Proposta de solução de Inteligência Artificial para detecção de fraudes nos procedimentos da Atenção Especializada à Saúde.

## Introdução

A proposta de desenvolvimento de uma solução para detecção de fraudes no SIA e no SIH, com foco nos procedimentos da Atenção Especializada à Saúde divide-se em 5 etapas:

1. Estudo de viabilidade para implantação dos 3 modelos: dois de ML (supervisionado e não supervisionado) e um de Deep Learning com Rede Neural de tipo Autoencoder Multilayer Perceptron (MLP), incluindo levantamento de requisitos e infraestrutura necessária e disponível.
2. Desenvolvimento de uma solução na versão Alpha com Inteligência Artificial de aprendizado de máquina supervisionado ou não supervisionado, a depender da acurácia e precisão dos modelos. Esta primeira versão pode ser desenvolvida tomando como base os mesmos critérios e método utilizado pela auditoria que identificou casos de fraude no estado do Maranhão.
3. Desenvolvimento de uma versão beta (1.0) com uma solução mais robusta de três modelos: (1.1) aprendizado de máquina não supervisionado; (1.2) aprendizado de máquina supervisionado; e (1.3) redes neurais (versão alpha de redes neurais)
4. Teste dos modelos em modo de desenvolvimento, desenho da solução (MLOps)
5. Implementação das 3 soluções em modo de produção em paralelo para avaliação de desempenho durante um ano, culminando com a avaliação do melhor modelo a ser melhorado para a versão 2.0 e implementado definitivamente.

## Propostas de nomes para a Inteligência Artificial de detecção de fraudes

*AGATHA - (Auditoria de Gerenciamento Autônomo de Transações e Históricos de Atividades)*

*AGNES - (Auditoria e Gerenciamento de Notificações e Estatísticas)*

*ALENA - (Análise de Linhas e Estatísticas em Auditoria Neural)*

*ALICIA - (Análise de Linhas e Identificação de Casos Irregulares Automatizada)*

*AMANDA - (Análise Multidimensional e Autônoma de Dados e Atividades)*

*ANA - (Análise Neural de Auditoria)*

*ARIA - (Auditoria Robótica de Informações e Atividades)*

*ARIEL - (Auditoria de Registros e Identificação de Exceções em Linhas)*

*ATLAS - (Análise de Transações e Laudos de Auditoria por Sistemas)*

*AURORA - (Auditoria Unificada de Registros e Operações Robóticas Autônomas)*

*IRIS - (Inteligência de Rastreamento e Identificação de Suspeitas)*

*ÍRIS - (Inteligência Robótica de Identificação de Suspeitas)*

*ISIS - (Inteligência de Supervisão e Identificação de Suspeitas)*

*LIA - (Inteligência Artificial de Auditoria)*

*RITA - (Robô de Inteligência em Auditoria)*

*RUBI - (Robô de Auditoria em Base de Informações)*

*SAFIRA - (Sistema de Análise e Fiscalização Inteligente de Registros e Atos)*

*SARAH - (Sistema Automatizado de Rastreamento e Auditoria em Histórico)*

## Base de dados proposta com variáveis para treinamento com dados do SIA, SIH, IBGE e SIOPS

|  |  |
| --- | --- |
| **Coluna** | **Descrição** |
| 1. CNES | Código do estabelecimento |
| 1. Mes\_Ano | Data que agrega os valores de quantidade e valor dos procedimentos por estabelecimento |
| 1. Estabelecimento | Nome fantasia do estabelecimento |
| 1. Razao\_Social | Razão social do estabelecimento |
| 1. UF | Unidade Federativa onde o estabelecimento está localizado |
| 1. Codigo\_Municipio | Código do município onde o estabelecimento está localizado |
| 1. Municipio | Nome do município onde o estabelecimento está localizado |
| 1. Codigo\_Procedimento | Código do procedimento realizado |
| 1. Procedimento | Descrição do procedimento realizado |
| 1. Valor\_Procedimentos | Valor total dos procedimentos realizados pelo estabelecimento por tipo de procedimento e por mês/ano |
| 1. Quantidade\_Procedimentos | Quantidade total de procedimentos realizados pelo estabelecimento por tipo de procedimento e por mês/ano |
| 1. Populacao | População do município onde o estabelecimento está localizado |
| 1. Procedimentos\_por\_10k\_habit | Quantidade total de procedimentos realizados pelo estabelecimento por tipo de procedimento e por 10.000 habitantes |
| 1. Valor\_Total\_10k\_habit | Valor total dos procedimentos realizados pelo estabelecimento por tipo de procedimento e por 10.000 habitantes |
| 1. Distancia\_Qtde\_Proced\_Med | Distância da quantidade total de procedimentos realizados pelo estabelecimento da mediana dos 120 meses |
| 1. Distancia\_Valor\_Proced\_Med | Distância dos valores totais dos procedimentos realizados pelo estabelecimento da mediana dos 120 meses |
| 1. Tipo\_Estabelecimento | Classificação do tipo de estabelecimento, como hospitais, clínicas, laboratórios, etc. |
| 1. Tamanho\_Municipio | Classificação do tamanho do município, como pequeno, médio e grande |
| 1. Emendas\_Parlamentares | Indica se o estabelecimento de saúde recebeu emendas parlamentares nos últimos anos |
| 1. IDH | Índice de Desenvolvimento Humano do município onde o estabelecimento está localizado |
| 1. Taxa\_Desemprego | Taxa de desemprego do município onde o estabelecimento está localizado |
| 1. Renda\_Per\_Capita | Renda per capita do município onde o estabelecimento está localizado |
| 1. Volume\_Atendimentos | Volume de atendimentos por estabelecimento de saúde |
| 1. Num\_Fiscalizacoes | Número de fiscalizações realizadas no estabelecimento de saúde |
| 1. Num\_Sancoes | Número de sanções aplicadas no estabelecimento de saúde |
| 1. Tipo\_Infração | Tipo de infração identificada durante a fiscalização |
| 1. Local\_Fiscalizacao | Local onde foi realizada a fiscalização |
| 1. Data\_Fiscalizacao | Data em que foi realizada a fiscalização |

Além das variáveis que podem ser obtidas no SIA e SIH, proponho a inclusão de variáveis complementares que podem ser obtidas pelo SIOP e outras fontes de dados públicas como ANVISA, IBGE, ATLAS do desenvolvimento humano, entre outras.

Com base nas variáveis disponíveis (variáveis 1 a 17) e em outros fatores que podem ser convertidos em variáveis a partir de bases de dados públicas do governo federal brasileiro, sugiro as seguintes variáveis adicionais que podem ser úteis para aumentar a eficiência do modelo de detecção de fraudes:

1. **Tipo de estabelecimento:** Adicionar uma variável que classifique os estabelecimentos de saúde de acordo com o tipo de serviço que prestam, como hospitais (por porte), clínicas, laboratórios, etc. Essa variável pode ajudar a identificar padrões de fraude específicos em cada tipo de estabelecimento.
2. **Tamanho do município:** Adicionar uma variável que classifique os municípios de acordo com sua população, como pequenos, médios e grandes. Essa variável pode ajudar a identificar padrões de fraude específicos em cada tamanho de município.
3. **Emendas parlamentares:** Adicionar uma variável que indique se o estabelecimento de saúde recebeu emendas parlamentares nos últimos anos. Essa variável pode ajudar a identificar padrões de fraude relacionados a possíveis favorecimentos políticos.
4. **Dados socioeconômicos:** Adicionar variáveis socioeconômicas, como índice de desenvolvimento humano (IDH), taxa de desemprego, renda per capita, etc. Essas variáveis podem ajudar a identificar padrões de fraude relacionados a áreas com maior vulnerabilidade socioeconômica.
5. **Volume de atendimentos:** Adicionar uma variável que indique o volume de atendimentos por estabelecimento de saúde. Essa variável pode ajudar a identificar padrões de fraude relacionados a estabelecimentos que realizam um número anormalmente alto de procedimentos.
6. **Dados de fiscalização:** Adicionar variáveis que indiquem o número de inspeções ou fiscalizações realizadas em cada estabelecimento de saúde. Essas variáveis podem ajudar a identificar padrões de fraude relacionados a estabelecimentos que recebem pouca ou nenhuma fiscalização.

Essas variáveis adicionais podem ser obtidas a partir de bases de dados públicas do governo federal e incorporadas ao modelo de autoencoder MLP para melhorar sua eficiência na detecção de fraudes em dados de saúde. No entanto, é importante lembrar que o modelo deve ser treinado com cuidado para evitar o overfitting e garantir a validade e a confiabilidade dos resultados.

Os dados de fiscalização podem ser obtidos em tempo real a partir do Portal da Transparência do Governo Federal, especificamente na seção de dados abertos do Sistema Integrado de Planejamento e Orçamento (SIOP).

O SIOP é um sistema que permite a gestão integrada do planejamento, orçamento e financeiro do Governo Federal. Em sua seção de dados abertos, é possível encontrar informações sobre a execução orçamentária dos órgãos públicos, incluindo dados de fiscalização.

Entre as informações disponíveis, destacam-se:

* Número de fiscalizações realizadas;
* Número de sanções aplicadas;
* Tipo de infração;
* Local da fiscalização;
* Data da fiscalização;
* Valor das multas aplicadas.

Os dados estão disponíveis em formato aberto e podem ser acessados por meio de APIs ou arquivos em formato CSV.

Além disso, é importante destacar que alguns órgãos públicos, como a Agência Nacional de Saúde Suplementar (ANS) e a Agência Nacional de Vigilância Sanitária (ANVISA), possuem seus próprios sistemas de fiscalização, nos quais é possível obter dados sobre a fiscalização em tempo real.

# Estratégia de implementação de pipelines de ETL e da aplicação de ML (versões alpha e beta 1.0) e NN (versão 2.0)

Para implementar uma solução completa em produção que possa processar milhões de linhas, é importante considerar os seguintes aspectos:

1. **Divisão em etapas e modularidade:** Dividir o código em etapas claras e modulares para que cada etapa possa ser executada independentemente em uma pipeline separada. Isso permite que cada etapa seja dimensionada separadamente e que o fluxo de dados possa ser controlado de forma eficiente entre as diferentes etapas.
2. **Orquestração de pipelines:** Usar uma ferramenta de orquestração de pipelines, como Apache Airflow ou Apache NiFi, para gerenciar e agendar a execução das diferentes etapas do pipeline. Isso permite que as etapas sejam executadas automaticamente em uma programação regular e que o fluxo de dados entre as diferentes etapas seja gerenciado automaticamente.
3. **Armazenamento de dados:** Armazenar os dados de entrada e saída em um local centralizado, como um banco de dados ou um sistema de arquivos distribuído, como o Hadoop Distributed File System (HDFS). Isso permite que os dados sejam acessados e processados de forma eficiente pelas diferentes etapas do pipeline e garante que os dados possam ser recuperados em caso de falha do sistema.
4. **Escalonamento de recursos:** Usar uma plataforma de gerenciamento de recursos, como Kubernetes ou Apache Mesos, para dimensionar automaticamente os recursos do sistema, como CPU, memória e armazenamento, para atender à carga de trabalho de processamento de dados em cada etapa do pipeline. Isso garante que o sistema possa lidar com um grande volume de dados sem sobrecarregar os recursos do sistema.
5. **Monitoramento e alertas:** Implementar um sistema de monitoramento e alertas para alertar os administradores do sistema sobre possíveis problemas ou falhas no pipeline. Isso permite que os problemas sejam identificados e resolvidos rapidamente antes que possam afetar a integridade ou a disponibilidade do sistema.

Com essas práticas recomendadas em mente, um possível esquema de pipeline para a solução proposta seria:

1. **Extração de dados:** uma pipeline separada para extrair dados dos sistemas de informações hospitalares e ambulatoriais do SUS e armazená-los em um banco de dados ou sistema de arquivos distribuído, como o HDFS.
2. **Pré-processamento de dados:** uma pipeline separada para pré-processar os dados, incluindo a transformação do formato de longo para largo, a adição de colunas adicionais e a normalização/padronização dos dados. Os dados pré-processados devem ser armazenados em um local centralizado, como um banco de dados ou sistema de arquivos distribuído.
3. **Treinamento do modelo:** uma pipeline separada para treinar o modelo de autoencoder MLP usando os dados pré-processados. O modelo treinado deve ser salvo em um local centralizado, como um banco de dados ou sistema de arquivos distribuído.
4. **Avaliação do modelo:** uma pipeline separada para avaliar o desempenho do modelo treinado usando um conjunto de dados de teste. Os resultados da avaliação devem ser armazenados em um local centralizado, como um banco de dados ou sistema de arquivos distribuído.
5. **Inferência em tempo real:** uma API RESTful para fornecer inferência em tempo real para os sistemas de saúde que desejam detectar fraudes em seus dados. A API deve usar o modelo de autoencoder MLP treinado para calcular a distância euclidiana entre os dados de entrada e as saídas do modelo. Se a distância for maior que um determinado limiar, um alerta de fraude deve ser gerado e registrado em um banco de dados.
6. **Monitoramento e alertas:** Implementar um sistema de monitoramento e alertas para alertar os administradores do sistema sobre possíveis problemas ou falhas nas diferentes etapas do pipeline. Isso pode ser feito usando ferramentas de monitoramento de sistema, como Prometheus, Grafana ou Zabbix.
7. **Escalonamento de recursos:** plataforma de gerenciamento de recursos, como Kubernetes ou Apache Mesos, para dimensionar automaticamente os recursos do sistema para atender à carga de trabalho de processamento de dados em cada etapa do pipeline. Isso garante que o sistema possa lidar com um grande volume de dados sem sobrecarregar os recursos do sistema.

Com essa arquitetura em mente, é possível implementar uma solução robusta para detecção de fraudes em grandes volumes de dados de saúde usando um modelo de autoencoder MLP.

## Sistema de monitoramento de fraudes e de desempenho do modelo

Para um painel de monitoramento da detecção de fraudes com o modelo proposto, sugiro os seguintes indicadores:

1. **Número de alertas de fraude:** O número total de alertas de fraude gerados pelo sistema em um determinado período de tempo.
2. **Percentual de alertas verdadeiros positivos**: A porcentagem de alertas de fraude que foram confirmados como fraudes reais em relação ao número total de alertas gerados no mesmo período de tempo.
3. **Percentual de alertas falsos positivos:** A porcentagem de alertas de fraude que foram identificados como falsos positivos em relação ao número total de alertas gerados no mesmo período de tempo.
4. **Tempo médio de resposta:** O tempo médio necessário para que um alerta de fraude seja identificado e registrado pelo sistema após a ocorrência da fraude.
5. **Distribuição geográfica das fraudes:** Um mapa que mostra a distribuição geográfica das fraudes detectadas pelo sistema, permitindo que as autoridades de saúde concentrem seus esforços de investigação onde as fraudes são mais frequentes.
6. **Histórico de alertas:** Uma lista dos alertas de fraude gerados pelo sistema nos últimos dias, incluindo informações como o local, data e hora da fraude, a magnitude da fraude e as ações tomadas pela equipe de saúde.
7. **Desempenho do modelo:** Uma seção que exibe o desempenho atual do modelo de autoencoder MLP, incluindo métricas como a acurácia, precisão e recall.

Além disso, o painel deve ser atualizado em tempo real, para que as autoridades de saúde possam responder rapidamente às fraudes detectadas pelo sistema.

A implementação do painel de monitoramento pode ser feita usando ferramentas de visualização de dados, como Tableau, Power BI ou Grafana, conectando-se ao banco de dados onde os alertas de fraude e as métricas de desempenho do modelo estão armazenados. O painel pode ser configurado para enviar alertas por email ou SMS quando o número de alertas de fraude ultrapassar um determinado limite ou quando houver um aumento no número de falsos positivos.

## Proposta de dashboard para monitoramento das possíveis fraudes detectadas e de desempenho do modelo

Para o monitoramento em tempo real das detecções de fraudes, seria interessante a criação de um dashboard que permita visualizar as informações de forma clara e intuitiva. Algumas possíveis visualizações e métricas que poderiam ser incluídas no dashboard são:

* Número de detecções de fraudes por mês
* Distribuição das detecções de fraudes por tipo de estabelecimento de saúde
* Distribuição das detecções de fraudes por tamanho do município
* Distribuição das detecções de fraudes por tipo de infração identificada
* Mapa geográfico com a localização dos estabelecimentos de saúde e das detecções de fraudes
* Gráfico de linha com a evolução do número de detecções de fraudes ao longo do tempo
* Gráfico de barras com o número de detecções de fraudes por UF
* Gráfico de pizza com a distribuição das detecções de fraudes por tipo de procedimento realizado
* Gráfico de dispersão com a relação entre o volume de atendimentos e o número de detecções de fraudes
* Tabela com os detalhes das detecções de fraudes, incluindo a data, o estabelecimento de saúde, o tipo de infração identificada e outras informações relevantes.

O dashboard poderia ser desenvolvido em uma plataforma de visualização de dados, como Qlik Sense, e ser atualizado automaticamente com os dados extraídos em tempo real da fonte de fiscalização. Também seria importante definir responsabilidades e protocolos claros para a tomada de ações em caso de detecções de fraudes, a fim de garantir a efetividade da solução.

O painel poderia ser dividido em diferentes seções, como:

1. **Dashboard**: Uma visão geral do painel, que inclui as principais métricas do modelo, como o número de alertas de fraude, o percentual de alertas verdadeiros positivos e falsos positivos, e o tempo médio de resposta.
2. **Mapa de calor:** Um mapa que mostra a distribuição geográfica das fraudes detectadas pelo sistema, com áreas em cores diferentes indicando a magnitude da fraude.
3. **Histórico de alertas:** Uma lista de alertas de fraude gerados pelo sistema nos últimos dias, que inclui informações como a data, hora e local da fraude, a magnitude da fraude, e as ações tomadas pela equipe de saúde.
4. **Desempenho do modelo:** Uma seção que exibe o desempenho atual do modelo de autoencoder MLP, incluindo métricas como a acurácia, precisão e recall.
5. **Alertas:** Uma seção que exibe alertas de fraude em tempo real, com uma descrição da fraude, o local e a magnitude, e sugestões de ações a serem tomadas pela equipe de saúde.
6. **Gráficos e tabelas:** Gráficos e tabelas adicionais que fornecem informações detalhadas sobre a distribuição de fraudes por região, tipo de estabelecimento e procedimento, e outras métricas relevantes.

O painel deve ser projetado de forma intuitiva e fácil de usar, com gráficos e tabelas claras e simples que permitem que as autoridades de saúde identifiquem rapidamente padrões e tendências nos dados. O layout do painel deve ser flexível e ajustável para que os usuários possam personalizar as informações exibidas de acordo com suas necessidades específicas.