PETROBRAS

DIAGNÓSTICO DOS DESAFIOS DA APLICAÇÃO DE TÉCNICAS DE CIÊNCIA DE DADOS NO PROCESSO EXPLORATÓRIO

Sumário Executivo

Relatores:

Daniel Thome de Paula – EXP/TPGG/TG

Dean Pereira de Melo – EXP/TAID/TID/GDGEO

Givanildo Santana do Nascimento – EXP/TPGG/TGEO

João Ribeiro Carrilho Junior – EXP/TPGG/TGEO

Laura Silveira Mastella – EXP/TAID/SDGG

Marcelo Delgado Landini de Mattos - EXP/TAID/TID/GDGEOF

Ricardo Augusto Rosa Fernandes – EXP (Coordenador do grupo de trabalho)

Rio de Janeiro
27 de janeiro de 2022

Este sumário condensa as informações constantes do relatório **Diagnóstico dos Desafios da Aplicação de técnicas de Ciência de Dados no processo Exploratório**, realizado pelo grupo de trabalho criado em 18/10/2021, por iniciativa dos gerentes gerais Rogerio Cunha, Jonilton Pessoa e Jefferson Kinzel, e pelo Coordenador de Projetos do EXP100 Otaviano Pessoa. O grupo desenvolveu um diagnóstico sobre o tema Ciência de Dados na Exploração, procurando identificar respostas para a seguinte questão chave:

Onde estão os bloqueios e gaps para que os projetos de ciência de dados agreguem valor ao macroprocesso explorar?

Np diagnóstico foi utilizado método de entrevista direta com perfis selecionados (25 respondentes), sempre com uma dupla de entrevistadores, com formações quase sempre diferentes para cada entrevistado, e tempo médio de 90 minutos. Ao todo, foram realizadas 29 sessões de entrevistas, entre os dias 16/11/2021 e 16/12/2021.

Três principais dimensões associadas ao tema "Ciência de Dados" foram objeto do diagnóstico: 1- **Tecnologia**, que abrange não apenas as ferramentas de ciência de dados (aplicações, recursos de processamento e visualização) e integração entre dados e aplicações mas também a infraestrutura para trabalhos dessa natureza, tanto dependente da TIC quanto dentro da Exploração; 2- **Processos**, que abrange os temas de qualidade e acesso a dados, gestão de dados e processos de trabalho; e 3 – **Pessoas**, que abrange os temas de Capacitação, Papéis e Responsabilidades, Gestão do Conhecimento e Comunicação.

Ao realizar o presente diagnóstico, esperamos poder contribuir em pelo menos 2 frentes:

1. Explicitar e registrar quais as "dores" e percepções dos entrevistados sobre temas críticos para os projetos de Ciência de dados, nas três vertentes principais citadas;
2. Estabelecer uma base de entendimento sobre a situação atual dos projetos que envolvem ciência de dados na Exploração, e poder propor os caminhos e mudanças necessárias para extrair máximo potencial desses projetos, que são um pilar de sustentação fundamental do programa estratégico EXP100.

No planejamento realizado pelo grupo em relação ao método e análise, havia a percepção de que cada tipo de dado apresentaria determinadas peculiaridades, isto é, seriam apresentados problemas específicos para determinados tipos de dados que representariam bloqueios para os cientistas de dados. Tal percepção se confirmou. Para perfis, dados sísmicos, dados de interpretação e dados de rocha, o problema em comum é a dúvida sobre a melhor versão do dado disponível. Os

dados de teste de formação foram os únicos considerados não estruturados. Os perfis brutos foram os únicos a serem considerados inconsistentes, não padronizados e com controle de qualidade problemático, contexto que se revelou mais crítico quando se discutiu curvas oriundas da interpretação petrofísica. O dado sísmico, elogiado em diversos critérios, teve citações de problemas de acesso, versionamento e limitação de processamento. Os dados de rocha, além da questão já citada, apresentaram mais comentários em relação a ferramentas de visualização. Os dados de intepretação são os que mais apresentaram problemas ficando bem acima da média, considerando todo o conjunto. Destacam-se questões de validação, acesso, versionamento, integração e preservação. Ainda que não seja o objetivo deste trabalho sugerir ações, vale investigar os motivos pelos quais processos de preservação das versões indicadas pelos intérpretes não estão (aparentemente) funcionando plenamente ou não estão incluindo o cientista de dados como um usuário deste dado.

Em relação aos bloqueios e gaps, por dimensão, foram identificados:

1) Tecnologia: não integração dos dados nos ambientes de análise, a compartimentação das soluções próprias e contratadas no processamento e interpretação dos dados e a não preservação e compartilhamento dos modelos de IA produzidos nas soluções.

2) Processos: versionamento dos dados, aspecto que de alguma forma incorpora problemas de qualidade de dados, metadados, e o próprio processo de versionamento em si que aparentemente não está estabelecido para a maior parte das áreas. Outro ponto crítico é o acesso a dados, que depende de contatos informais sendo também prejudicado pelos diversos sistemas de informação que não necessariamente seguem uma política similar de concessão de acesso.

3) Pessoas: troca de conhecimento limitada a grupos, dependentes de iniciativas pontuais de pessoas e falta de ambientes colaborativos; há confusão entre cientista de dados e analistas de dados bem como a falta compreensão dos limites e potencialidades da ciência de dados e, somado a isso, verifica-se resistência por parte de alguns grupos. Finalmente, os participantes apontam que a questão da comunicação das soluções de IA precisa melhorar, uma vez que se restringem a pequenos grupos e pessoas e aquelas ligadas ao EXP100.

Um capítulo específico trata da infraestrutura, fundamentalmente associada a TIC, onde diversos pontos críticos são identificados, o mais latente refere-se à falta de pessoal para atendimento aos projetos em tempo hábil. São reportadas, ainda, limitações de hardware e área de armazenamento, para projetos que envolvam modelos de ML/IA. No tema de software e MLOps constata-se ausência

de infraestrutura de software adequada, passando por containers, MLFlow, kubernetes, CI/CD, configurações de proxy e atualizações de sistema operacional – aspectos que só se tornam importantes para profissionais que estejam com alto nível de maturidade no processo de *Machine Learning* e que apesar de serem minoria, representam grandes expectativas de retorno para a Companhia. Outro ponto crítico identificado é que o ambiente de nuvem está, ainda, indisponível para o negócio de forma ampla. Essas dificuldades estimulam soluções próprias e locais, que colocam em risco a preservação do conhecimento e escalabilidade das soluções, além de serem ineficientes e custosas.

Enquanto as respostas pontuadas foram classificadas segundo o critério quantitativo, os comentários na parte dissertativa das respostas geraram uma classificação qualitativa, baseada em manifestações negativas, que permitem capturar a criticidade de cada tema. Uma análise combinando as avaliações quantitativa e qualitativa, utilizando a média dessas avaliações, resultou na estratificação dos pontos mais críticos em escala de 1 a 9 (maior = mais crítico), mostradas na Tabela 1 abaixo, que corresponde a Tabela 5 do relatório. Os pontos identificados serão objeto de discussão interna da EXP para planejamento de ações de correção, seja com resoluções internas ou em articulação com a TIC.

Tabela 1 – Ranqueamento das 7 perguntas que obtiveram respostas com problemas mais críticos.

Questão	Média das avaliações quali-quanti
Questão 08 - Melhor versão dos dados é claramente identificada?	9
Questão 20 - Os modelos de IA produzidos são preservados e compartilhados adequadamente?	9
Questão 04 - Dados estão facilmente acessíveis?	8
Questão 06 - Os dados necessários para Soluções de IA (treinamento e inferência) estão acessíveis e completos?	8
Questão 10 - Perfis de acesso são fornecidos e obtidos com transparência?	8
Questão 12 - Dados estão integrados para uso em ambiente de análise?	8
Questão 36 - Pessoas e equipes que trabalham com ciência de dados trocam conhecimento de forma organizada e sistematizada?	8