Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Arcson de Melo Assunção

kNN para seleção de Operadora Móvel Celular

Curitiba
2020

Arcson de Melo Assunção

kNN para seleção de Operadora Móvel Celular

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Luiz Eduardo Soares de Oliveira

Curitiba
2020

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

kNN para seleção de Operadora Móvel Celular

Arcson de Melo Assunção[1]
Luiz Eduardo Soares de Oliveira, Ph.D[2]

Resumo

Atualmente existem bilhões de dispositivos que para funcionarem dependem das redes das operadoras de telefonia móvel, tais como telefones celulares, rastreadores, sistemas de monitoramento eletrônico, redes elétricas inteligentes, fazendas inteligentes, meios de pagamento e muitos outros. Esse artigo tem o objetivo de apresentar uma nova metodologia para auxiliar na escolha da melhor operadora, afim de que os sistemas que dependem desse tipo de comunicação, mitiguem problemas, e se mantenham o máximo de tempo possível conectados e em funcionamento. Foram propostos dois modelos e realizada uma comparação entre eles. Ambos utilizam um classificador kNN, um considerando somente as medidas georreferenciadas e a identificação da operadora e outro acrescentando um fator de penalidade considerando a intensidade do sinal. Após a realização de 50 mil experimentos, 25 mil com cada modelo, obtivemos o resultado de 89,0% para o modelo penalizado, e 89,4% para o modelo sem a penalização. O que nos leva a conclusão de que o nível de sinal não é um fator preponderante na escolha da melhor operadora. **Palavras-chave**: knn, nearest neighbor, classificação, supervisionada, iot, m2m, sinal celular, melhor operadora, qualidade de conexão, escolha de operadora, python, webscrap, ETL.

Abstract

Today there are billions of devices that work depending on the networks of mobile operators, such as cell phones, trackers, electronic monitoring systems, smart grids, smart farms, payment systems and many others. This article aims to present a new methodology to assist in choosing the best operator, in order that systems that depend on this type of communication, mitigate problems, and stay connected as long as possible and in operation. Two models were proposed and a comparison was made between them. Both use a kNN classifier, one considering only the georeferenced measures and operator identification and the other adding a penalty factor considering the signal strength. After performing 50 thousand experiments, 25 thousand with each model, we obtained the result of 89.0% for the penalized model, and 89.4% for the model without the penalty. Which leads us to the conclusion that the signal level is not a major factor in choosing the best operator. Versão em inglês do resumo.

Keywords: *knn, nearest neighbor, classification, supervioned, iot, m2m, signal cellular, best carrier, connection quality, choose carrier, python, webscrap, ETL*

1 Introdução

A tecnologia tem crescido de forma exponencial, causando uma revolução em todas as áreas da vida humana. Dentre muitas forças motrizes que impulsionam esse crescimento, três em especial se destacam de forma bem significativa: a miniaturização de dispositivos, o barateamento de componentes eletrônicos e o aumento da conectividade. Esses três vetores principais permitiram a criação de uma infinidade de soluções antes consideradas como tecnologicamente ou economicamente inviáveis.

No aspecto da miniaturização e do aumento de capacidade, podemos observar a evolução dos processadores, dos dispositivos eletrônicos e computadores, que tiveram seus tamanhos reduzidos expressivamente ao longo do tempo. Segundo Gordon Moore [1] fundador da Intel, de 1965 à 1975, a quantidade de transistores em um processador dobraria a cada 18 meses conforme a Figura 1. E isso de fato ocorreu, devido a evolução tecnológica e produção em larga escala, esse artigo ficou mundialmente conhecido como a "Lei de Moore". Essa previsão se mantém válida até os dias atuais se considerarmos o intervalo de aproximadamente dois anos. Diversos estudos estimam que ela continuará válida até 2021/2022, por conta de limitações físicas, pois recentemente os transistores atingiram a escala de 7 nanômetros, porém até essa barreira pode ser superada segundo Chen no artigo "Atomic level deposition to extend. Moore's law and beyond"[2].

Um outro ponto fundamental é o crescimento exponencial dos meios de comunicação. Podemos evidenciar esse processo através do crescimento da quantidade de aparelhos celulares. O primeiro aparelho comercial foi lançado em 1983 [3], e menos de quatro décadas depois existem cerca de 14 bilhões de dispositivos celulares no mundo conforme a projeção exibida na Figura 2.

A comunicação celular móvel em suas primeiras versões permitia somente tráfego de voz. Atualmente exis-

[1]Aluno do programa de Especialização em Data Science & Big Data, arcson@gmail.com.
[2]Professor do Departamento de Informática - DINF/UFPR.

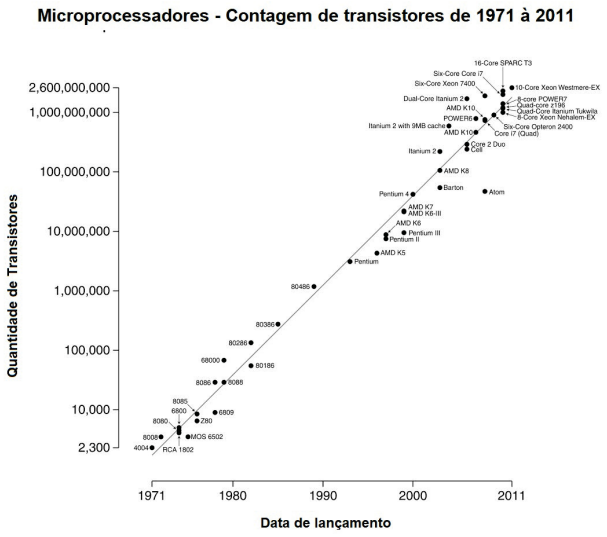


Figura 1: Lei de Moore 1965 até 2011

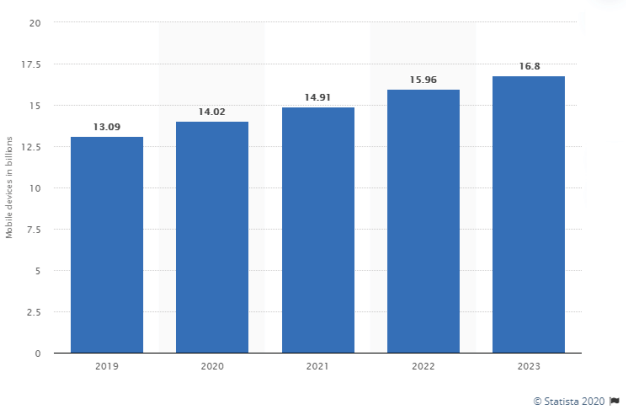


Figura 2: Projeção do número de dispositivos móveis no mundo de 2019 a 2023 (em bilhões)
[4]

tem velocidades de comunicação que chegam na casa dos Gigabits conforme a tabela 1. Além da velocidade, a abrangência de cobertura da conectividade se expandiu consideravelmente. Hoje já contamos com mais de 750 operadoras distintas[5], chegando a todos os países, segundo a GSMA [5].

Tabela 1: Velocidades XG [6]

Tecnologia	Ano de Lançamento	Velocidade
1G	1981	N/A
2G	1991	0,4 Mbps
3G	1998	21,6 Mbps
4G	2008	1.000,0 Mbps
5G	2018	20.000,0 Mbps

O custo dos componentes eletrônicos é um fator determinante na viabilização de novas soluções tecnológicas. Ano após ano temos observado um aumento na produção de componentes eletrônicos. Além do surgimento de novos materiais, componentes e tecnologias. A junção desses fatores trouxeram uma redução de custos expressiva, o que consequentemente abriu portas para o desenvolvimento de inúmeras soluções tecnológicas. Como exemplo, hoje temos mais dispositivos de comunicação móvel [7] segundo o Teleco do que pessoas no Brasil [8] segundo o IBGE. Há 30 anos atrás, era algo inconcebível, pois cada dispositivo custava dezenas de milhares de reais.

A principal rede para conexão mundial é a internet, que nos últimos anos tem sido utilizada para conexão de computadores e celulares, mas a quantidade de outros tipos de dispositivos conectados a essa rede já é tão grande, que já estamos vivendo numa era denominada de "Internet of Things"(IOT) ou "Internet das Coisas". Olhando ao nosso redor percebemos diversos tipos desses dispositivos e tecnologias, como por exemplo, a Smart TV, Smart Watch, assistentes pessoais eletrônicos (como Amazon Alexa™, e Google Home™), lâmpadas com internet sem fio, portas com trancas eletrônicas, dentre muitos outros. Esses dispositivos conectados a rede, permitem a interatividade e automação, como a ligação de uma lâmpada por comando de voz, ligar uma televisão por um aparelho celular, e até mesmo, acessar uma câmera de uma geladeira quando se faz uma compra no mercado, para saber o que está faltando.

Saindo do âmbito residencial, temos outas possibilidades de automação e otimização ainda maiores. Para o setor logístico, podemos escolher em tempo real rotas mais inteligentes e econômicas. Nas fábricas, podemos automatizar todo o processo produtivo deixando-o mais eficiente e veloz, gerando economia e aumento de produtividade. Atualmente na indústria de meios de pagamento a maior parte dos processamentos desses pagamentos é realizada através de dispositivos eletrônicos. Temos uma infinidade de outros setores como redes elétricas inteligentes, gestão de mobilidade urbana, e muitos outros. Na Figura 3 temos as principais áreas apontadas por um estudo do IoT Analytics [9].

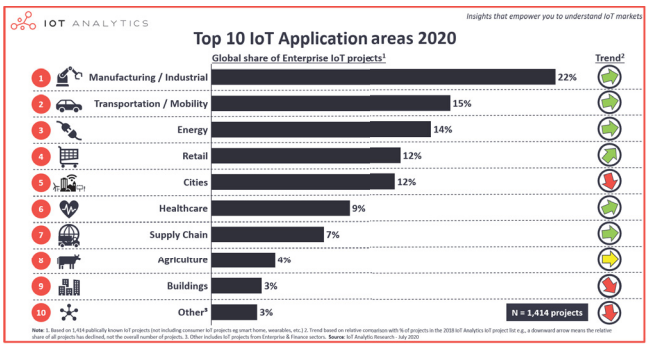


Figura 3: Areas IoT
[9]

- Essa é uma lista de tópicos apenas para ilustrar como construir.
- Coloquei apenas do itens que é o suficiente para você entender.

– Mas se a lista for hierárquica, então é só repetir o ambiente.
– Bem fácil.

2 Descrição do problema

A criação de diversos sistemas de internet das coisas (IOT) traz uma necessidade intrínseca no aspecto da conectividade, pois a maior parte desses dispositivos estão distribuídos em diversas localidades, e muitos necessitam de mobilidade. Quando estamos em casa, na universidade ou no trabalho, a conectividade em geral é provida por um link de internet de uma operadora e uma rede sem fio (Wi-Fi) de curto alcance, o que pode ser resolvido com relativa facilidade. Mas quando pensamos em bilhões de dispositivos espalhados por todo o globo, é um grande desafio fazer com que possam se conectar a uma rede.

Dentre todas as opções existentes, a categoria mais utilizada para esse tipo de conectividade são as redes sem fio distribuída e de longo alcance, como ilustrado na Figura 4. Dentro dessa categoria temos algumas opções interessantes. As redes via satélite tem cobertura em praticamente qualquer localidade, mas em geral são muito custosas, o que a inviabiliza para a maior parte dos dispositivos. As redes do tipo SigFox são uma opção economicamente interessante, mas possuem baixa cobertura e uma capacidade de tráfego muito pequena (dois bytes por pacote, e até 144 pacotes por dia). A rede LoRa possui um alcance interessante e um bom custo, mas em geral são redes privadas, em que a própria empresa precisa montar a infraestrutura necessária. E por último temos as redes celulares móveis tradicionais, derivadas da tecnologia GSM (2G, 3G, 4G e 5G), que se destacaram e se consolidaram como a opção mais utilizada mundialmente, tanto por pessoas, quanto por objetos para se conectar a rede. Isso principalmente pela sua relação de custo benefício, tendo um baixo custo de utilização, uma boa cobertura e boa capacidade de tráfego.

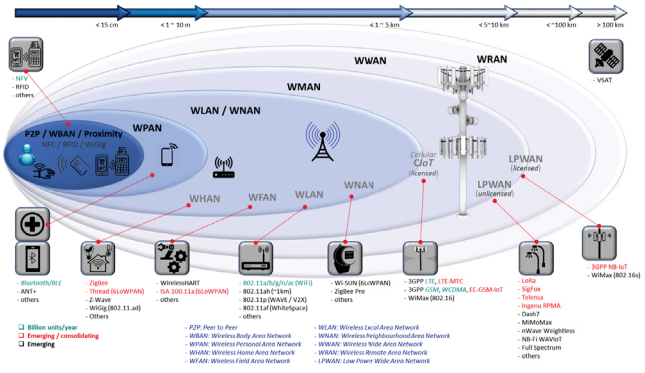


Figura 4: Tecnolgias de rede sem fio [10]

Uma vez que a maior parte dos dispositivos utilizam tecnologias celulares móveis (2G a 5G) para esse tipo de comunicação, é relativamente fácil para as prestadoras de serviço alcançar a cobertura nacional em países que possuem uma pequena extensão territorial. Porém, em países de extensões continentais, como no caso do Brasil, é um grande desafio, que se torna ainda maior se considerarmos que cada operadora tem sua infraestrutura própria, que na maioria das vezes nos leva a optar por uma operadora que funcione bem na localidade na qual necessitamos o serviço.

Alguns desses sistemas de internet das coisas se tornaram essenciais em diversos setores. Como exemplo, no setor de rastreamento que monitora o deslocamento de carga, permitindo além da gestão em tempo real, a recuperação do ativo no caso de roubo ou furto. Em redes elétricas inteligentes, além de permitir a medição remota do consumo de energia, em caso de acidentes, como o rompimento de uma rede de alta tensão, é possível realizar remotamente a interrupção da energia, preservando qualquer ameaça a vida humana. Esses exemplos evidenciam a importância da comunicação nos dispositivos móveis.

Atualmente a escolha da operadora para prestação do serviço de conectividade ocorre em sua maioria das vezes de forma empírica, por impressões passadas ou por indicações de terceiros. Em todos os casos citados não são baseados em dados concretos e massivos, e por muitas das vezes, levam a decisões equivocadas. Um recurso possível é realizar uma medição da qualidade do sinal no local onde se deseja instalar o equipamento, o que na maior parte das vezes é inviável pela relação de custo.

O processo de conectividade das redes celulares é feito através de equipamentos de radiofrequência, instalados normalmente em torres ou prédios. Esses equipamentos emitem ondas que se propagam pelo ar, essas ondas na maior parte dos casos são perturbadas pelo meio, através da colisão com objetos, condições climáticas, interferências e entre outras circunstâncias. A construção de um prédio por exemplo, poderia mudar a dispersão dessas ondas de forma significativa, fazendo com que um determinado local que possuía boa condição de conectividade passe a apresentar problemas de conexão. A conectividade sem fio é uma ciência complexa de ser aferida com precisão através de simulações ou cálculos matemáticos.

3 Proposta de Solução

O objetivo desse artigo é apresentar uma metodologia para realizar a escolha de uma operadora de telefonia móvel de forma mais assertiva e baseada em dados. Além disso, busca realizar a comparação entre duas metodologias, e identificar se a intensidade do sinal de conexão é um dado relevante no processo de predição.

O primeiro passo para aplicação da técnica é a obtenção dos dados. Preferencialmente a informação para esse tipo de análise deve ser gerada por alguns milhares de dispositivos móveis, que se utilizem da conectividade de diversas operadoras. Quanto maior a área geográfica coberta e maior a densidade de informações, melhor será a predição do modelo. Esse tipo de informação normal-

mente é conhecida como "Crowd Source Data", podendo ser gerada a partir de fontes distintas, como rastreadores veiculares, aplicativos de celulares móveis, máquinas de cartões de crédito, dentre outras.

Um ponto fundamental necessário nos dados é possuir a identificação da geolocalização no momento em que a informação foi coletada. Para a realização desse artigo foi utilizado o modelo de georreferenciamento baseado em latitude e longitude.

Os dados necessários para aplicação da metodologia devem ter a seguinte estrutura:

- Latitude - em graus
- Longitude - em graus
- Operadora - Operadora utilizada na conexão
- Nível do sinal Conexão - Nível de sinal no momento da coleta
- Nível de precisão GPS footnote - Nível de precisão do Global Position System (GPS)

A metodologia se baseia na aplicação de uma técnica de machine learning chamada Nearest Neighbor (kNN) [11]. Ele busca classificar uma nova amostra com base em sua similaridade em relação as demais existentes na população já conhecida. Para fazer isso, o algoritmo calcula a distância Euclidiana entre a amostra desejada e cada uma das amostras, para uma ou mais características, e encontra os elementos que possuem a menor distância em relação ao elemento que se deseja classificar. Uma premissa é que os elementos de sua população já conhecida estejam previamente identificados. No contexto desse artigo, significa dizer que a base de dados deve possuir as informações da geolocalização e da operadora.

4 Preparação do ambiente

4.1 Coleta de dados

Os dados utilizados para realização do estudo são oriundos de cerca de dez mil rastreadores veiculares, que geraram aproximadamente setenta milhões de registros únicos georreferenciados no período analisado.

Esses dados foram disponibilizados através de arquivos em uma página web. Para realizar a coleta das informações foi necessário criar uma aplicação com um framework da linguagem Python chamado BeatifulSoup.

Os arquivos baixados estavam em formato de texto e possuíam todas as informações de comunicação interna do veículo. Esse canal de comunicação é conhecido como barramento CAM, ele tráfega informações sobre a parte elétrica, nível de combustível, e dentre todas essas informações, os dados do rastreamento veicular.

Informações não mandatórias para o funcionamento do modelo, mas podem ser utilizadas para geração de modelos alternativos de maior acurácia.

4.2 Processamento de dados

Para extração dos dados necessários para a realização do presente estudo, foi necessário entender o funcionamento do protocolo e implementar um mecanismo de processamento desses dados. A solução escolhida para essa parte do processamento é chamada "Spoon"(anteriormente Kettle). Atualmente a ferramenta faz parte da suíte de aplicações do Pentaho, uma das suítes opensource mais utilizadas, na qual recebeu o nome de Pentaho Data Integration (PDI). Essa aplicação pertence a uma categoria de softwares chamados ETL (Extraction, Tranformation and Load ou em português Extração, Transformação e Carga). Esse tipo de aplicação permite a criação de diagramas de processamento de dados, na qual é possível de forma simples tratar de grandes transformações de dados, através de um fluxo de informações bem intuitivo, realizando pequenas transformações em cada estágio por onde a informação passa.

Quando se tem um volume expressivo de dados, nem sempre é possível processar e armazenar todo o conteúdo num único ciclo. Para isso, é necessário fazer um processamento dos dados em pequenas porções, comumente chamadas de "batchs". Nos dados utilizados nesse artigo existiam cerca de 300 mil arquivos, totalizando aproximadamente 400 milhões de linhas de dados.

Normalmente os dados necessitam de uma série de tratamento, antes de serem utilizadas nos modelo de machine learning. Situações como remoção de registros com dados faltantes, normalização de dados para evitar dominância de determinadas características, dentre outros. Segue uma das transformações implementada no PDI para tratar os dados do presente artigo na Figura 5.

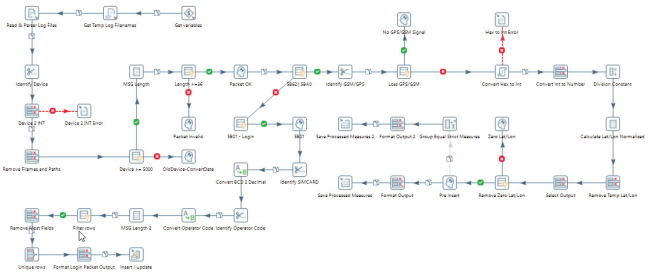


Figura 5: Pentaho Data Integration

4.3 Armazenamento de dados

Para facilitar todo o processo de consulta foi utilizado um banco de dados MySQL.

Terminado o processamento, após o descarte dos dados incompletos ou incorretos, chegamos ao resultando de aproximadamente 70 milhões de medições válidas. Após essa primeira etapa de armazenamento, foram observadas muitas medidas similares (com mesma localização, equipamento e data), ocasionadas principalmente pelo fato do veículo permanecer parado em diversos momentos. Para realizar a agregação dessas medidas e reduzir o volume de dados, foram realizados testes com a agregação de média e mediana

nos campos "GSM"(intensidade do sinal da conexão) e "GPS"(precisão da localização), porém ambos os resultados foram muito parecidos, optamos então por realizar somente a média. Essa agregação consolidou os dados em 20 milhões de medições. Podemos observar a estrutura utilizada no artigo na Tabela 2 e uma amostra dos dados na Tabela 3.

Tabela 2: Estrutura de dados

Coluna	Tipo	Descrição
id	bitint	identificado da linha
timestamp	datetime	date e hora do registro
date	double	data do registro
lat	double	latitude
lon	varchar	longitude
operator	int	Operadora
device	int	Dispositivo criptografado
gsm	int	Intensidade de sinal
gps	int	Precisão do GPS
weight	int	Medidas agregadas

Tabela 3: Amostra da base de dados

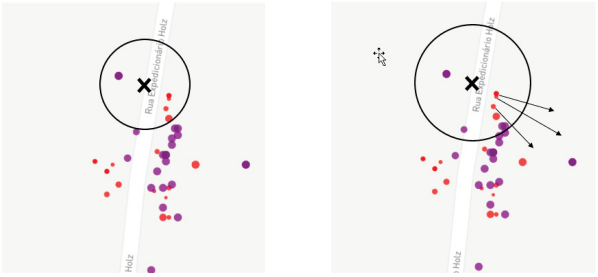
lat	lon	operator	gsm	gps
-23.767217	-46.698067	VIVO	8	12
-23.631333	-46.710227	VIVO	9	34
-23.708053	-46.68847	VIVO	10	16
-30.098697	-51.12632	CLARO	8	18
-30.110997	-51.152503	CLARO	10	12
-29.984623	-51.10687	CLARO	10	8
-30.08344	-51.117537	CLARO	9	11
-30.060803	-51.179513	CLARO	9	23

5 Análise de dados e método de validação

5.1 Análise de dados

Além de recomendar a melhor operadora para uma determinada localidade, um dos objetivos do presente artigo é validar se a intensidade do sinal da conexão é um dado relevante no processo de recomendação. Por conta disso, a principal análise realizada nessa etapa foi em relação a distribuição amostral da intensidade do sinal de conectividade.

Como o método de classificação escolhido (kNN) trabalha com a distância Euclidiana, para utilizar a intensidade do sinal como um fator penalizador, aumentamos a distância das amostras da população em relação a amostra a ser predita em função da intensidade do sinal. Para que se tenha uma visão mais clara desse efeito, apresentamos uma amostra da base de dados onde esse processo de penalização altera a classe predita. Nessas imagens a cor representa a classe e o tamanho do marcador o nível de sinal. Na Figura 6(a) temos o cenário de aplicação do algoritmo KNN tradicional, e na Figura 6(b) imagem temos o cenário do kNN com penalização em função do sinal.



(a) KNN Simples (b) KNN Penalizado

Figura 6: KNN Comparativo

Os equipamentos que realizam a aferição desses dados registram o sinal numa escala de 0 a 255, sendo 0 a ausência total de sinal e 255 o sinal máximo. Entretanto a distribuição das amostras não é linear. A empresa que criou o hardware do equipamento tem a seguinte convenção em relação a intensidade do sinal:

- Ruim - Menor ou igual a 10
- Médio - de 10 a 15
- Bom - Acima de 15

Inicialmente foram realizados testes com penalizações considerando esses parâmetros, porém isso penalizava de forma igual cerca de 90% das amostras, constatando-se que esse não foi um bom método.

Ao observar a distribuição amostral através de um histograma conforme a Figura 7, percebemos uma concentração muito grande das amostras entre os valores sete e quatorze. Por conta disso, optou-se por penalizar progressivamente esses valores, descartando os valores abaixo de seis, considerados como muito ruins, e não penalizar os valores iguais ou maiores que quinze, conforme a Tabela 4. O histograma representado na Figura 7 teve suas extremidades agregadas, valores menores que seis foram agregados na primeira barra e os maiores que quinze em na última barra.

A penalização progressiva foi aplicada conforme a Tabela 4.

Para realizar uma predição, o método kNN realiza um cálculo de distância para cada amostra da base. Nos experimentos do artigo seriam 20 milhões de cálculos para cada execução. Entretanto se levarmos em conta a conectividade de antenas celulares, o sinal só alcança uma certa distância da torre. Então para mitigar esses cálculos excessivos, foi implementada uma busca geográfica na base de dados. Para classificar um determinado ponto, somente são trazidos para o algoritmo os dados num raio de três quilômetros em relação ao ponto escolhido. Isso reduz drasticamente os recursos computacionais necessários para realização de cada predição, sem nenhum impacto na qualidade do resultado.

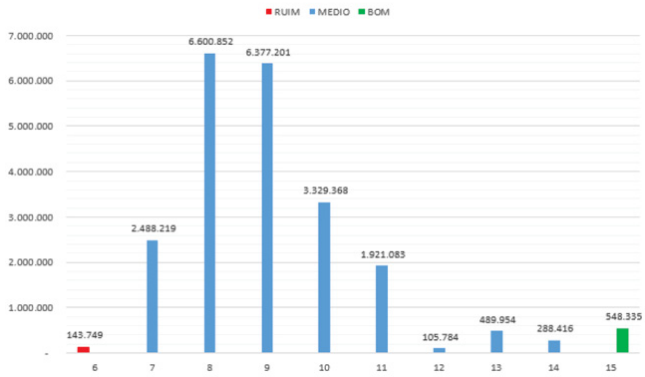


Figura 7: Histograma do nível de sinal

Tabela 4: Tabela de Penalidades

Valor inicial	Valor final	Penalidade
0	6	Descarte
7	7	900%
8	8	800%
9	9	700%
10	10	600%
11	11	500%
12	12	400%
13	13	300%
14	14	200%
15	255	0%

Por fim, a metodologia de penalização possui os seguintes passos:

1. Escolha da localidade a ser predita
2. Busca dos dados num raio de três quilômetros
3. Cálculo da distância Euclidiana do ponto a ser predito e de cada amostra da janela
4. Aplicação da penalidade nas amostras conforme a intensidade do sinal
5. Seleção da classe dominante conforme o parâmetro K previamente definido

5.2 Método de otimização e validação

O método proposto para avaliar se o modelo está conseguindo acertar as classificações é similar a técnica leave-one-out. Após a escolha de uma medida aleatória da base dados, na qual já temos a identificação da classe correspondente, excluímos momentaneamente essa medida dos dados a serem analisados, e realiza-se o processo de predição para determinar a classe. Em seguida, ocorre a comparação entre a classe predita e a classe real. Caso sejam iguais, o modelo irá considerar como um acerto, caso sejam diferentes o modelo irá considerar como um erro. Esse processo é repetido inúmeras vezes, até se obter a acurácia média reportada. Essa acurácia também é utilizada como referência para ajustar os parâmetros do modelo, que determinará se isso aumenta ou diminui sua assertividade.

Na utilização do algoritmo kNN, um ponto fundamental é a escolha do parâmetro K, que está relacionado a quantidade de vizinhos ou amostras mais próximas que o classificador irá considerar para determinar a classe dominante, e consequentemente, a classe do novo elemento a ser predito. Na estrutura de dados utilizada nesse artigo temos somente duas classes, que estão balanceadas na proporção de 70% e 30%, em bases desbalanceadas normalmente é recomendável um K menor, pois o K maior sempre irá tendenciar para a escolha da classe dominante.

O K mais adequado para essa amostra de dados foi obtido através da criação de um search grid, ou seja, a realização de testes com diversos K, iniciando com o valor de 5, e incrementando mais unidades 5, até chegar a 300. Para validação entre os modelos, foram selecionadas mil amostras aleatórias através da distribuição normal, e as mesmas amostras foram submetidas aos classificadores com os diversos K.

Dentre todos os valores do parâmetro K na análise, o que trouxe o melhor resultados foi o parâmetro 5. Seguem os 10 primeiros resultados encontrados na Tabela 5.

Tabela 5: K Search Grid

K	Acurácia
5	87,7%
10	86,5%
15	86,2%
20	86,5%
25	86,7%
30	86,5%
35	86,9%
40	87,3%
45	87,2%
50	86,7%

6 Resultados e Discussões

Um dos principais objetivos do artigo é identificar se existe relevância significativa em utilizar a intensidade do sinal na escolha da melhor operadora. Para realizar a comparação entre os dois modelos foram executados 50 mil experimentos, sendo 25 mil em cada metodologia, todas as amostras foram obtidas através de uma distribuição uniforme. Os experimentos ocorrem em baterias de mil testes, executando sequencialmente cada amostra selecionada com o modelo do kNN tradicional e o modelo do kNN penalizado. O Resultado da bateria de testes realizadas nos dois modelos pode ser observado na Tabela 6

Tem um efeito de afastar as amostras do ponto que se deseja prever, quanto mais fraca a intensidade do sinal.

Tabela 6: Acurária do KNN vs KNN Penalizado

Teste inicial	Teste Final	KNN	KNN penalizada
1	1000	88,5%	87,0%
1.001	2.000	89,3%	89,7%
2.001	3.000	87,8%	88,1%
3.001	4.000	87,7%	88,1%
4.001	5.000	89,3%	89,4%
5.001	6.000	87,0%	85,8%
6.001	7.000	91,3%	89,9%
7.001	8.000	91,0%	90,3%
8.001	9.000	90,0%	89,4%
9.001	10.000	91,5%	90,4%
10.001	11.000	89,8%	88,9%
11.001	12.000	89,0%	88,7%
12.001	13.000	88,5%	88,6%
13.001	14.000	88,3%	88,8%
14.001	15.000	91,4%	90,2%
15.001	16.000	90,0%	89,7%
16.001	17.000	90,8%	89,9%
17.001	18.000	90,1%	89,6%
18.001	19.000	90,8%	89,8%
19.001	20.000	90,1%	89,6%
20.001	21.000	89,4%	89,7%
21.001	22.000	88,3%	87,7%
22.001	23.000	90,6%	89,3%
23.001	24.000	87,0%	88,0%
24.001	25.000	88,6%	88,3%
Média		**89,4%**	**89,0%**

A acurácia do modelo kNN tradicional (não penalizado) teve um desempenho levemente superior na maior parte dos testes realizados, se comparado ao modelo que utiliza a intensidade do sinal como um fator penalizador, portanto a informação do nível de sinal foi considerada não preponderante na recomendação da melhor operadora no modelo proposto. Se considerarmos também que o modelo penalizado tem um custo computacional maior, isso reforça ainda mais a escolha do modelo kNN tradicional para o cenário apresentado. Isso se deve principalmente ao fato da localização geográfica das amostras ter um peso muito mais expressivo na tomada de decisão, se comparado a intensidade do sinal.

É importante ressaltar que essa conclusão não significa que a informação de intensidade de sinal não possui relevância em outros contextos ou bases de informação.

7 Trabalhos futuros

Existem melhorias posteriores que poderão ser realizadas no modelo, assim como trabalhos relacionados relevantes. Uma extensão interessante a ser executada é levar em consideração a precisão do GPS como um fator de penalização da localização da amostra. Um outro ponto é considerar os pesos da consolidação da amostra, cerca de 70 milhões de amostras foram consolidadas em 20 milhões, e o algoritmo atualmente implementado não está levando em conta a quantidade de amostras consolidadas em cada registro.

Outra possibilidade é realizar a inferência da localização aproximada das estações rádio bases (também conhecidas com ERB) a partir dos dados, pois os terminais tendem a possuir um sinal mais intenso a medida que se aproximam da antena, e isso acaba estimulando a geração de "clusters"de medidas com sinal mais intenso ao redor de antenas. O cruzamento dessas informações com bases públicas da ANATEL (Agência Nacional de Telecomunicações) também podem auxiliar no mapeamento e localização das antenas. A informação de localização das antenas, pode ser utilizada como entrada no modelo de classificação, melhorando a assertividade do modelo ao levar em conta mais essa dimensão de informação.

Agradecimentos

Gostaria de agradecer a minha esposa Tamiris por todo apoio e incentivo para meus projetos. Ao professor Luiz Eduardo Soares de Oliveira pela orientação no processo de concepção e elaboração, tanto dos experimentos, quanto do artigo. E a toda equipe de professores da Especialização de Data Science e Big Data da Universidade Federal do Paraná (UFPR) que através da união dos departamentos de Matemática (DMAT), Estatística (DEST) e Informática (DINF) estruturou a especialização que trouxe o embasamento necessário para realização deste artigo.

Referências

[1] Gordon Moore. Cramming more components onto integrated circuits, 1965.

[2] Rong Chen et al. Atomic level deposition to extend moore's law and beyond. *International Journal of Extreme Manufacturing*, 2(2):022002, 2020.

[3] Michael J Thacker and Wesley W Wilson. Telephony choices and the evolution of cell phones. *Journal of Regulatory Economics*, 48(1):1–25, 2015.

[4] STATISTA.COM (org.). Forecast number of mobile devices worldwide from 2019 to 2023 (in billions). 2019 (accessed August 01, 2020).

[5] gsma.com. Gsm coverage global network. 2020 (accessed August 24, 2020).

[6] Marcus L Roberts, Michael A Temple, Robert F Mills, and Richard A Raines. Evolution of the air interface of cellular communications systems toward 4g realization. *IEEE Communications Surveys & Tutorials*, 8(1):2–23, 2006.

[7] telecom.com. Estatísticas de celulares no brasil. 2020 (accessed August 31, 2020).

[8] ibge.gov.br. Projeção da população do brasil e das unidades da federação. 2020 (accessed August 31, 2020).

[9] iot analytics.com. Areas iot. 2020 (accessed August 24, 2020).

[10] celplan.com.br. Tecnologias wifi. 2020 (accessed August 24, 2020).

[11] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.