Documentação de análise de impacto ético

Em um mundo cada vez mais impulsionado por dados e tomada de decisão com base nos mesmos, a ética tornou-se um pilar fundamental para garantir que essas inovações promovam benefícios sociais, respeitem os direitos individuais e evitem consequências prejudiciais. A crescente intersecção entre tecnologia e dados requer uma abordagem cuidadosa e responsável, destacando a importância da análise de impacto ético. Esta documentação tem como objetivo fornecer um guia abrangente e sistemático para avaliar os impactos éticos associados a projetos, processos ou tecnologias que envolvem a manipulação e interpretação de dados. Ao abordar princípios fundamentais e estratégias práticas, esta documentação visa, também, a exploração de requisitos importantes para toda curadoria e utilização coerente das fontes de dados, com uma visão ética, assegurando que inovações tecnológicas sejam guiadas por princípios éticos e promovam um impacto positivo em todas as esferas da sociedade.

1. Privacidade e proteção de dados

Sobre os dados coletados, vale ressaltar que uma parte deles é de natureza pública, enquanto outra parcela é fornecida diretamente pelo cliente. No caso desta última, a coleta é realizada de forma segura por meio de uma API dedicada, assegurando o tráfego protegido dessas informações até a fase de armazenamento na nuvem. Para tal finalidade, adotamos uma abordagem robusta que envolve a utilização do Data Lake da AWS, mais especificamente o Amazon S3, em conjunto com um Data Warehouse, o Amazon Redshift. É relevante salientar que, no caso específico do Amazon S3, empregamos a criptografia padrão oferecida pela plataforma, assegurando assim a proteção dos dados armazenados desde o momento da coleta até a persistência na nuvem. Essa arquitetura meticulosamente planejada não apenas

facilita a eficácia operacional, mas também assegura a integridade e confidencialidade dos dados em cada etapa do processo, garantindo, assim, a segurança abrangente do sistema.

É essencial garantir que as organizações responsáveis pelos dados utilizados no projeto trabalham com informações que não são diretamente ligadas a indivíduos específicos. Isso inclui dados sobre a população, sobre o ambiente econômico e sobre o poder aquisitivo de uma região. É preciso validar se essas informações são coletadas e utilizadas para fins de pesquisa e análise, não para identificar ou rastrear indivíduos específicos.

Para isso, é necessário que se realize uma análise de cada fonte utilizada. Essas análises serão feitas considerando pontos como: se as informações podem ser consideradas sensíveis, ou seja, se são informações como dados de saúde, informações financeiras pessoais ou detalhes sobre comportamento pessoal; se foi garantido que a fonte está de acordo com o LGPD, o que pode ser feito checando por exemplo se os titulares dos dados têm ou não o direito de solicitar acesso aos mesmos; entre outras.

A ANVISA possibilita a análise de poder aquisitivo por região, uma vez que exerce a vigilância sanitária sobre produtos e serviços. Esta fonte apresenta dados de cunho econômico que não se caracterizam, portanto, como dados pessoais. O que garante a proteção dos dados é a chamada “Política de Proteção de Dados Pessoais” que seguem. A política visa, de acordo com o Ministério da Saúde, “garantir o cumprimento das normas relacionadas à privacidade, à transparência, ao acesso às informações públicas e à proteção das liberdades e dos direitos fundamentais dos indivíduos”.

O CNPJ armazena informações cadastrais das pessoas jurídicas e de outras entidades. O CNPJ em si é um dado não pessoal, pois ele é um número único que identifica uma entidade. Porém, as informações que podem ser obtidas a partir dele podem não ser. Por exemplo, a situação cadastral do CNPJ pode indicar se a empresa está em conformidade com suas obrigações fiscais e tributárias.

Apesar disso, ele pode ser classificado como uma fonte de dados não pessoais pois contém informações sobre a situação cadastral da empresa, que são relacionadas ao negócio da empresa, não a indivíduos. Ele não contém informações que possam ser usadas para identificar uma pessoa individualmente e, por fim, não contém informações que possam ser consideradas sensíveis.

O BACEN permite acesso a informações sobre taxa selic, inflação e o valor da moeda em dólar. Estes dados são de cunho estritamente econômico, não representando, portanto, ameaça a privacidade de terceiros.

O POF e o IBGE têm acesso a informações relacionadas ao poder aquisitivo de cada região, como a composição dos orçamentos domésticos e as condições de vida da população brasileira. Os dados que o POF fornece não são utilizados para análise de indivíduos específicos, portanto pode ser considerada uma fonte de dados não pessoais, assim como aquelas utilizadas do IBGE.

2. Equidade e justiça

Ao trabalhar com grandes volumes de dados, é necessário considerar os possíveis impactos em grupos específicos e buscar formas de minimizar as disparidades. O uso ético de dados em arquiteturas de big data envolve garantir que todos os grupos se beneficiem do uso desses dados e que ninguém seja prejudicado ou desfavorecido pelo uso indevido ou enviesado desses dados.

Por exemplo, o uso de dados do IBGE pode revelar disparidades socioeconômicas entre diferentes grupos, que podem ser amplificadas se as análises geradas não forem feitas de forma justa e equitativa. Da mesma forma, os dados da ANVISA podem ter implicações significativas para a saúde pública e podem impactar desproporcionalmente grupos vulneráveis se não forem manuseados de forma ética (De Freitas Saldanha et al., 2021).

Para minimizar as disparidades no uso de Big Data, é vital garantir que o pré-processamento dos dados seja realizado de maneira justa e equitativa . Isso pode envolver a garantia de que todos os grupos estejam adequadamente representados nos dados, a consideração de vieses potenciais nos dados e a implementação de técnicas de análise robustas para minimizar a chance de resultados tendenciosos. Garantir também que dados que podem ser relevantes para alguma minoria, por exemplo, não sejam removidos.

A visualização dos dados, a partir do infográfico, também tem que ser feita pensando nos princípios de ética, uma vez que a distorção de dados, mesmo que seja feita de forma acidental, pode acabar apoiando alguma narrativa falsa e intensificando uma possível desinformação. Adotar uma abordagem ética implica em evitar manipulações que possam distorcer a compreensão dos dados, selecionar adequadamente os métodos de visualização para representar fielmente as informações e fornecer contexto apropriado para evitar interpretações equivocadas.

Além disso, é crucial considerar o impacto potencial nas comunidades marginalizadas ao lidar com dados sensíveis. O uso indevido dessas informações pode resultar em discriminação e reforçar desigualdades existentes. Para evitar isso, é imperativo implementar medidas rigorosas de segurança e privacidade, como a anonimização eficaz dos dados, para garantir que a identidade de grupos específicos não seja comprometida.

Em resumo, ao lidar com dados provenientes de fontes como BACEN, CNPJ, POF, IBGE, ANVISA, é fundamental adotar uma abordagem ética em todas as fases do processo, desde a coleta até a visualização, considerando cuidadosamente os impactos potenciais em grupos específicos e implementando medidas para minimizar disparidades e proteger a privacidade.

3. Transparência e consentimento informado

Primeiramente é importante ressaltar que o LGPD estabelece diretrizes sobre a coleta e o tratamento de dados pessoais, assegurando que o consentimento seja obtido. Todas as partes envolvidas, seja na pesquisa do IBGE ou no fornecimento de dados para o CNPJ, devem operar em conformidade com essas regulamentações.

O IBGE tem como princípio a transparência dos dados coletados. Antes de realizar qualquer pesquisa, o IBGE informa os objetivos das informações obtidas, garantindo que todos os participantes estejam cientes do propósito da coleta. No caso da POF, o IBGE adota medidas rigorosas para assegurar que os participantes estejam cientes da natureza da pesquisa, dos benefícios sociais associados e dos métodos de coleta de dados. Além disso, é assegurado que os dados individuais sejam tratados com sigilo e confidencialidade, sendo utilizados apenas para fins estatísticos.

Quanto ao CNPJ, a obtenção de consentimento é uma parte fundamental do processo. As empresas e organizações que fornecem informações para o CNPJ são informadas sobre a necessidade e o propósito da coleta desses dados. O acesso a essas informações é estritamente regulamentado e destinado a fins específicos, como fiscalização e transparência.

A API do cliente disponibiliza um CSV com todas as vendas realizadas por CNPJ, apresentando a data, o valor e a quantidade atrelados a cada um. A este CSV se aplicam, portanto, as mesmas especificações citadas sobre as pesquisas do CNPJ.

Por fim, os dados coletados da ANVISA e do BACEN correspondem de maneira geral a dados econômicos não pessoais, situação na qual a transparência e o consentimento informado são garantidos por conta da natureza destes dados.

4. Responsabilidade social

No cenário dinâmico do avanço do big data, a responsabilidade social torna-se um elemento central que orienta o impacto desta tecnologia na sociedade. À medida que os dados se tornam cada vez mais abundantes, a adoção de práticas éticas e socialmente responsáveis torna-se não só mais um requisito, mas também crítica para mitigar as desigualdades, promover a inclusão e garantir que o desenvolvimento tecnológico tenha um impacto que beneficie a todos.

Ao analisar o impacto social dos projetos, nos comprometemos com uma avaliação criteriosa, focando em como essas iniciativas repercutem nas comunidades e no meio ambiente. O compromisso do projeto com as comunidades e meio ambiente é o mais importante para o grupo, principalmente para seguir as normas de regras da ODS. Além disso, não só enfatizando os impactos positivos esperados, mas também implementando medidas preventivas para evitar quaisquer impactos negativos.

O foco principal é garantir que os nossos esforços contribuam de forma eficaz e positiva para questões globais prementes, como a redução da desigualdade, o combate às alterações climáticas e a promoção da inovação tecnológica responsável.

Esta abordagem reflete o nosso forte compromisso não só em impulsionar a inovação e a eficiência, mas também em promover o progresso social e ambiental de uma forma ética e sustentável. Trabalhamos para integrar práticas sustentáveis e éticas em todas as fases dos projetos, traçando um caminho que não só antecipa as necessidades atuais, mas também visa construir um futuro mais equitativo e sustentável.

Concluindo, com o avanço do big data, a responsabilidade social tornou-se crucial. Ao alinhar os projetos de dados com os ODS, priorizamos não apenas a inovação e a eficiência, mas também o impacto positivo nas questões globais. A nossa abordagem visa não só antecipar as necessidades atuais, mas também construir um futuro mais equitativo e

sustentável, integrando práticas éticas em todas as fases dos projetos. Desta forma, não só impulsionamos o progresso tecnológico, mas também promovemos o progresso social e ambiental ético e duradouro.

5. Viés e discriminação

O viés é um tópico importante quando tratamos sobre dados, especialmente quando considerando os mesmo na temática de ciência de dados e modelos de predição ou classificação. Neste sentido, o viés se baseia na distorção nos dados, ou tendência desvirtuada ou preconceituosa em relação ao conhecimento e insights extraídos após análise dos dados.

Esses vieses podem ser definidos em três principais categorias, o viés de amostragem, o viés humano e viés algorítmico (Luis B, 2023). Respectivamente, o primeiro se refere aos tipos de dados provenientes de amostras bem pequenas (pode ser até pelo nicho de pesquisa performado), que não representam a população no geral. O viés humano, é um dos vieses mais recorrentes quanto estamos no processo de tratamento, análise e interpretação dos dados. Esse viés se dá especialmente pela opinião, possíveis preconceitos e estereótipos que possam guiar a interpretação e inferência dos dados, assim como influenciar nas decisões e resultados finais. Por fim, o viés algorítmico se reflete ao processo de aprendizado de máquina, principalmente em modelos de aprendizado supervisionado, ser feito a partir de conjuntos de dados refletem desigualdades sociais, discriminação ou estereótipos existentes (isso acontece muito com dados antigos).

Sendo assim, é muito importante que tenhamos isso em mente, sempre que formos tratar com dados, especialmente, aqueles no qual não houve uma curadoria na forma em que foi feita a pesquisa, em base de dados estatísticos e dados públicos (dados abertos de livre

acesso). Principalmente, tendo em vista que esses vieses podem surgir de forma intencional ou não intencional.

Com base em nossas fontes de dados, podemos examinar as potenciais fontes de viés e estratégias para mitigá-las em cada frente:

1. **Dados da ANVISA:**
    - Possível Viés: Pode haver viés socioeconômico nas petições alimentícias, favorecendo áreas com maior poder aquisitivo.
    - Mitigação: Normalizar os dados em relação à população de cada região, considerar indicadores socioeconômicos locais e aplicar técnicas estatísticas para ajuste de viés.
2. **Dados do Bacen:**
    - Possível Viés: Dados financeiros podem refletir desigualdades econômicas e impactar a análise de diferentes regiões.
    - Mitigação: Normalizar os dados em relação à população, considerar fatores socioeconômicos, e adotar técnicas de correção para equilibrar possíveis distorções.
3. **Dados de CNPJ:**
    - Possível Viés: Pode haver desigualdades na representação de diferentes canais de venda, favorecendo alguns em detrimento de outros.
    - Mitigação: Analisar representatividade proporcional dos canais de venda, considerar dados adicionais sobre distribuição demográfica e aplicar técnicas de ajuste.
4. **Dados da POF e IBGE:**
    - Possível Viés: As amostras podem não ser totalmente representativas, resultando em visões distorcidas das condições de vida.
    - Mitigação: Validar a representatividade das amostras, corrigir distorções conhecidas e incorporar dados adicionais para enriquecer a análise.

5. **Dados coletados com a API do parceiro:**

- Possível Viés: Pode haver distorções nas dinâmicas de vendas fictícias que não reflitam a realidade do mercado.
- Mitigação: Validar a consistência dos dados, considerar fatores contextuais que possam influenciar as vendas fictícias e aplicar técnicas de ajuste.

Na estratégia geral de mitigação, é fundamental adotar medidas que assegurem a integridade e imparcialidade das análises de dados. A transparência é importante, exigindo que os métodos de coleta, processamento e análise sejam compreensíveis e acessíveis a todos os envolvidos. O segundo ponto principal é a revisão ética, realizada de forma regular, para identificar potenciais pontos de discriminação ou exclusão involuntária, garantindo que os resultados sejam éticos e equitativos. Além disso, é imperativo fomentar a diversidade na equipe responsável pela análise, tendo em vista que uma equipe com perspectivas variadas contribui para uma abordagem mais abrangente e sensível às diferentes nuances presentes nos dados. Consequentemente, a inclusão de diversas vozes na tomada de decisões promove uma análise mais completa e objetiva. A reavaliação contínua é outra peça-chave na estratégia de mitigação. Estabelecer protocolos que permitam revisar e ajustar as análises à medida que novos insights ou preocupações éticas surgem é fundamental. Esse processo dinâmico assegura que as análises permaneçam alinhadas com os padrões éticos mais recentes e com a evolução do contexto em que estão inseridas. Em conjunto, essas abordagens formam uma base sólida para a realização de análises de dados éticas e equitativas, mitigando potenciais riscos de viés e discriminação.

Em conclusão, a análise de dados enfrenta desafios significativos relacionados ao viés, especialmente nas áreas da ciência de dados e modelos de predição. O viés, caracterizado pela distorção nos dados e tendências preconceituosas, pode manifestar-se em diferentes

formas, como o viés de amostragem, o viés humano e o viés algorítmico. Cada uma dessas categorias apresenta riscos específicos que podem comprometer a precisão e a equidade das análises.

Ao examinar as fontes de dados específicas, identificamos possíveis fontes de viés em cada frente. Contudo, é crucial adotar uma abordagem proativa para mitigar esses vieses e garantir a integridade das análises. A estratégia geral de mitigação destaca a importância da transparência nos métodos de coleta e análise, da revisão ética regular para identificar possíveis discriminações, da promoção da diversidade na equipe para trazer perspectivas variadas e da reavaliação contínua para ajustar as análises conforme necessário.

Essas medidas, quando implementadas de forma coordenada, formam uma base robusta para a realização de análises éticas e equitativas, minimizando os riscos associados ao viés e à discriminação nos dados. Assim, ao enfrentar os desafios inerentes à análise de dados, é essencial manter um compromisso contínuo com a transparência, a ética e a diversidade para garantir resultados embasados e sem, ou com pouco viés.

6. Referências

Situação Cadastral CNPJ: entenda o que é e como funciona. (2023, December 7). Blog C6 Bank. https://www.c6bank.com.br/blog/situacao-cadastral-cnpj

Conheça a Política de Proteção de Dados Pessoais da Anvisa. (2023, October 23). *Agência Nacional De Vigilância Sanitária - Anvisa*. https://www.gov.br/anvisa/pt-br/assuntos/noticias-anvisa/2023/conheca-a-politica-de-protecao-de-dados-pessoais-da-anvisa

De Freitas Saldanha, R., Barcellos, C., & De Moraes Pedroso, M. (2021). Ciência de dados e big data: o que isso significa para estudos populacionais e da saúde? *Cadernos Saúde Coletiva*, *29*(spe), 51–58. https://doi.org/10.1590/1414-462x202199010305

Totvs, E. (2023, May 3). *Big Data: o que é, como funciona e como aplicar?* TOTVS. https://www.totvs.com/blog/inovacoes/big-data

B, L. (2023, July 24). *O viés na Ciência de Dados*. https://www.linkedin.com/pulse/o-vi%25C3%25A9s-na-ci%25C3%25AAncia-de-dados-luis-balero/?trackingId=