AVALIAÇÃO DA VIABILIDADE DE UM SISTEMA DE IDENTIFICAÇÃO DE CHAMADAS FALSAS EM CHAMADAS DE EMERGÊNCIA

Robert Cristiano Almeida Viana
Prof.ª Dra. Luciana Balieiro Cosme

Sumário

Introdução

- Ministério da Saúde publicou duas portarias em 2003:
  - GM/MS nº 1863: Política Nacional de Atenção às Urgências
  - GM/MS nº 1864: Componente pré-hospitalar móvel
- Disponível em todo território nacional;
- Serviço gratuito;
- Grandes impactos positivos.

Problemas

- Crescimento alarmante de trotes e chamados falsos;
- Ocupação das linhas telefônicas;
- Mobilização indevida de recursos;
- Prejuízos financeiros ao estado;
- Dificuldade para punir os malfeitores.

Desafios

- Cerca de dois meses de pesquisas e consultas a literatura;
- Conteúdo das chamadas sensíveis a privacidade;
- Dificuldades para se obter parcerias com os gestores desses dados;
- Bases de dados públicas com poucas informações e em outros idiomas.

Motivação

- Um estudo de 2014 apontou que a soma dos gastos chega a R$ 1 bilhão por ano.
- Aproximadamente R$ 500,00 por chamado emergencial incompleto;
- Mesmo com triagem, impossível bloquear todas as ligações indevidas;
- Campanhas de conscientização não geram grandes impactos;
- Uma área que carece de ferramentas computacionais.

Objetivos gerais

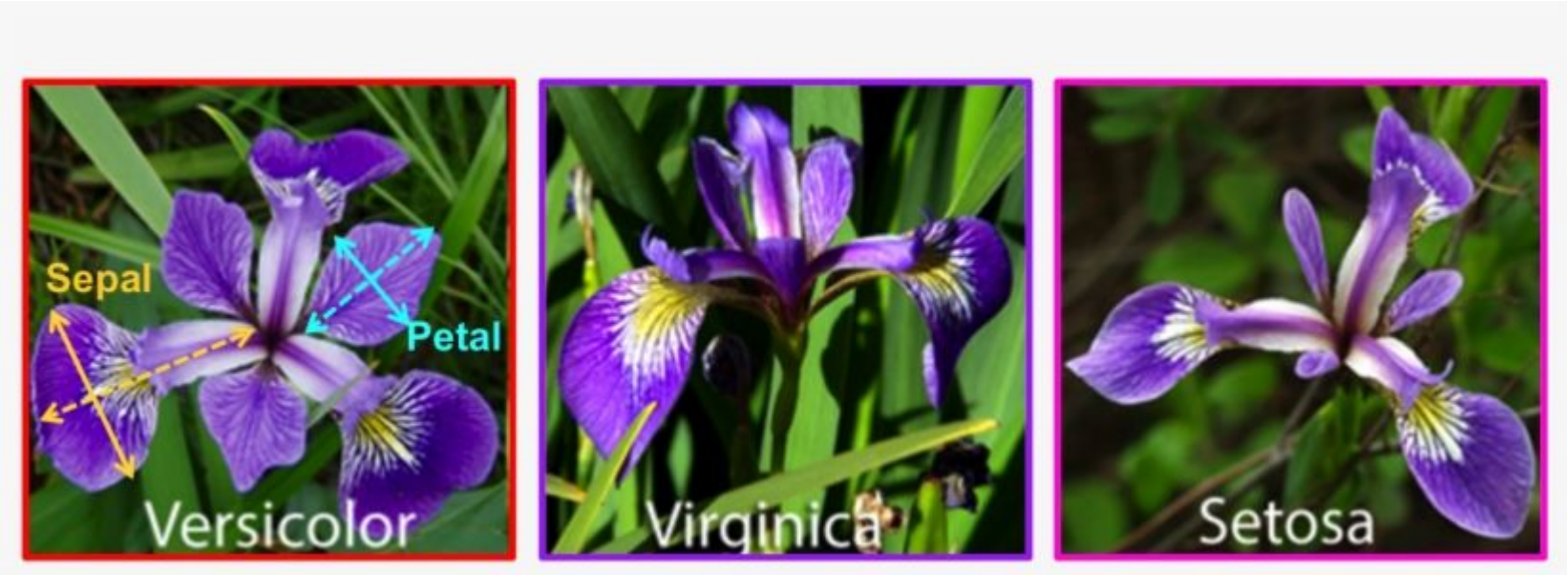
- Avaliar a viabilidade de se produzir sistemas computacionais capazes de identificar corretamente chamadas falsas;
- Produzir documentos e argumentos para estimular uma maior participação e auxílio dos gestores dos dados;
- Elencar quais dados são suficientes ou necessários para a identificação correta.

Sumário

Classificação de dados

- Abrange inúmeras aplicações em diversos cenários no nosso dia a dia:
  - Diagnóstico de doenças;
  - Identificação de objetos;
  - Categorização de seres vivos e espécies.
- Consiste em determinar um ou vários rótulos a um objeto de análise, baseado em um conjunto de características extraídas dele;
- De acordo com um conhecimento prévio, adquirido por uma fase de treinamento.

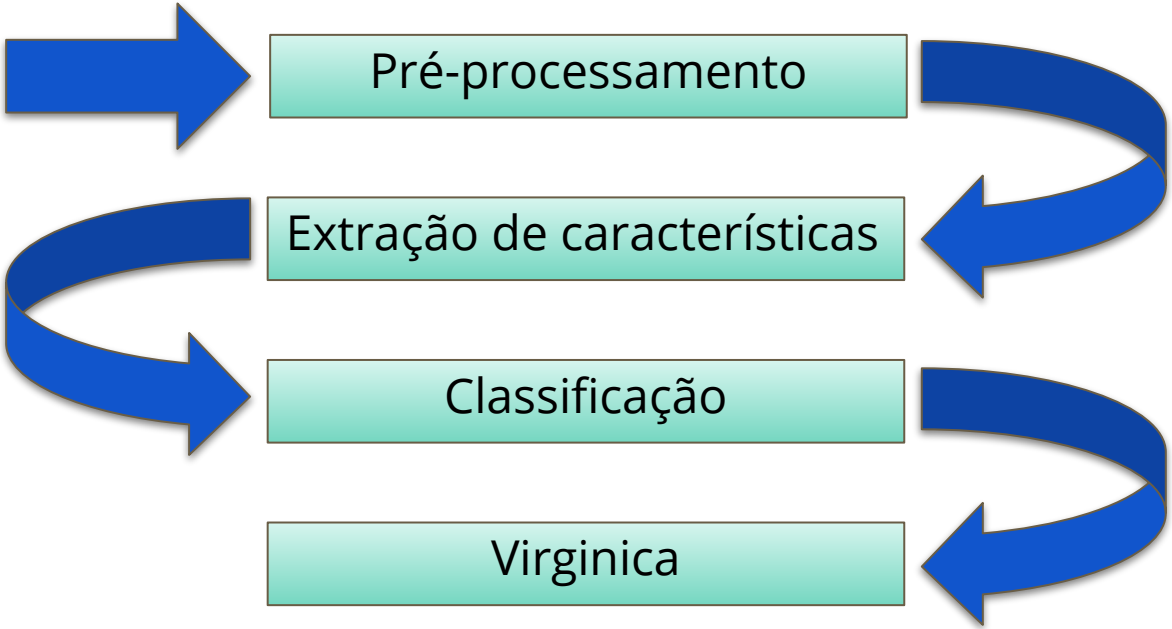
Problema da flor de Iris



Classificação de dados

- Versicolor
- Virginica
- Setosa

Classificação de dados



Algoritmos de classificação

1. **Redes Neurais Artificiais**
2. K-vizinhos mais próximos

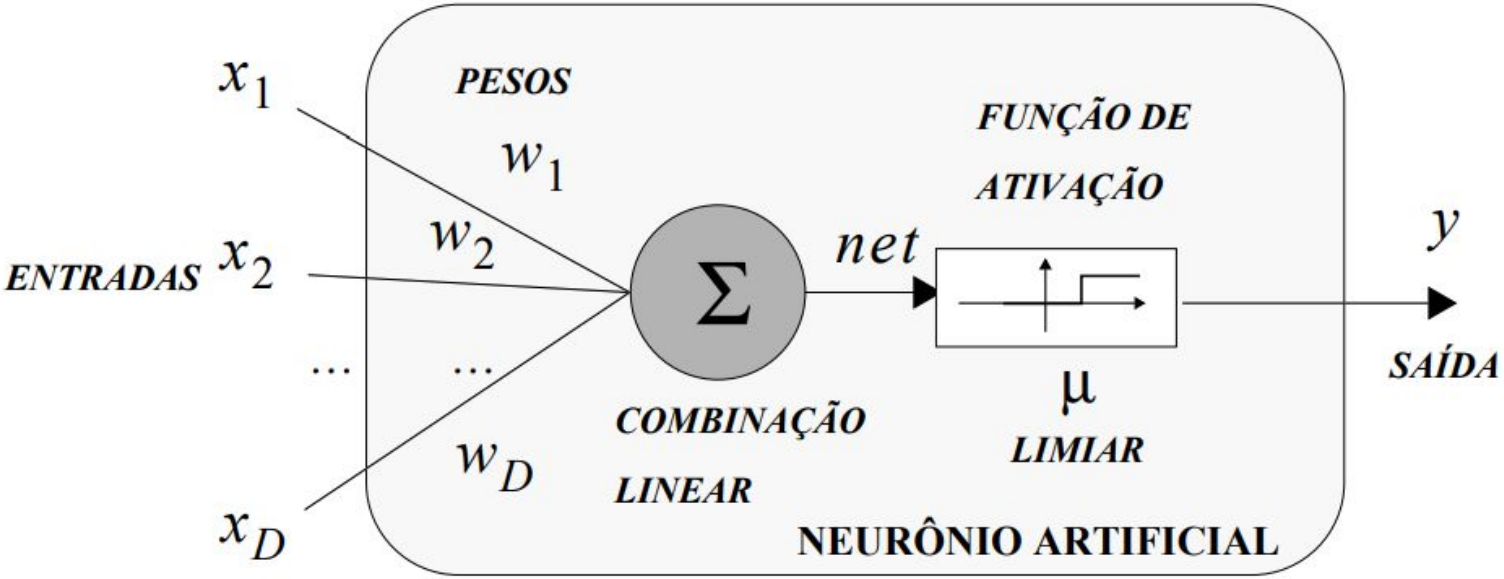
Redes Neurais Artificiais - RNA

- É uma tentativa de copiar a estrutura e o funcionamento do cérebro humano para uma estrutura artificial, transformando as redes neurais biológicas em redes neurais artificiais;
- Por ser uma ferramenta flexível, poderosa e possuir uma grande capacidade de processamento, apresenta resultados excepcionais nas mais diversas aplicações da literatura;

Neurônio artificial

- Cada neurônio é considerado como uma unidade básica de processamento;
- Quando estimulados por sinais de entrada, emitem sinais de saída, como reação;
- Um neurônio é conectado a outro neurônio através de uma conexão sináptica;
- A principal função de um neurônio é armazenar conhecimento.

Neurônio artificial

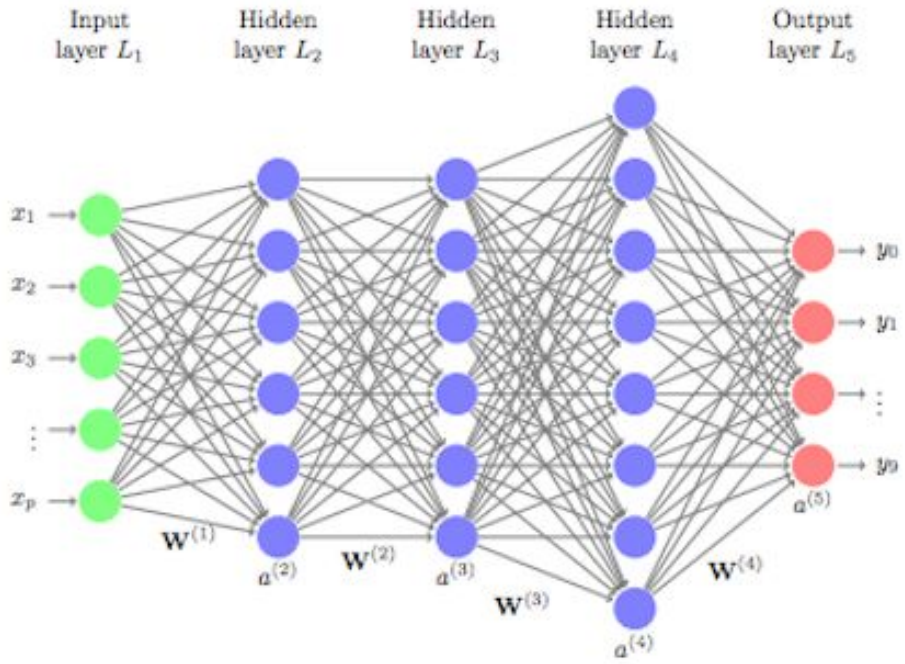


Redes neurais de multicamadas

- Como geralmente a capacidade de processamento de um único neurônio não soluciona tarefas de classificação mais complexas;
- Surge a necessidade do uso conjunto de outros neurônios;
- Cada camada que não esteja conectada às entradas ou saídas, é chamada de camada intermediária ou oculta.

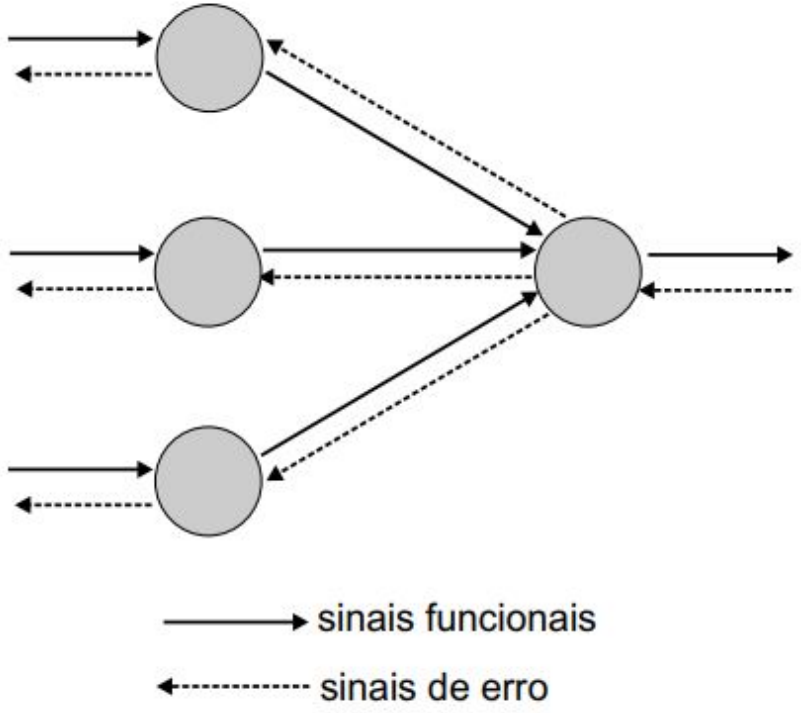
Feed-Forward Network - FNN

- É uma rede multicamadas unidirecional;
- Comumente utilizadas para o reconhecimento de padrões e classificação de dados;
- Topologia mais utilizada na área de aprendizado de máquina.



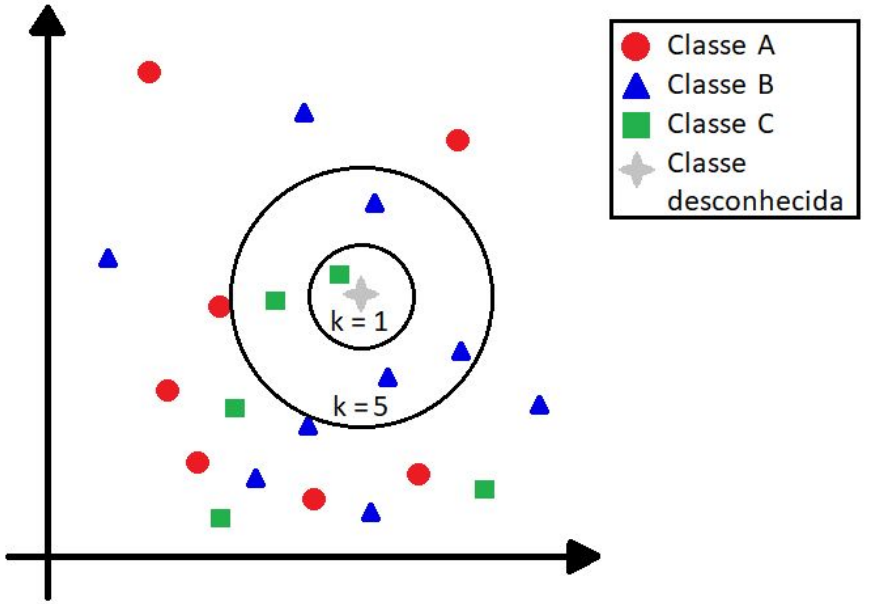
Processo de aprendizagem FNN

- Um método amplamente utilizado para o treinamento para redes FNN é o algoritmo de *Backpropagation*;
- O objetivo é ajustar os pesos sinápticos da rede, minimizando o erro gerado pelo processamento da rede.



K-vizinhos mais próximos

- Têm como objetivo determinar um rótulo à uma amostra, baseando-se em outras amostras vizinhas;
- Ao final, por votação majoritária, a amostra se torna uma da classe dominante entre os *k* vizinhos mais próximos;
- Resultados dependentes da escolha do *k*.



Processamento de linguagem natural

- Têm como objetivo tratar os aspectos presentes da comunicação humana;
- O PLN consiste no emprego de um conjunto de técnicas computacionais para aprender, entender e reproduzir uma linguagem natural;
- Durante o processo de aprendizado, o PLN pode utilizar de conhecimentos linguísticos e métodos estatísticos para resolvê-los;
- Uma das tarefas possíveis no PLN é a classificação de texto.

Classificação de texto

- Antes de qualquer operação sobre os dados textuais, é necessário verificar a normatização do texto;
- Convertê-lo a um formato padrão, eliminando erros ortográficos e redundância, facilitando a manipulação dos dados;
- Sub-etapas comuns da normalização são:
  - Tokenization
  - Lemmatization

Tokenization

- No processo de *tokenization*, deseja-se separar as palavras de um texto, tratando-as isoladamente;
- Exemplo: "*Incêndio em casa no Todos os Santos*";
- Importante o tratamento das palavras no contexto em que foram inseridas;

Lemmatization

- É o processo de deflexionar uma palavra para o seu lema;
- Exemplo: "*estudou, estudava, estudaria, estudará*". São formas do mesmo lema: estudar.
- Útil quando deseja-se ver o uso das palavras sem considerar o contexto de flexão ou usar-los como índices;
- Um processo potencialmente custoso, de acordo com o tamanho da entrada.
- Como alternativa, surge a análise morfológica dos dados: *stemming.*

Classificação de texto

- Além de uma etapa de normalização dos dados, uma etapa de extração e seleção de características é essencial:
  - Reduz a dimensão do problema;
  - Aumenta a velocidade de tratamento;
  - Aumento de precisão na predição e classificação;
  - Facilita a visualização dos dados.
- Uma vez escolhido um conjunto de características, é possível aplicar técnicas de classificação de texto, tais como as RNA e o KNN;
- Portanto, decidir a qual conjunto de classes previamente conhecidas o texto possui.

Sumário

Classificadores de texto

- **Uysal & Gunal (2012):** propuseram uma nova metodologia para extração de características em bases textuais:
  - Revista online Reuters
  - Publicações em fóruns de discussão
  - Mensagens de texto - SMS
  - Conteúdo de *e-mail*
- RNA se destacou dentre os algoritmos, com altas taxas de acerto, superiores a 90%.

Classificadores de texto

- **Joachims (1998):** identificou um padrão presente em bases textuais e deduziu que qualquer algoritmo de classificação que trabalhasse bem nessas circunstâncias, também poderia classificar eficientemente um texto;
- KNN mostrou-se um classificador de texto sólido.

Chamadas emergenciais falsas

- **Waseem et al. (2010):** uma análise nas chamadas emergenciais de Punjab no Paquistão;
- Cerca de 97% do total de chamadas foram falsas:
  - 91% como trote;
  - 7% por buscas de informações;
  - 1% por engano;
  - 0,13% forjou uma situação inexistente.
- Implantação de um sistema de monitoramento de chamadas e o bloqueio das linhas;

Chamadas emergenciais falsas

- **Rashford & Isoardi (2010):** Após os altos índices de trotes, alterações na legislação vigente da Austrália foram feitas, aplicando penas mais severas:
  - Até 1 ano de prisão;
  - Até $ 10.000 de multa.
- Aplicação de um sistema manual: Sistema de Prioridades para Despacho Médico - MPDS;
- Carecendo ainda do uso de ferramentas computacionais para automatizar a identificação de falsos chamados;

Chamadas emergenciais falsas

- **Kuroiwa et al. (2004) :** propuseram um sistema de identificação automática de trotes e chamadas falsas, baseado no reconhecimento de fala;
- O usuário é solicitado a repetir uma palavra em japonês, com o intuito de validar se ele realmente compreende o idioma da região a qual está ligando;
- 97% de taxa de acerto e 93% de taxa de rejeição.

Privacidade dos dados e ética

- Por sua especificidade, os conteúdos e relatórios das bases de dados públicas disponíveis são rasos, carecendo de detalhes;
- Informações que possam identificar, de alguma forma, o usuário, são omitidas/ocultas;
- Além da obrigação legal contida nos códigos de ética profissional, como também no código penal brasileiro, faz-se necessário uma etapa sujeita a um Comitê de Ética em Pesquisa;
- Gestores dos dados, desconhecem o possível potencial que tais ferramentas poderiam proporcionar ao processo.

Sumário

Metodologia

- Após consultas em repositórios de artigos e publicações acadêmicas, não foram encontrados trabalhos sobre sistemas que identifiquem chamadas falsas baseadas no conteúdo da chamada;
- Logo, por essa incógnita sobre a viabilidade, este trabalho se propõem a levantar a viabilidade desses sistemas, com base no conteúdo da chamada e por meio de classificadores;
- Ao fim, com os resultados obtidos, será possível mensurar preliminarmente a viabilidade de tais sistemas;

Metodologia

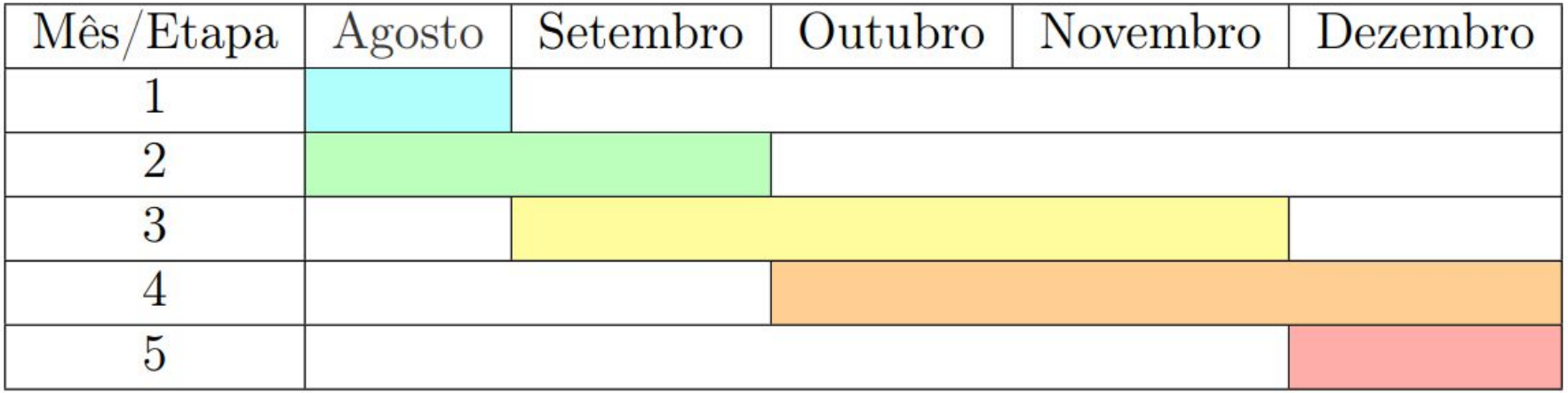
- Passos necessários para alcançar os objetivos:
  1. Obter, analisar e compreender a base de dados;
  2. Identificar, selecionar e extrair as características mais importantes, que possam influenciar na classificação;
  3. Implementar e treinar os algoritmos de classificação, ajustando seus parâmetros empiricamente;
  4. Executar os algoritmos para dados de teste e avaliar a predição alcançada por cada;
  5. Escrever, apresentar os resultados e concluir;

Base de dados

- ***Call Data*:** do centro de comunicações do departamento de polícia de Seattle, EUA;
- Representa os relatórios das chamadas emergenciais originados pela comunidade local;
- Armazenado cerca de 95% de todas as chamadas recebidas;
- Existe cerca de 3,9 milhões de ocorrências únicas, extraídas desde 2009;
- Atualizada diariamente.

Cronograma de atividades

1. Análises sobre a base de dados;
2. Seleção e extração de características;
3. Implementação dos algoritmos e testes;
4. Validação de resultados e discussão;
5. Escrita, conclusão e apresentação.



Mês/Etapa	Agosto	Setembro	Outubro	Novembro	Dezembro
1					
2					
3					
4					
5					

Referências bibliográficas

- Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. Computing in Science & Engineering, 9(3):90–95.
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning, pp. 137–142. Springer.
- Kuroiwa, S.; Naito, M.; Nakamura, M.; Sakayori, S. & Mukasa, T. (2004). Automatic prank call rejection system for home country direct based on speech recognition technology. Electronics and Communications in Japan (Part II: Electronics), 87(4):44–52.
- Loper, E. & Bird, S. (2002). NLTK: the natural language toolkit. CoRR, cs.CL/0205028.

Referências bibliográficas

- McCulloch, W. S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4):115–133.
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830.
- Rashford, S. & Isoardi, K. (2010). Optimizing the appropriate use of the emergency call system, and dealing with hoax callers. Emergency Medicine Australasia, 22(5):366– 367.

Referências bibliográficas

- Rauber, T. W. (2005). Redes neurais artificiais. Universidade Federal do Espírito Santo.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6):386.
- Uysal, A. K. & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. Knowledge-Based Systems, 36:226–235.
- Van Der Walt, S.; Colbert, S. C. & Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. Computing in Science & Engineering, 13(2):22.

Referências bibliográficas

- van Rossum, G. (1995). Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam.
- Waseem, H.; Durrani, M. & Naseer, R. (2010). Prank calls: a major burden for an emergency medical service. Emergency Medicine Australasia, 22(5):480–480.
- Wes McKinney (2010). Pandas: Data Structures for Statistical Computing in Python
- Widrow, B. & Hoff, M. E. (1960). Adaptive switching circuits. Technical report, Stanford Univ Ca Stanford Electronics Labs.

OBRIGADO!

“If it wasn’t hard, everyone would do it. It’s the hard that makes it great”
(Tom Hanks)