou Internamento, que indica se o paciente foi encaminhado para internamento após a consulta.

A área de interesse desse projeto se refere às doenças catalogadas no CID-10 como Doenças do Aparelho Respiratório (J00-J99), que podem ser desdobradas de acordo com os grupos de doenças destinado para cada código do CID-10 abordado.

O total de atendimentos de doenças do aparelho respiratório representaram, no período de junho a agosto de 2016, o principal motivo de consultas médicas (17,9%) e o segundo cenário em número de internamentos (15,6%), para todas as faixas etárias. Para a faixa etária pediátrica (até 14 anos), o número é ainda maior, sendo 37,6% no número de consultas e 46,9% em internamentos.

B. Dados climáticos

Os dados metereológicos utilizados foram disponibilizados pela The Weather Company, através da Weather Underground, que fornece uma API gratuita para desenvolvedores.

O ponto de referência para esse estudo é o bairro Centro da cidade de Curitiba. As variáveis climáticas utilizadas nessa pesquisa são a temperatura média, umidade relativa do ar média do dia e a amplitude térmica, as demais nove características climáticas, como velocidade do vento, foram desprezadas.

C. Remoção de anomalias

Esse passo consistiu na eliminação de 31 colunas, que representam variáveis que não foram utilizadas nesse estudo, como localização da unidade de atendimento e dados pessoais dos pacientes.

Após a eliminação das colunas desnecessárias, foram excluídos os registros duplicados e com preenchimento incompleto – sem data de atendimento e sem código CID-10. Além disso, foram calculadas as idades dos pacientes, gerando uma nova coluna à tabela, para que a análise fosse aprofundada nesse nível.

Os arquivos de entrada dessa etapa são divididos por período trimestral e totalizam 2,81GB de dados, com 7.757.527 registros. A manipulação desses dados, que durou aproximadamente 3,5 minutos, resultou em 301.915 registros, o que representa 3,9% da amostra inicial. Dessa forma, a execução das próximas etapas do processo de análise exploratória foi otimizada.

D. Agregação dos dados

Os dados metereológicos foram agrupados aos registros de atendimento, utilizando-se a data de atendimento como parâmetro de junção e de agregação.

Além das variáveis mencionadas acima, também foram calculadas as médias de temperatura, umidade e amplitude térmica dos últimos sete dias e armazenadas em um novo campo. Com essas variáveis, pretendeu-se afirmar a hipótese de que os atendimentos não variam conforme as características do dia em questão ou do dia anterior, e sim, com a influência das características dos últimos dias.

Após a agregação dos valores, os dados foram separados pelos códigos CID-10 que se iniciam com J, o que indica que o paciente foi diagnosticado com uma doença respiratória. Obteve-se, assim, duas bases de dados para fins comparativos: total de atendimentos e somente atendimentos de doenças respiratórias.

E. Normalização dos dados

O processo utilizado foi o de normalização por desvio padrão, que constitui-se em calcular a média do volume de atendimento de cada dia da semana, subtrair do total de cada dia e dividir o resultado dessa operação pelo desvio, como mostra a equação (1).

$$x' = \frac{x - x_m}{\sigma} \quad (1)$$

Onde, x é o volume de atendimento, x_m é a média de atendimento do dia da semana em questão, σ é o desvio padrão e x' é o valor resultante da normalização, que será considerado para aquele dia. Dessa forma, ao subtrair a média e dividir pelo desvio padrão, manteve-se apenas a variação de cada dia.

Com a normalização, pretendeu-se eliminar a possibilidade dos resultados serem influenciados pelos dias da semana e, sendo assim, a escala foi alterada.

4. Metodologia

A. Pré-processamento de dados

O pré-processamento de dados consistiu na execução de três etapas:

1) Remoção de anomalias

Foi observado um fenômeno em que nas segundas e terças-feiras o volume de atendimentos aumentava significativamente. Após análise, identificou-se que grande parte desses atendimentos eram de consultas que não possuíam CID-10, ou seja, eram de atendimentos a pessoas que não estavam doentes.

Além disso, foram removidos dados duplicados, visto que muitas vezes eram cadastrados dados sobrepostos, duplicando ou até triplicando o mesmo registro.

Também foram removidos registros com formato de data inválido. Esse efeito pôde ser visto pelo fato de os dados serem cadastrados manualmente e, não necessariamente, seguindo um padrão.

2) Integração dos dados

Foram agrupados os dados climáticos aos dados de atendimento a partir da coluna de data, ou seja, foi feita a junção dos registros climáticos e de atendimento quando possuíam a mesma data de ocorrência.

3) Transformação de dados

Foi feita a alteração em alguns formatos de datas, pois nem todos seguiam o mesmo padrão. Além disso foram adicionadas algumas novas colunas calculadas, como idade, temperatura média, umidade média e amplitude térmica média da última semana.

B. Regressão Linear

A partir das características selecionadas (temperatura média, umidade média e amplitude térmica média da última semana), decidiu-se criar um modelo de regressão linear com o objetivo de tentar estimar o volume de consultas de um ou mais dias posteriores baseado nas variáveis climáticas dos dias anteriores.

Primeiramente foi feita a divisão de 80% dos dados para treinamento e 20% para a validação. Esses dados foram selecionados de maneira aleatória.

O modelo foi ajustado a partir dos dados de treinamento e validado utilizando os dados de teste.

Por fim foi calculado o erro quadrado mínimo dos valores previstos, bem como o coeficiente de determinação e o valor p de cada característica para avaliar quão bem o modelo escolhido pode ou não prever o total de atendimentos futuro.

5. Resultados

Para a criação do modelo foi, primeiramente, calculada a similaridade do cosseno centralizado, ou coeficiente de correlação de Pearson, entre as características e o alvo (total de atendimentos normalizado). Tais coeficientes são descritos na tabela 1 e são utilizados para dizer se há uma possibilidade de existir uma relação linear ou não entre duas variáveis.

TABELA 1. COEFICIENTES DE CORRELAÇÃO DE PEARSON ENTRE AS CARACTERÍSTICAS E O ALVO

	Total Normalizado
Temperatura mínima do dia	-0,176277
Média de temperatura dos últimos 7 dias	-0,289704
Média de umidade dos últimos 7 dias	-0,058195
Média da amplitude térmica dos últimos 7 dias	0,038121

Os coeficientes negativos indicam que a as variáveis tendem a descrescer. Nesse caso seria a indicação, por exemplo, de que quando a temperatura mínima diminui, o total normalizado aumenta.

Segundo [Evans, 1996], a correlação do valor mais significativo encontrado (-0,289704) é considerada fraca, mas dada a complexidade do problema proposto, as correlações foram consideradas razoáveis para continuar as análises.

Para uma análise visual, foi traçado o gráfico da figura 1, representando os valores total de atendimentos comparados com a temperatura média dos últimos 7 dias referente ao dia em questão, que foi a característica o coeficiente de correlação mais representativo. Além disso foi traçada a reta que melhor se ajustou nos pontos representados no gráfico.

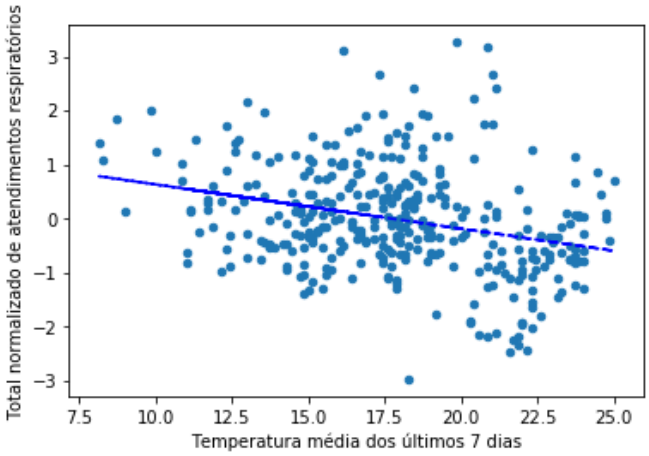

Figura 1. Temperatura média da última semana pelo total normalizado de atendimentos de doenças respiratórias

Pode-se notar que a reta representada na figura 1 tende a decrescer. Apesar de ter uma inclinação leve, pode ser um indicativo de o total de atendimentos aumentar à medida que a temperatura média da semana anterior diminui.

Com a separação dos dados de treinamento e de teste, foram utilizados os 80% referentes aos dados de treinamento para criar o modelo utilizando o método dos mínimos quadrados ordinários.

Os valores-p, ou probabilidades de significância, de cada variável são apresentados na tabela 2, bem como os coeficientes e o erro padrão de cada característica no modelo.

TABELA 2: VALORES-P, COEFICIENTES E ERROS PADRÃO DAS CARACTERÍSTICA DO MODELO

	Valor-p	Coeficiente	Erro padrão
Temperatura mínima do dia	0,149	0,0667	0,046
Média de temperatura dos últimos 7 dias	0,0388	-0,114	0,054
Média de umidade dos últimos 7 dias	0,1309	0,0543	0,036
Média da amplitude térmica dos últimos 7 dias	0,1076	0,1877	0,115

Considerando um nível de significância de 5%, somente seria rejeitada a hipótese nula utilizando a característica de média de temperatura da última semana. Além disso, o coeficiente dessa característica é de -0,114, o coeficiente com maior significância se for

considerado o erro padrão, visto que a característica de média da amplitude térmica da última semana é mais alta, porém o valor-p acima de 5% não rejeitaria a hipótese nula e o erro padrão seria de 11,5%.

O coeficiente de determinação, também chamado de R^2, encontrado foi de 0,104. Isso significa que 10,4% da variável dependente (total normalizado de atendimentos) consegue ser explicado pelo modelo criado a partir das outras características.

Então, foi utilizado o modelo para tentar prever os dados de teste e validar o modelo. O gráfico da imagem 2 mostra a comparação entre os valores reais e os previstos.

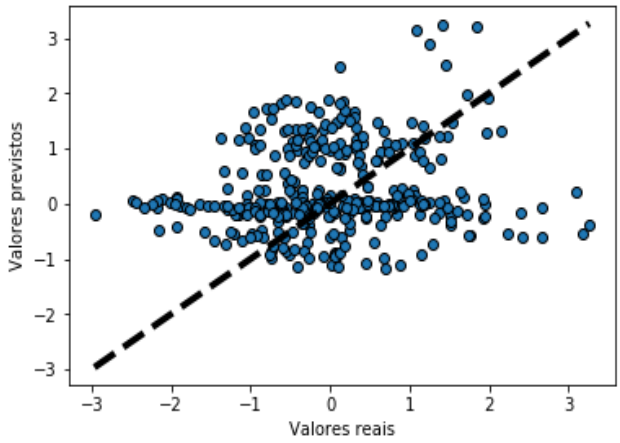

Figura 2: Comparação entre valores previstos e valores reais do total de atendimentos normalizado

A linha tracejada representa a posição ideal dos pontos caso a predição fosse 100% assertiva e foi inserida para melhorar a visualização da distribuição dos pontos.

6. Conclusão

O modelo desenvolvido foi capaz de capturar cerca de 10% da viariação de atendimentos por doenças respiratórias. Considerando que um dia típico tem média de 350 atendimentos em Curitiba, o modelo é capaz de prever cerca de 10 casos para mais ou para menos, Consideramos o resultado relevante para este modelo inicial desenvolvido, sobretudo considerando-se a complexidade do problema.

Diversas variáveis não foram consideradas neste modelo inicial e pretendemos empregá-las em futuros refinamentos do modelo. Por exemplo, seria importante tratar a agregação da temperatura em janelas diferentes (usar 7 dias foi uma decisão arbitrária e de relevância não avaliada). Também pretendemos agregar informações sobre poluição do ar e tratar com maior granularidade a idade e diagnóstico do paciente. Também pretendemos estender o perído de dados coletados.

Infelizmente, diversas variáveis importantes associadas ao problema são difíceis de se caracterizar, como por exemplo as variantes dos vírus presentes em um determinado período. De acordo com mutações genéticas imprevisíveis, varia-se de ano a ano a taxa de transmissibilidade, gravidade de sintomas e época de disseminação dos vírus.

A despeito da complexidade do problema, acreditamos que os resultados obtidos são promissores e já poderiam ser considerados para o plenejamento de atendimentos. Esperamos também aumentar significativamente a precisão do modelo a partir dos refinamentos previstos.

7. Referências

B VIVEIROS, JOSÉ. Universidade de Coimbra: A Influência das Alterações Climáticas nas Patologias Respiratórias. Disponível em <http://estudogeral.sib.uc.pt/bitstream/10316/29245/1/A%20Influ%C3%AAncia%20das%20Altera%C3%A7%C3%B5es%20Clim%C3%A1ticas%20nas%2o fazendo0Patologias%20Respirat%C3%B3rias.pdf>

CONNECTED SMART CITIES. O que é uma cidade inteligente? Disponível em <http://www.connectedsmartcities.com.br/index.php/afinal-o-que-e-uma-cidade-inteligente>

SILVA, BRIGIANE. VANDERLINE, MARCOS. Inteligência Artificial, Aprendizado de Máquina. Disponível em <http://www.ceavi.udesc.br/arquivos/id_submenu/387/brigiane_machado_da_silva__marcos_vanderlinde.pdf>

PREFEITURA DE CURITIBA. Infecções e Doenças Respiratórias. Disponível em <http://www.saude.curitiba.pr.gov.br/index.php/programas/saude-da-crianca/infeccoes-e-alergias-respiratorias>

HOLLEMAN, DONALD. BOWLING, RENEE. GATHY, CHARLANE. Predicting Daily Visits to a Walk-in Clinic and Emergency Department Using Calendar and Weather Data. Disponível em <https://www.researchgate.net/publication/14457443_Predicting_daily_visits_to_a_walk-in_clinic_and_emergency_department_using_calendar_and_weather_data>

GONÇALVES, FÁBIO. COELHO, MICHELINE. Universidade Federal de Santa Maria: Variação da morbidade de doenças respiratórias em função da variação da temperatura entre os meses de abril e maio em São Paulo. Disponível em <https://periodicos.ufsm.br/cienciaenatura/article/viewFile/9500/5649>

Evans, J. D. Straightforward statistics for the behavioral sciences. Pacific Grove, CA: Brooks/Cole Publishing, 1996.