Relatório Técnico – PNAD COVID

1. Introdução

A pandemia de COVID-19 representou um dos maiores desafios sanitários, sociais e econômicos da história recente, impactando diretamente a vida da população brasileira em diferentes dimensões. Nesse contexto, compreender o comportamento da sociedade diante da crise, bem como os sintomas clínicos mais reportados e as condições socioeconômicas associadas, torna-se essencial para subsidiar políticas públicas e estratégias de resposta em eventuais novos surtos.

O presente trabalho foi desenvolvido no âmbito do Tech Challenge – Fase 3, com o objetivo de analisar os dados da PNAD-COVID19, pesquisa realizada pelo IBGE durante o período crítico da pandemia. Essa base de dados constitui uma fonte oficial, ampla e confiável para compreender a dinâmica da população brasileira frente à COVID-19, abarcando aspectos de saúde, comportamento e economia.

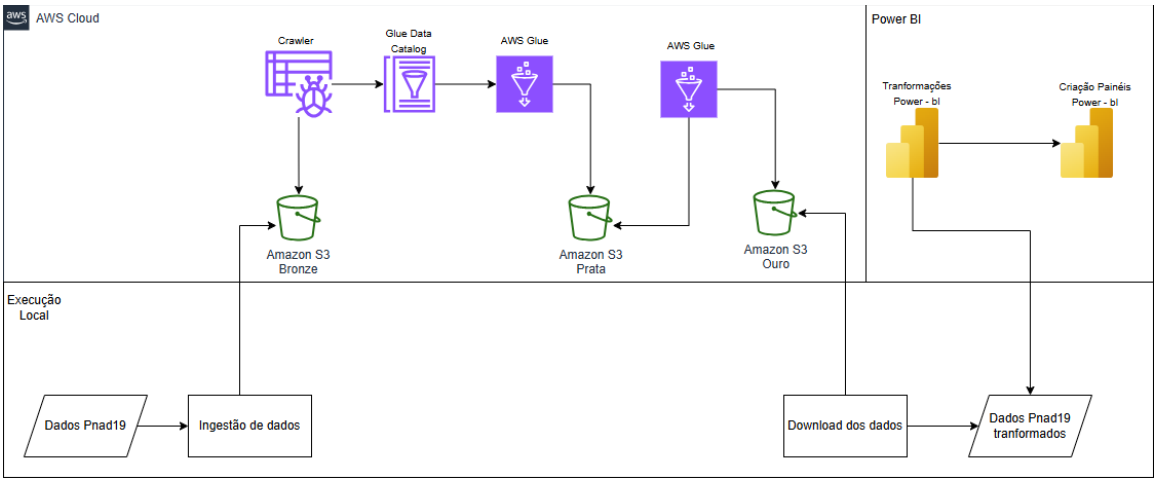
A proposta consiste em selecionar um conjunto reduzido de perguntas da pesquisa, limitado a 20 questões, que sejam capazes de fornecer uma visão abrangente sobre três dimensões centrais:

- **Características clínicas:** identificação e análise dos principais sintomas relatados;
- **Comportamento da população:** medidas de prevenção, restrição de mobilidade e acesso a serviços de saúde;
- **Características econômicas:** situação ocupacional, rendimentos e impactos financeiros no período.

A partir dessa organização, a análise busca não apenas descrever os achados, mas também gerar insights estratégicos para a tomada de decisão de um grande hospital parceiro. O objetivo é apoiar o planejamento de medidas preventivas e operacionais em caso de novos surtos de COVID-19, garantindo maior agilidade na resposta e efetividade na alocação de recursos.

2. Arquitetura do Projeto

A arquitetura do projeto é dividida em dez etapas sequenciais, que abrangem desde a coleta dos dados brutos até a visualização final das informações. A espinha dorsal da solução é o Amazon S3, que serve como um repositório centralizado de dados (Data Lake) organizado em três camadas distintas: Bronze (dados brutos), Prata (dados processados e limpos) e Ouro (dados agregados e prontos para análise).



1ª Parte: Coleta de Dados

Nesta fase inicial, os dados da PNAD COVID-19 foram coletados manualmente do site do IBGE. Foram selecionados os três últimos meses da pesquisa para garantir a relevância e a atualidade da análise.

- **Ferramenta/Fonte:**
    - **IBGE (Instituto Brasileiro de Geografia e Estatística):** É a principal fonte de dados estatísticos do Brasil. A PNAD COVID-19 foi uma pesquisa fundamental para monitorar os impactos da pandemia na população brasileira, coletando informações sobre saúde e mercado de trabalho.

2ª Parte: Estruturação do Data Lake no AWS S3

A base para o armazenamento de dados foi criada no Amazon S3, utilizando uma arquitetura de Data Lake com três camadas lógicas para organizar os dados em diferentes estágios de processamento.

- **Serviço Utilizado:**
  - **Amazon S3 (Simple Storage Service):** É um serviço de armazenamento de objetos altamente escalável, seguro e durável. No projeto, ele funciona como um repositório central (Data Lake) para todos os dados, independentemente do formato. A divisão em camadas (Bronze, Prata e Ouro) é uma prática recomendada que facilita a governança e o gerenciamento do ciclo de vida dos dados.
    - **Camada Bronze:** Armazena os dados brutos, exatamente como foram coletados do IBGE, no formato CSV. Esta camada serve como uma fonte de verdade, garantindo que os dados originais sejam preservados.
    - **Camada Prata (Silver):** Contém dados que passaram por um processo de limpeza, enriquecimento e padronização. As transformações realizadas aqui preparam os dados para análises mais complexas.
    - **Camada Ouro (Gold):** É a camada final, onde os dados estão refinados, agregados e prontos para o consumo por ferramentas de análise e Business Intelligence (BI). Os dados nesta camada são otimizados para consultas rápidas e eficientes.

3ª Parte: Carregamento dos Dados na Camada Bronze

Os arquivos CSV coletados na primeira etapa foram carregados manualmente na camada Bronze do bucket S3. Este passo marca a ingestão dos dados brutos no ambiente de nuvem.

4ª Parte: Catalogação dos Dados com AWS Crawler e Data Catalog

Com os dados na camada Bronze, foi necessário criar um catálogo de metadados para que eles pudessem ser consultados. O AWS Crawler foi utilizado para escanear os dados e inferir o esquema, que foi então armazenado no AWS Glue Data Catalog. Durante este processo, os dados foram convertidos para o formato Parquet.

- **Serviços Utilizados:**
    - **AWS Glue Crawler:** É um serviço que rastreia automaticamente seus repositórios de dados (como o S3), identifica os formatos de dados e infere esquemas e partições. Ele popula o AWS Glue Data Catalog com essas informações, tornando os dados consultáveis.
    - **AWS Glue Data Catalog:** Atua como um repositório central de metadados. Ele armazena informações sobre a localização, o esquema e as propriedades dos seus dados, funcionando como um catálogo para todos os serviços de análise da AWS, como Athena e Glue.
    - **Formato Parquet:** É um formato de arquivo colunar de código aberto, otimizado para análise de big data. Ele oferece compressão eficiente e melhora significativamente o desempenho das consultas em comparação com formatos baseados em linhas, como o CSV.

5ª Parte: Exploração dos Dados com AWS Athena

Antes de realizar transformações mais complexas, os dados na camada Bronze (já catalogados) foram explorados utilizando o AWS Athena. Esta análise exploratória permitiu entender a estrutura, o conteúdo e a qualidade dos dados, ajudando a planejar as etapas de limpeza e transformação subsequentes.

- **Serviço Utilizado:**
    - **Amazon Athena:** É um serviço de consulta interativa que facilita a análise de dados diretamente no Amazon S3 usando SQL padrão. O Athena é "serverless", o que significa que não há infraestrutura para gerenciar. Foi fundamental para a validação rápida dos dados e para o planejamento das transformações a serem executadas pelo AWS Glue.

6ª Parte: Processamento e Limpeza dos Dados (Camada Prata)

Nesta etapa, um trabalho (job) do AWS Glue foi executado sobre os dados da camada Bronze. O script de ETL (Extração, Transformação e Carga) realizou tarefas de limpeza, como a remoção de registros duplicados e a renomeação de variáveis para nomes mais intuitivos e padronizados. O resultado deste processamento foi salvo na camada Prata do S3, no formato Parquet.

- **Serviço Utilizado:**

- **AWS Glue:** É um serviço de ETL totalmente gerenciado que simplifica a preparação e o carregamento de dados para análise. Ele permite criar, executar e monitorar jobs de ETL em um ambiente serverless, utilizando Apache Spark. Nesta fase, o Glue foi essencial para garantir a qualidade e a consistência dos dados.

7ª Parte: Refinamento dos Dados para Análise (Camada Ouro)

Um segundo job do AWS Glue foi executado, desta vez lendo os dados da camada Prata. O objetivo era refinar ainda mais o conjunto de dados, selecionando apenas as variáveis de interesse para o projeto de análise final. Este subconjunto otimizado foi salvo na camada Ouro, também em formato Parquet, pronto para consumo.

8ª Parte: Extração dos Dados para o Power BI

Os arquivos em formato Parquet, armazenados na camada Ouro do S3, foram baixados para um ambiente local para serem carregados no Power BI.

9ª Parte: Modelagem e Estruturação de Dados no Power BI

Dentro do Power BI, os dados carregados da camada Ouro foram estruturados seguindo um modelo estrela (Star Schema), uma abordagem otimizada para análise e Business Intelligence.

Os relacionamentos entre a tabela fato e as tabelas de dimensão foram estabelecidos, permitindo que as análises sejam facilmente segmentadas e filtradas por diferentes perspectivas. Além da criação do modelo, foram realizadas transformações adicionais, como a criação de colunas calculadas, para enriquecer o modelo e preparar os dados para a camada de visualização. Essa estrutura otimiza a performance das consultas e simplifica a criação de relatórios interativos.

10ª Parte: Criação dos Dashboards

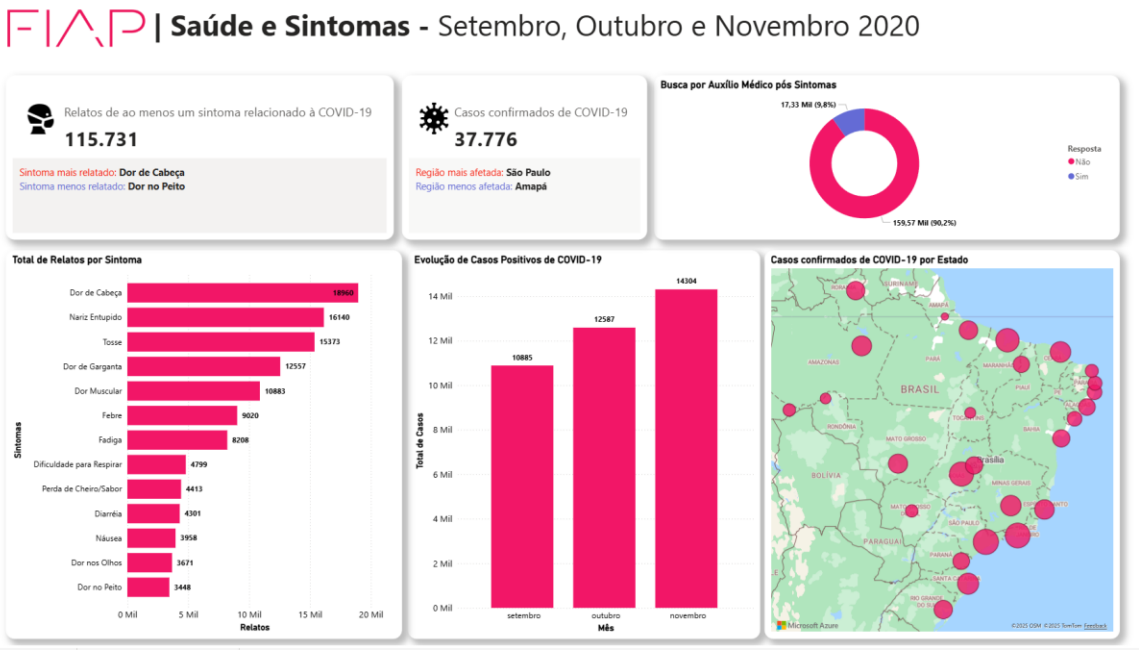
A etapa final consistiu na criação de dashboards interativos no Power BI. Utilizando os dados modelados, foram desenvolvidos gráficos, tabelas e outros elementos visuais para apresentar os insights extraídos da PNAD COVID-19 de forma clara e objetiva.

- **Ferramenta Utilizada:**
    - **Microsoft Power BI:** É uma plataforma de Business Intelligence que oferece ferramentas para agregar, analisar, visualizar e compartilhar dados. Sua capacidade de se conectar a diversas fontes de dados

(incluindo arquivos Parquet) e suas robustas funcionalidades de modelagem e visualização o tornaram a escolha ideal para a camada de apresentação deste projeto.

3. Storytelling dos Painéis

3.1 Painel 1 – Saúde e Sintomas



Este primeiro painel apresenta um panorama epidemiológico e clínico dos últimos três meses da pesquisa (setembro, outubro e novembro de 2020). A análise foca na identificação dos sintomas mais prevalentes, na evolução do número de casos e no comportamento da população em relação à busca por auxílio médico. Estes insights são fundamentais para que o hospital parceiro possa compreender o perfil do paciente com COVID-19 e o contexto da pandemia no período.

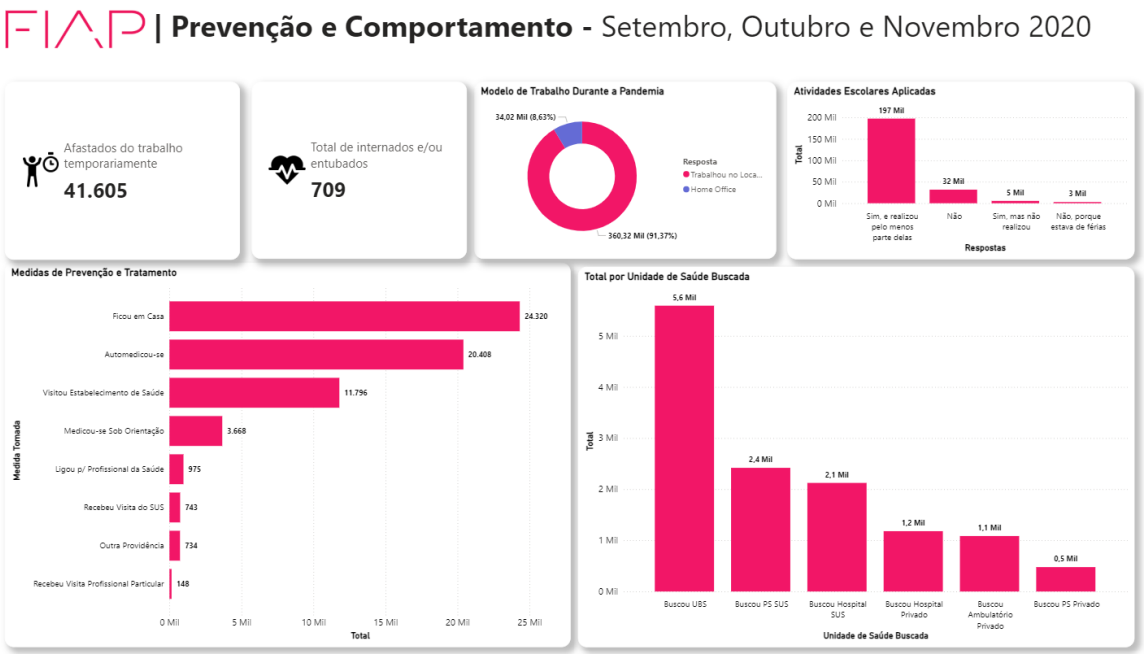
Durante o trimestre analisado, 115.731 pessoas relataram ao menos um sintoma associado à COVID-19. A análise detalhada desses relatos, apresentada no gráfico "Total de Relatos por Sintoma", revela um perfil de apresentação da doença marcado por sintomas comuns e facilmente confundíveis com outras viroses respiratórias. Sendo Dor de Cabeça (18.960 relatos), Nariz Entupido ou Coriza (16.140) e Tosse (15.373) são os sintomas mais comuns,

O gráfico "Busca por Auxílio Médico pós Sintomas" revela uma discrepância imensa entre sentir um sintoma e procurar ajuda profissional. Esse comportamento da população decorreu de diversos fatores, entre eles a recomendação de procura dos serviços de saúde, em aparição de sintomas mais severos [1] de um dos ministros no início da dissipação do vírus da COVID em território nacional. A maior parte dos indivíduos entrevistado com sintomas (90,2%), optou por não procurar um médico ou serviço de saúde. Apenas 9,8% buscaram auxílio. Este fenômeno "iceberg epidemiológico", demonstra que o número de pacientes que chegam aos hospitais e unidades de saúde representa apenas a ponta do problema, existe um vasto contingente de pessoas doentes na comunidade que não entram nas estatísticas oficiais, funcionando como um reservatório de transmissão do vírus. Para o hospital, isso sinaliza o risco de chegadas tardias, com quadros de saúde já agravados, e a possibilidade de surtos repentinos, assim gerando um engarrafamento no gargalo de recepção de sintomáticos.

É possível notar no painel que dos mais de 115 mil sintomáticos, 37.776 foram confirmados como casos de COVID-19. Além da tendência de crescimento de casos ao decorrer dos meses, indicando a aceleração e crescimento constante de casos confirmados. No mapa de indicadores afirma-se que a pandemia não se distribuiu de forma homogênea pelo território. São Paulo foi a região mais afetada, concentrando o maior número de casos, enquanto estados como o Amapá registraram menor impacto. Esse grupamento de casos geográficos, é vital para o planejamento de recursos, pois demonstra que hospitais em grandes centros urbanos, como São Paulo, estiveram no epicentro da crise.

[1] https://www12.senado.leg.br/radio/1/noticia/2021/05/04/mandetta-diz-que-orientacao-para-busca-de-hospitais-apenas-apos-sintomas-era-para-evitar-disseminacao-da-covid-19

3.2 Painel 2 – Prevenção e Comportamento



O painel de **Prevenção e Comportamento**, referente ao período de setembro a novembro de 2020, mostra como a população respondeu às orientações preventivas e como adaptou sua rotina em meio às restrições impostas pela pandemia.

No campo das **medidas de prevenção e tratamento**, cerca de **41% dos entrevistados** relataram que permaneceram em casa como forma de proteção, enquanto aproximadamente **34% recorreram à automedicação**. Apenas **20%** buscaram atendimento em estabelecimentos de saúde, e menos de **6%** seguiram tratamento sob orientação médica direta. O uso dos canais formais de orientação foi ainda mais limitado: pouco mais de **1,5% ligaram para profissionais de saúde**, e menos de **2% receberam atendimento do SUS em domicílio**. Esses dados revelam que, embora o isolamento tenha sido amplamente adotado, houve uma forte dependência de soluções individuais, como a automedicação, e uma baixa utilização do acompanhamento médico remoto ou institucional, o que fragilizou a capacidade de monitoramento e resposta do sistema de saúde.

Quanto ao **modelo de trabalho**, apenas **8,6% da população ativa** conseguiu exercer suas funções em **home office**, enquanto a grande maioria, cerca de **91,4%**, continuou trabalhando de forma presencial. Isso evidencia as barreiras estruturais e setoriais para

o trabalho remoto, sobretudo em ocupações informais e essenciais, que não permitiam essa flexibilização, ampliando a exposição ao contágio.

No setor educacional, os dados reforçam uma realidade desigual: **83% dos estudantes** relataram ter realizado pelo menos parte das atividades escolares à distância, enquanto **14% não tiveram qualquer continuidade pedagógica**. Apenas uma minoria relatou interrupção por férias ou motivos externos.

A análise do **acesso aos serviços de saúde** confirma a centralidade do sistema público: **quase 60% das buscas por atendimento ocorreram em UBSs**, seguidas pelos **prontos-socorros e hospitais do SUS (cerca de 40% somados)**. Já o setor privado respondeu por menos de **15% dos atendimentos**, indicando que, mesmo diante da crise, a maioria da população permaneceu dependente do atendimento público.

Para a gestão hospitalar, são indicados 2 pontos críticos:

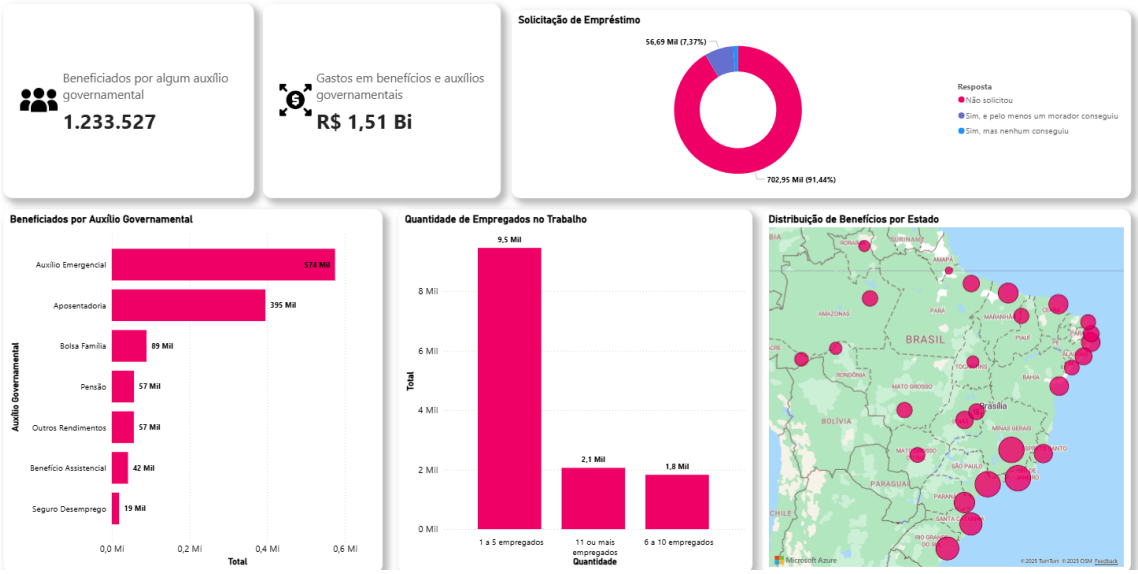
1. A necessidade de campanhas de comunicação mais efetivas para reduzir a automedicação e incentivar o uso dos canais formais de saúde.

2. A importância de estratégias de suporte às populações que não podem adotar o trabalho remoto ou interromper atividades escolares, reduzindo desigualdades e prevenindo riscos adicionais.

Resumindo o painel mostra que, embora medidas individuais tenham sido adotadas, o comportamento coletivo ainda refletia limitações de acesso a infraestrutura adequada para o momento, desigualdade social e forte dependência do SUS, fatores decisivos para o planejamento de futuras ações preventivas.

3.3 Painel 3 – Socioeconomia e Impactos

| Socioeconomia e Impactos - Setembro, Outubro e Novembro 2020



A análise do painel "Socioeconomia e Impactos", referente ao período de setembro a novembro de 2020, apresenta um retrato complexo do país em um momento crucial, revelando uma narrativa de duas realidades paralelas e conflitantes que coexistiram durante a pandemia. A compreensão dessa dualidade é um pilar para transformar dados históricos em inteligência preditiva para o planejamento estratégico de futuras crises em instituições hospitalares. A narrativa que emerge dos dados é a de uma intervenção governamental massiva que, ao mesmo tempo que proveu alívio, também mascarou a profunda agonia do verdadeiro motor da economia: as pequenas empresas e médias empresas.

De um lado, o painel destaca uma injeção de R$ 1,51 bilhão em auxílios governamentais, uma medida que sustentou milhões de famílias e evitou um colapso econômico-social imediato. O Auxílio Emergencial, com R$ 576 milhões, foi a principal ferramenta dessa estratégia, funcionando como uma medida de saúde pública em sua essência. Estudos da época confirmam que, para alguns beneficiários, o valor recebido auxiliou as perdas de renda, tirando temporariamente 13,1 milhões de brasileiros da linha da pobreza, com o pagamento entre R$600 e R$1200 para cada família dependendo de cada caso. Essa intervenção, garantiu um piso mínimo de segurança alimentar e condições de vida, o que provavelmente preveniu um surto ainda maior de doenças relacionadas à desnutrição, à miséria e às complicações da covid 19. Para a gestão hospitalar, a lição é clara: em uma crise futura, a ausência de um programa de suporte de renda similar deve ser interpretada

como um alerta vermelho, sinalizando um aumento iminente na chegada de pacientes com quadros clínicos agravados por extrema vulnerabilidade social, em decorrência da necessidade de buscar alternativas para ter uma renda mínima que garante ao menos a sua alimentação e de sua família.

De outro lado, enquanto o auxílio sustentava o consumo, os dados revelam a crise existencial das microempresas. Responsáveis por 9,3 milhões de empregos, essas empresas, que formam a espinha dorsal do mercado de trabalho, foram atingidas de forma desproporcional. Pesquisas do Sebrae de 2020 indicam que as microempresas tiveram uma das maiores taxas de mortalidade, com muitas atribuindo o fechamento diretamente à pandemia. O fechamento em massa desses negócios representa um fator de risco clínico direto para a população, implicando a potencial perda de cobertura de saúde e o aumento do estresse crônico, um gatilho poderoso para transtornos de ansiedade, depressão e agravamento de doenças preexistentes. Assim, o monitoramento da saúde das micro e pequenas empresas locais oferece um valioso indicador preditivo sobre a futura demanda por serviços de saúde mental e tratamento de doenças crônicas.

A peça que conecta essa narrativa é o colapso no acesso ao capital, exposto nos dados sobre solicitação de empréstimo. Simultaneamente, o endividamento das famílias atingiu o maior patamar em 11 anos. Para uma instituição de saúde, esta falta de crédito é o mecanismo que pode transformar uma crise econômica em uma crise de saúde, ou agravar mais ainda uma crise de saúde pública, pois famílias e empresários sob estresse financeiro extremo tendem a negligenciar cuidados preventivos, chegando aos hospitais com quadros mais graves e complexos, o que eleva o custo e a complexidade do tratamento.

Em suma, a análise revela uma economia que sobreviveu por conta de políticas públicas necessárias como o auxílio emergencial, mas que, entretanto, sofria com a falência de pequenas empresas por falta da disponibilização de crédito. Sendo assim, o ponto de atenção fundamental é que a saúde econômica da comunidade é um sinal vital que deve ser monitorado, o qual, em caso de não ação do Estado com políticas públicas de contingenciamento de crise, pode trazer consequências mais drásticas para o controle de futuras pandemias e endemias, causando uma sobrecarga ainda maior no sistema de saúde público e privado. Com base nisso, as ações recomendadas convergem para uma postura proativa com o: monitoramento ativo de indicadores econômicos locais, como

desemprego e a saúde das MPEs, como um sistema de alerta precoce; o fortalecimento da linha de cuidado em saúde mental para se preparar para picos de demanda em tempos de instabilidade; Construção de materiais digitais de conscientização acerca das condições de saúde do país, contendo cartilhas de prevenção e cuidados em caso de infecção; e a integração da avaliação socioeconômica nos protocolos de assistência social para identificar e apoiar pacientes em vulnerabilidade, prevenindo o agravamento de suas condições de saúde em decorrência de uma maior vulnerabilidade social.

4. Conclusão

A análise da PNAD COVID-19 permitiu compreender de forma abrangente os impactos da pandemia sobre a saúde, o comportamento e a situação socioeconômica da população brasileira. Os dados evidenciaram não apenas a prevalência de sintomas relacionados à doença, mas também a baixa procura por atendimento médico e testagem, indicando um cenário de subnotificação. Observou-se ainda que, apesar da adesão parcial a medidas de prevenção e isolamento, fatores estruturais como a necessidade de trabalho presencial e a desigualdade de acesso a serviços de saúde limitaram a efetividade dessas estratégias. No campo socioeconômico, destacou-se a relevância do auxílio emergencial como mecanismo de suporte às famílias mais vulneráveis, embora micro e pequenas empresas tenham sido fortemente impactadas. Em síntese, o estudo evidencia a complexidade da crise enfrentada, reforçando a importância de políticas públicas integradas de saúde e proteção social, bem como da utilização de dados para orientar decisões em situações de emergências sanitárias.