**IA PARA PROBABILIDADE DE ALAGAMENTOS NO BUTANTÃ**

SISTEMAS DE CONTROLE E INTELIGÊNCIA ARTIFICIAL

Gustavo Da Silva Hossein - 821148931

Izabelle Silva de Souza - 821144894

João Pedro Silva de Oliveira - 821225047

Nathalia Cabral Rodrigues de Almeida - 823130648

Ricardo Brito Ponticceli Prieto - 821129414

São Paulo

2025

**SUMÁRIO**

1. **Introdução**
2. **Objetivo  
     
    2.1 Objetivo Geral  
     
    2.2 Objetivos Específicos**
3. **Fundamentos Teóricos  
     
    3.1 Inundações Urbanas, Drenagem e Previsibilidade  
     
    3.2 Intensidade de Chuva, Limiares e Classes de Risco  
     
    3.3 Dados Pluviométricos, Séries Temporais e API  
     
    3.4 Algoritmos de Aprendizado de Máquina: Regressão Logística e Random Forest  
     
    3.5 Desbalanceamento de Classes e Oversampling**
4. **Materiais e Métodos  
     
    4.1 Área de Estudo  
     
    4.2 Fontes de Dados  
     
    4.3 Pré-processamento da Chuva (CGE Butantã)  
     
    4.4 Dataset do Modelo de 3 Classes (Risco por Chuva)  
     
    4.5 Construção das Labels com Ocorrência de Alagamento (GeoSampa + CGE)  
     
    4.6 Filtragem para Dias Chuvosos e Features do Modelo Binário  
     
    4.7 Oversampling e Treino dos Modelos Binários**
5. **Discussão  
     
    5.1 Por que Dois Modelos?  
     
    5.2 Dificuldade Intrínseca de Prever Alagamentos  
     
    5.3 Ganhos e Limitações dos Resultados  
     
    5.4 Dados Adicionais que Poderiam Melhorar a Acurácia  
     
    5.5 Aspectos Didáticos e Narrativos do Projeto**
6. **Encerramento**
7. **Referências**

**Relatório Técnico – Protótipo de IA para Predição de Risco de Inundação no Butantã**

**1. Introdução**

As inundações urbanas em grandes cidades brasileiras, como São Paulo, vêm se tornando mais frequentes e intensas, em grande parte devido à combinação entre mudanças climáticas, urbanização acelerada e infraestrutura de drenagem muitas vezes insuficiente. Estudos apontam aumento tanto na média de chuva anual quanto no número de eventos extremos (por exemplo, dias com mais de 80–100 mm de chuva em 24h) na capital paulista.

Ao mesmo tempo, prever com precisão quando um alagamento vai ocorrer em um ponto específico é extremamente difícil. Não basta chover “muito”: o resultado depende de fatores locais como:

* capacidade das galerias de drenagem;
* estado de manutenção de bueiros e bocas de lobo;
* presença de lixo e entulho;
* grau de impermeabilização do solo;
* ocupação de áreas de várzea e fundos de vale.

Esses fatores, em geral, **não aparecem diretamente em bases públicas abertas**, o que limita qualquer modelo puramente estatístico.

Diante desse cenário, este trabalho desenvolve um protótipo de sistema de IA para **predição de risco de inundação no bairro do Butantã**, na cidade de São Paulo, usando somente dados públicos. A escolha de focar em uma região específica traz vantagens:

* reduz a necessidade de calibrar diferentes tipos de solo e usos em várias bacias;
* permite trabalhar com **um único posto pluviométrico** (CGE Butantã) e uma **camada de ocorrências de alagamento** (GeoSampa);
* viabiliza a entrega de um MVP funcional dentro dos limites de tempo e escopo da disciplina “Sistemas de Controle e IA”.

Durante o desenvolvimento, ficou claro que **não há dados suficientes hoje para treinar um modelo robusto de ocorrência de alagamentos** puramente a partir de chuva, e que a natureza do fenômeno é altamente condicionada por fatores externos (infraestrutura, manutenção, ocupação, etc.). Isso levou à adoção de **dois modelos complementares**, com papéis diferentes:

1. Um modelo de **3 classes de risco de inundação por intensidade de chuva**, baseado em faixas de precipitação diária (chuva acumulada em 24h).
2. Um modelo **binário condicionado a dias chuvosos**, treinado com ocorrências reais de alagamento do GeoSampa e chuva do CGE, usando **engenharia de features + oversampling** para lidar com a raridade de eventos.

Ambos foram encapsulados em protótipos de interface em **Streamlit**, compondo um MVP didático de sistema de alerta local para o Butantã.

**2. Objetivo**

**Objetivo geral**

Desenvolver e documentar um protótipo de sistema de IA para **avaliação de risco de inundação** no bairro do Butantã, combinando classificadores baseados em chuva com dados reais de ocorrências de alagamento, utilizando dados públicos e ferramentas de código aberto.

**Objetivos específicos**

1. Construir uma série histórica diária de chuva (2020–2025) para o posto CGE Butantã a partir de planilhas oficiais.
2. Definir faixas de chuva diária que representem três classes de risco local:
   * Sem risco;
   * Risco de Inundação Transitável;
   * Risco de Inundação Intransitável.
3. Treinar um modelo de classificação (Regressão Logística e Random Forest) para **aprender** essas três classes de risco a partir de chuva e lags, e disponibilizá-lo em um app Streamlit.
4. Integrar dados de **ocorrências de alagamento** do GeoSampa (camada risco\_ocorrencia\_alagamento) com a chuva do CGE Butantã, gerando uma série de “dias com evento / sem evento” para a subprefeitura do Butantã.
5. Desenvolver um segundo modelo binário para **“alarme de alagamento em dias chuvosos”**, com engenharia de features (lags, acumulados, API) e oversampling para mitigar o desbalanceamento extremo entre dias com e sem alagamento.
6. Discutir criticamente as limitações de dados, a dificuldade física de prever alagamentos e propor caminhos futuros de melhoria (mais dados, novas bases, ajustes de limiares, etc.).

**3. Fundamentos teóricos**

**3.1 Inundações urbanas, drenagem e previsibilidade**

Inundações urbanas ocorrem quando o escoamento superficial de água de chuva excede:

* a capacidade de infiltração do solo; e/ou
* a capacidade de escoamento do sistema de drenagem (galerias, bocas de lobo, canais, rios).

Em áreas densamente urbanizadas, como o Butantã e região, a grande quantidade de superfícies impermeáveis (asfalto, lajes, telhados) combinada com redes de drenagem muitas vezes subdimensionadas ou mal mantidas favorece o acúmulo rápido de água em vias e depressões topográficas.

Não existe uma relação simples do tipo **“chove X mm ⇒ alaga”**, porque:

* a resposta depende da **capacidade hidráulica** do sistema de drenagem;
* bueiros podem estar **entupidos** por lixo e folhas;
* a topografia local (baixadas, fundos de vale) canaliza o escoamento;
* a manutenção (limpeza, desassoreamento) muda no tempo;
* a ocupação irregular de margens e várzeas piora a condição.

Por isso, muitos estudos de inundação urbana usam **modelos híbridos**, combinando:

* modelos físicos/hidrológicos (chuva–vazão, escoamento em bacias, hidráulica de galerias);
* modelos estatísticos ou de aprendizado de máquina para **ajustar** o risco em função de dados históricos.

Neste projeto, por limitação de dados e escopo, foi seguida **uma abordagem puramente estatística** (aprendizado supervisionado) aplicada a séries de chuva + ocorrência de alagamentos.

**3.2 Intensidade de chuva, limiares e classes de risco**

Órgãos de meteorologia como o INMET classificam chuvas por faixas de intensidade e acumulado. De forma simplificada:

* chuvas fracas: acumulados diários baixos, com pouco impacto;
* chuvas moderadas a fortes: acumulados diários na faixa de **20–50 mm**;
* chuvas muito fortes / extremas: acumulados acima de **60–70 mm/24h**, frequentemente associados a impactos significativos.

Estudos de precipitação extrema no Brasil mostram que:

* o **percentil 90** de chuva diária (eventos intensos) costuma cair na faixa 20–40 mm em diversas regiões;
* o **percentil 99** (eventos muito extremos) costuma ficar acima de 60–70 mm/dia.

Em contexto urbano, trabalhos que analisam chuva e impacto (alagamentos, movimentos de massa) frequentemente apontam **limiares em torno de 60 mm/24h** como associados a danos significativos e ocorrência de alagamentos em cidades como Recife, Salvador e Curitiba.

Com base nessa literatura, nas classificações de órgãos oficiais e na prática em grandes centros urbanos, foram definidas as seguintes faixas de risco para o Butantã:

* **Classe 0 – Sem risco:**  
  chuva em 24h < 20 mm
* **Classe 1 – Risco de Inundação Transitável:**  
  20 mm ≤ chuva em 24h < 60 mm
* **Classe 2 – Risco de Inundação Intransitável:**  
  chuva em 24h ≥ 60 mm

Esses limites **não garantem** que todo dia com 60 mm resultará em inundação intransitável, mas funcionam como um **indicador de severidade** razoável para um sistema de alerta de risco, alinhado a práticas de classificação de extremos.

**3.3 Dados pluviométricos, séries temporais e API**

A chuva é medida em milímetros (mm), que correspondem à **altura de lâmina de água** sobre uma área. No projeto foi utilizada **chuva diária (24h)**, acumulada no posto do CGE Butantã, no período de 2020 a 2025, a partir de planilha oficial.

Como o alagamento não depende apenas da chuva do dia, mas também:

* da **chuva dos dias anteriores**;
* do **estado de saturação** do solo / sistema;

a literatura costuma empregar:

* **lags** de 1, 2 e 3 dias (chuva\_lag1, chuva\_lag2, chuva\_lag3);
* **acumulados** em janelas móveis (acum\_3d, acum\_7d, acum\_30d);
* **API – Antecedent Precipitation Index**, índice que resume a chuva antecedente com decaimento exponencial:

onde:

* : índice de precipitação antecedente no dia ;
* : chuva do dia (mm);
* : fator de decaimento (0 < k < 1).

Neste projeto adotou-se **k = 0,9**, valor típico encontrado na literatura para solos com resposta intermediária, apenas como referência simplificada.

**3.4 Algoritmos de aprendizado de máquina: Regressão Logística e Random Forest**

Foram escolhidos dois algoritmos amplamente usados em problemas de classificação binária e multiclasse, inclusive em aplicações hidrológicas:

**Regressão Logística**

A **Regressão Logística** modela a probabilidade de uma classe (por exemplo, Y=1) como:

onde é o vetor de features (chuva, lags, etc.) e são os coeficientes aprendidos.

Para mais de duas classes (como no modelo de 3 classes de risco), utiliza-se a versão **multinomial** (softmax).

Vantagens:

* modelo relativamente simples e interpretável;
* robusto com poucos dados;
* boa baseline para comparação com modelos mais complexos.

**Random Forest**

A **Random Forest** é um método de ensemble baseado em:

* construção de muitas **árvores de decisão**, cada uma treinada em uma amostra bootstrap do conjunto de treino (bagging);
* em cada nó, considera apenas uma amostra aleatória de variáveis para divisão.

Essa aleatoriedade aumenta a diversidade das árvores, o que, por sua vez, melhora a generalização do ensemble.

Vantagens:

* geralmente boa performance “de saída”, com pouca necessidade de ajuste fino;
* captura **não linearidades** e **interações** entre variáveis;
* robusta a ruídos e a escalas diferentes das features.

Diversos trabalhos de mapeamento de suscetibilidade a inundação utilizam Random Forest com sucesso, inclusive em bacias urbanas brasileiras.

**3.5 Desbalanceamento de classes e oversampling**

Um problema central do modelo binário é o **desbalanceamento extremo** entre dias com e sem alagamento:

* cerca de **2131 dias** na série 2020–2025 (chuva no CGE Butantã);
* apenas **10 dias com ocorrência registrada** de alagamento na subprefeitura do Butantã (flag\_evento=1) na camada GeoSampa.

Se treinarmos um modelo diretamente assim, ele tende a aprender:

“Sempre responda 0 (sem alagamento)”

e alcança uma acurácia alta (porque quase todos os dias são de fato 0), mas **nunca prevê eventos reais** – recall = 0 para a classe 1.

Para mitigar isso, utilizou-se **oversampling da classe minoritária** apenas no conjunto de treino:

1. Contar (dias com evento) e (sem evento).
2. Definir um alvo de proporção, por exemplo target\_ratio = 0.2, significando ~20% de exemplos positivos no treino.
3. Calcular o número desejado de positivos:
4. Replicar (com reposição) exemplos positivos até atingir .

Isso **não cria novos tipos de eventos**, mas força o algoritmo a “prestar atenção” na classe positiva durante o ajuste dos parâmetros.

**4. Materiais e Métodos**

**4.1 Área de estudo**

A área de estudo é a **Subprefeitura do Butantã**, na Zona Oeste da cidade de São Paulo. A região inclui:

* áreas residenciais;
* importantes eixos viários;
* grandes áreas institucionais, como a USP, que também influenciam a dinâmica de tráfego e drenagem.

Para este MVP, o Butantã foi considerado como **uma única unidade de risco**, associada a:

* um **posto pluviométrico** do **CGE Butantã**;
* registros de alagamentos georreferenciados da camada geoportal:risco\_ocorrencia\_alagamento do GeoSampa, filtrados por nm\_subprefeitura ≈ “Butantã”.

**4.2 Fontes de dados**

**a) Chuva – CGE-SP (CGE Butantã)**

* Fonte institucional: Centro de Gerenciamento de Emergências Climáticas da Prefeitura de São Paulo (CGE-SP).
* Arquivo tratado no projeto:  
  data/processed/CGESP-PROCESSADOS/CGESP\_BUTANTA\_2020\_2025.xlsx
* Estrutura:
  + colunas de **Ano**, **Mês** (em português: Jan, Fev, etc.), **Posto**;
  + colunas de dias (1–31) contendo o acumulado diário de chuva (mm).

**b) Ocorrências de alagamento – GeoSampa**

* Portal GeoSampa – Sistema de informações geográficas da Prefeitura de São Paulo.
* Serviço WFS:  
  http://wfs.geosampa.prefeitura.sp.gov.br/geoserver/geoportal/wfs
* Camada utilizada: geoportal:risco\_ocorrencia\_alagamento.
* Campos relevantes:
  + dt\_ocorrencia – data da ocorrência;
  + nm\_subprefeitura – nome da subprefeitura (usado para filtrar Butantã);
  + geometry – coordenadas do ponto de alagamento (não exploradas neste MVP, mas úteis futuramente).

**4.3 Pré-processamento da chuva (CGE Butantã)**

Implementado em Python (por exemplo, em carregar\_chuva\_cge):

1. **Leitura do Excel:**  
   Uso de pandas.read\_excel na aba "BUTANTA\_2020\_2025".
2. **Conversão de mês textual para numérico:**
3. mes\_map = {
4. "Jan": 1, "Fev": 2, "Mar": 3, "Abr": 4, "Mai": 5, "Jun": 6,
5. "Jul": 7, "Ago": 8, "Set": 9, "Out": 10, "Nov": 11, "Dez": 12
6. }
7. df\_raw["MesNum"] = df\_raw["Mes"].map(mes\_map)
8. **“Derretimento” das colunas de dia:**
   * identificar colunas inteiras de 1 a 31;
   * usar pd.melt:
9. df\_long = df\_raw.melt(
10. id\_vars=["Ano", "Mes", "MesNum", "Posto"],
11. value\_vars=day\_cols,
12. var\_name="Dia",
13. value\_name="chuva\_mm"
14. )
15. **Construção da data:**
16. df\_long["Dia"] = df\_long["Dia"].astype(int)
17. df\_long["data"] = pd.to\_datetime(
18. dict(year=df\_long["Ano"], month=df\_long["MesNum"], day=df\_long["Dia"]),
19. errors="coerce"
20. )
21. **Limpeza e ordenação:**
    * remover linhas com data ou chuva\_mm nulos;
    * ordenar por data e resetar o índice.

Resultado: df\_long com colunas data (datetime) e chuva\_mm (float), com aproximadamente **2131 dias**.

**4.4 Dataset do modelo de 3 classes (risco por chuva)**

Implementado na função montar\_dataset\_ml(df\_long):

1. **Classificação da chuva em 3 classes de risco:**
2. def classificar\_risco\_chuva(mm: float) -> int:
3. if mm < 20.0:
4. return 0
5. elif mm < 60.0:
6. return 1
7. else:
8. return 2

Gera a coluna classe\_risco ∈ {0,1,2}.

1. **Índice temporal:**
2. df = df\_long.set\_index("data").sort\_index()
3. **Lags de chuva:**
4. df["chuva\_lag1"] = df["chuva\_mm"].shift(1)
5. df["chuva\_lag2"] = df["chuva\_mm"].shift(2)
6. df["chuva\_lag3"] = df["chuva\_mm"].shift(3)
7. **Variáveis de calendário:**
8. df["mes"] = df.index.month
9. df["dia\_semana"] = df.index.dayofweek # 0 = segunda
10. **Remoção de linhas com NaN nos lags:**

As primeiras três linhas (sem lags completos) são removidas.

Resultado final: df\_ml com colunas:

* chuva\_mm, chuva\_lag1, chuva\_lag2, chuva\_lag3,
* mes, dia\_semana,
* classe\_risco.

Tamanho aproximado: **2128 linhas x 12 colunas**.

**4.5 Construção das labels com ocorrência de alagamento (GeoSampa + CGE)**

Implementado em src/data/build\_labels\_geosampa.py:

1. **Leitura da chuva:**
   * utiliza a mesma função carregar\_chuva\_cge, produzindo série diária 2020–2025.
2. **Consulta WFS ao GeoSampa:**
   * Requisição:
   * service=WFS&version=1.0.0&request=GetFeature
   * &typeName=geoportal:risco\_ocorrencia\_alagamento
   * &outputFormat=application/json
   * Carregamento em GeoDataFrame com geopandas.read\_file.
   * Foram lidos cerca de **255 registros** de ocorrências para o município.
3. **Filtro para a subprefeitura do Butantã:**
   * inspeção das colunas mostra nm\_subprefeitura;
   * filtro:
   * gdf\_butanta = gdf[gdf["nm\_subprefeitura"]
   * .str.contains("BUTANTÃ", case=False, na=False)]
   * resultado: **13 registros** associados ao Butantã.
4. **Agregação diária:**
   * converter dt\_ocorrencia para datetime.date;
   * agrupar:
   * eventos\_dia = gdf\_butanta.groupby("data").size().reset\_index(name="n\_eventos")
   * eventos\_dia["flag\_evento"] = (eventos\_dia["n\_eventos"] > 0).astype(int)
5. **Junção com a série de chuva:**
   * df\_labels = df\_chuva.merge(eventos\_dia, on="data", how="left");
   * preencher NaNs em n\_eventos com 0, e flag\_evento com 0.

Resultado: labels\_chuva\_alagamento\_butanta.csv com colunas:

* data, chuva\_mm, n\_eventos, flag\_evento
* 2131 linhas (um dia por linha).

**4.6 Filtragem para dias chuvosos e features do modelo binário**

No script src/training/treino\_modelo\_evento\_butanta.py:

1. **Filtragem por limiar mínimo de chuva:**
   * objetivo: focar em dias com chuva minimamente relevante;
   * adotou-se RAIN\_MIN = 5.0 mm:
2. df\_sub = df\_labels[df\_labels["chuva\_mm"] >= 5.0].copy()
   * resultado: **397 dias** (chuva ≥ 5 mm), dos quais 7 com evento (flag\_evento = 1) e 390 sem.
3. **Engenharia de features:**

Sobre df\_sub foram calculadas:

* + chuva\_lag1, chuva\_lag2, chuva\_lag3;
  + acum\_3d, acum\_7d, acum\_30d (janelas móveis da chuva);
  + API usando ;
  + ano, mes, dia, dia\_semana.

Após remoção das primeiras linhas por conta dos lags, o conjunto final ficou com **394 dias**, sendo 7 com evento e 387 sem.

1. **Divisão treino/teste:**
   * divisão 70/30, estratificada quando possível;
   * treino: 275 dias (5 com evento, 270 sem);
   * teste: 119 dias (2 com evento, 117 sem).

**4.7 Oversampling e treino dos modelos binários**

No conjunto de treino (275 exemplos):

* classe 1 (com evento): 5 dias;
* classe 0 (sem evento): 270 dias.

Aplicou-se **oversampling** simples da classe positiva até atingir cerca de 20% de exemplos positivos:

* antes: 5 positivos, 270 negativos;
* depois: 67 positivos, 270 negativos (337 exemplos no treino).

Foram treinados dois modelos:

1. **Regressão Logística (binária):**
   * treinada sem class\_weight, pois o oversampling já equilibrou parcialmente as classes.
2. **Random Forest (binária):**
   * n\_estimators = 300, random\_state = 42,
   * também sem class\_weight.

O conjunto de teste permaneceu com a distribuição real (**2 eventos, 117 não eventos**).

**5. Discussão**

**5.1 Por que dois modelos?**

Os dois modelos têm **propósitos diferentes**:

1. **Modelo de 3 classes (chuva ⇒ risco)**
   * Pergunta principal:

“Dado um acumulado de chuva diária, quão grave pode ser o risco de inundação?”

* + Depende apenas de chuva\_mm e lags, com definições:
    - 0 – Sem risco: chuva < 20 mm;
    - 1 – Transitável: 20–60 mm;
    - 2 – Intransitável: ≥ 60 mm.
  + É, na prática, uma **régua de severidade de chuva**, coerente com classificações de eventos intensos na literatura.
  + Não depende de dados de ocorrência de alagamento, o que é fundamental já que esses registros são escassos.

1. **Modelo binário (dias chuvosos ⇒ alagamento sim/não)**
   * Pergunta principal:

“Entre os dias chuvosos, quais efetivamente tiveram alagamento no Butantã?”

* + Usa **dados reais de ocorrências** (flag\_evento) do GeoSampa, associados à chuva do CGE.
  + Precisa lidar com um cenário em que os **eventos são raríssimos** (apenas 7–10 dias com alagamento no período).
  + Mesmo com oversampling, continua limitado, mas permite estudar padrões de chuva antecedente que antecedem dias com alagamento.

Essa divisão é consistente com a distinção entre:

* **perigo/hazard** (chuva intensa, condição hidrológica desfavorável);
* **evento de impacto** (alagamento registrado), que também depende de vulnerabilidade, exposição e manutenção.

**5.2 Dificuldade intrínseca de prever alagamentos**

Mesmo se ignorássemos limitações de dados, o problema é difícil por natureza:

* São Paulo tem observado aumento de eventos extremos (chuva ≥ 80–100 mm/dia), mas nem toda chuva extrema resulta em alagamento em todos os bairros.
* Muitos alagamentos se devem a **problemas de manutenção**: bueiros entupidos, galerias assoreadas, lixo acumulado.
* A infraestrutura de drenagem (piscinões, galerias, canais) tem desempenho variável; obras recentes podem aliviar ou concentrar problemas.

Logo, **chuva forte é condição necessária, mas não suficiente** para produzir um alagamento. Um modelo que veja apenas chuva estará sempre “enxergando” só metade da história.

O protótipo desenvolvido aqui assume justamente isso:

* o modelo de 3 classes é uma **primeira camada de alerta** (chuva ⇒ nível de risco potencial);
* o modelo binário tenta capturar **assinaturas de chuva** associadas a dias com eventos reais, mas já nas condições atuais enfrenta sinais fracos e poucos exemplos.

**5.3 Ganhos e limitações dos resultados**

**Modelo de 3 classes**

A matriz de confusão da Regressão Logística para o período de teste foi:

com acurácia ≈ 0,991 e métricas muito altas em todas as classes.

A Random Forest, por sua vez, apresentou matriz completamente diagonal:

ou seja, **acurácia de 100%** no período de teste.

Isso não significa que “resolvemos” o problema de inundação, e sim que:

* o modelo está **reaprendendo a mesma regra** que definimos (sempre que chuva\_mm < 20, classe 0; entre 20 e 60, classe 1; etc.);
* como chuva\_mm está nas features, a tarefa é praticamente aprender os limiares que nós já determinamos.

Na prática, o modelo de 3 classes é uma forma organizada de consolidar a lógica de risco por faixas de chuva, deixando espaço para **no futuro adicionar mais variáveis** (por exemplo, API, uso do solo) sem mudar o fluxo de uso.

**Modelo binário condicionado**

Com filtragem para dias chuvosos (chuva ≥ 5 mm) e oversampling, a última execução produziu:

* treino: 337 exemplos (67 com evento, 270 sem);
* teste: 119 exemplos (2 com evento, 117 sem).

A Regressão Logística obteve:

* Matriz de confusão:
* Para a classe de evento (1):
  + precisão ≈ 0,17;
  + recall ≈ 0,5 (acertou 1 de 2 eventos).
* Acurácia global ≈ 95%.

O Random Forest, mesmo após oversampling, se manteve muito conservador, prevendo quase sempre “sem evento” (classe 0), com acurácia alta mas recall ≈ 0 para a classe de interesse.

Em resumo:

* o modelo binário **não é confiável** como previsor operacional, mas já consegue, com oversampling e condicionamento a dias chuvosos, ao menos “tocar” alguns eventos reais, em vez de simplesmente ignorá-los.

**5.4 Dados adicionais que poderiam melhorar a acurácia**

Ao longo da pesquisa, foram identificadas bases abertas que podem ser integradas para melhorar os modelos:

* **Camadas de risco hidrológico** do GeoSampa: áreas de risco de inundação próximas a córregos.
* **Manchas de inundação modeladas** (tempos de retorno 5, 25, 100 anos, etc.) para trechos de cursos d’água.
* **Camadas de uso e ocupação do solo, impermeabilização e altimetria**, frequentemente usadas como features em mapeamentos de suscetibilidade.
* **Séries mais longas e completas** de ocorrências de alagamento da Defesa Civil, se disponíveis (anos anteriores e posteriores a 2025).

Do ponto de vista de modelagem, poderiam ser explorados:

* oversampling mais sofisticado (por exemplo, SMOTE);
* modelos de **gradient boosting** (XGBoost, LightGBM) ou **redes neurais recorrentes (LSTM/GRU)** para lidar melhor com dependência temporal;
* calibração de limiares de decisão (probabilidade de corte) para privilegiar **recall** da classe de evento, aceitando mais falsos positivos, o que é adequado em sistemas de proteção à vida.

**5.5 Aspectos didáticos e narrativos do projeto**

Além dos números, o projeto tem valor como **estudo de caso** em “Sistemas de Controle e IA”:

* mostra como **definir um recorte viável** (um bairro, em vez da cidade toda);
* explicita o “choque de realidade” entre a **ideia ideal** (prever alagamento com alta precisão) e a **disponibilidade real de dados** (7–10 eventos em 5 anos);
* guia o estudante por todo o **pipeline de um projeto de IA aplicado**:
  + obtenção de dados de diferentes fontes (CGE, GeoSampa);
  + pré-processamento e limpeza;
  + engenharia de features (lags, acumulados, API);
  + escolha de modelos;
  + avaliação com métricas adequadas;
  + salvamento do modelo (joblib) e criação de um frontend simples (Streamlit).
* reforça a importância de **explicar as decisões** (por exemplo, por que usar 20 e 60 mm como limiares) e conectá-las à literatura e a fontes oficiais.

**6. Encerramento**

Este relatório apresentou o desenvolvimento de um **MVP de IA para predição de risco de inundação no bairro do Butantã**, construído a partir de duas abordagens complementares:

1. Um **classificador de 3 classes de risco**, baseado em intensidade de chuva diária e lags, com limiares em 20 e 60 mm/24h, fundamentados em classificações de chuva forte e estudos de precipitação extrema no Brasil.
2. Um **modelo binário condicionado a dias chuvosos (chuva ≥ 5 mm)**, que utiliza dados reais de ocorrências de alagamento do GeoSampa e engenharia de features (lags, acumulados, API, calendário), aliado a oversampling da classe positiva para enfrentar o desbalanceamento extremo.

Do ponto de vista quantitativo:

* o modelo de 3 classes atinge desempenho quase perfeito – o que é esperado, pois ele está basicamente reproduzindo uma regra que nós mesmos definimos;
* o modelo binário ainda é frágil, pois conta com pouquíssimos exemplos positivos, mas já ilustra a importância de oversampling e de restrição a dias chuvosos.

Do ponto de vista conceitual, o trabalho reforça três mensagens importantes:

1. **Sem dados, não há milagre de IA:** algoritmos não substituem séries históricas longas, dados de drenagem, manutenção e uso do solo.
2. **Prever alagamentos é intrinsecamente difícil**, pois o fenômeno depende de um conjunto complexo de fatores físicos, operacionais e sociais.
3. Protótipos bem documentados, como este, são essenciais para **organizar o problema**, testar hipóteses (<i>e se</i> chover X hoje e Y ontem?) e criar uma base sobre a qual projetos futuros mais robustos possam ser construídos.

Como próximos passos, recomenda-se:

* ampliar o período de análise, incorporando mais anos de chuva e ocorrências;
* integrar camadas de risco hidrológico, manchas de inundação e uso do solo;
* experimentar modelos mais sofisticados e calibração de limiares focada em recall;
* conectar o protótipo a **APIs de previsão de chuva** (nowcasting) para evoluir de um sistema retrospectivo para um sistema de alerta quase em tempo real.

**Referências (principais, em estilo ABNT simplificado)**

Obs.: aqui estão as principais. Você pode formatar direitinho na ABNT no Word (itálico em títulos, etc.).

**ANDRADE, Jackeline Soares.** Eventos extremos de precipitação pluvial na cidade de Itacoatiara-AM. 2023. Dissertação (Mestrado em Geografia) – Universidade Federal do Amazonas, Manaus, 2023. Disponível em: <https://tede.ufam.edu.br/bitstream/tede/9467/5/DISS\_JackelineAndrade\_PPGEOG.pdf>. Acesso em: 23 nov. 2025.

**DUARTE, Cristiana Coutinho.** Eventos extremos de chuva e análise da suscetibilidade a movimentos de massa e inundações na Região Metropolitana do Recife. 2016. Tese (Doutorado em Geografia) – Universidade Federal de Pernambuco, Recife, 2016. Disponível em: <https://repositorio.ufpe.br/bitstream/123456789/16971/1/TESE\_VERS%C3%83O%20FINAL\_CRISTIANA%20COUTINHO%20DUARTE\_GEOGRAFIA\_BIBLIOTECA.%2025%20jan.%202016.pdf>. Acesso em: 23 nov. 2025.

**GONÇALVES, B. H.** Estudo de episódios extremos de chuva e impactos urbanos. Referido em: DUARTE, C. C. (2016).

**MARTINS, Alessandro Marques.** Determinação de limiares de precipitações extremas na bacia do rio Sapucaí. Revista Brasileira de Geomorfologia, v. 19, n. 3, p. 547–561, 2018. Disponível em: <https://www.abge.org.br/downloads/rev-DETERMINACAO%20DE%20LIMIARES.pdf>. Acesso em: 23 nov. 2025.

**NOBOA, Caique Salvador.** Engenharia de features para construção de um modelo de previsão de eventos hidrometeorológicos em Curitiba. 2023. Trabalho de Conclusão de Curso – Universidade Tecnológica Federal do Paraná, Curitiba, 2023. Disponível em: <https://riut.utfpr.edu.br/jspui/bitstream/1/35538/1/previsao.pdf>. Acesso em: 23 nov. 2025.

**SILVA, Luiz Felipe da.** Aplicação de algoritmos de aprendizado de máquina para o mapeamento de suscetibilidade a inundações da bacia hidrográfica do Rio Itajaí. 2025. Trabalho de Conclusão de Curso – Universidade Federal de Santa Catarina, Florianópolis, 2025. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/266382/TCC\_LuizFelipeDaSilva.pdf>. Acesso em: 23 nov. 2025.

**MAKITA, F. H.** Mapeamento de suscetibilidade a inundação usando modelo Random Forest. São José dos Campos: INPE, 2023. Disponível em: <https://marte2.sid.inpe.br/col/sid.inpe.br/marte2/2023/04.25.10.50/doc/156263\_compressed.pdf>. Acesso em: 23 nov. 2025.

**PENNA, Ana Caroline.** Extremos de chuva na cidade de São Paulo. 2020. Dissertação – INPE, São José dos Campos, 2020. Disponível em: <https://mtc-m21c.sid.inpe.br/col/sid.inpe.br/mtc-m21c/2020/07.08.18.51/doc/Ana%20Caroline%20Penna.pdf>. Acesso em: 23 nov. 2025.

**PORTAL MULTIPLIX.** Meteorologista do Inmet explica como classificar a intensidade das chuvas. 2021. Disponível em: <https://www.portalmultiplix.com/noticias/cotidiano/meteorologista-do-inmet-explica-como-classificar-a-intensidade-das-chuvas>. Acesso em: 23 nov. 2025.

**PREFEITURA DO MUNICÍPIO DE SÃO PAULO.** Centro de Gerenciamento de Emergências Climáticas – CGE-SP. Disponível em: <https://www.cgesp.org/>. Acesso em: 23 nov. 2025.

**PREFEITURA DO MUNICÍPIO DE SÃO PAULO.** GeoSampa – Sistema de informações geográficas. Metadados da camada de alagamentos. Disponível em: <https://metadados.geosampa.prefeitura.sp.gov.br/geonetwork>. Acesso em: 23 nov. 2025.

**PREFEITURA DO RIO DE JANEIRO; FUNDAÇÃO CEPERJ.** Inundação: dimensionamento, alternativas tecnológicas, políticas públicas e estudos afins na prevenção e controle. Rio de Janeiro: CEPERJ, 2014.

**TERRA / BBC News Brasil.** Por que chuvas em São Paulo têm provocado tantos estragos. 13 mar. 2025. Disponível em: <https://www.terra.com.br/noticias/brasil/por-que-chuvas-em-sao-paulo-tem-provocado-tantos-estragos>. Acesso em: 23 nov. 2025.