PÓS-GRADUAÇÃO - **MBA em Data Science e Analytics**

Universidade de São Paulo (USP)/ ESALQ

2025 – CONCLUÍDO

Após o mestrado acadêmico, diante da expectativa de ingressar no mercado e aplicar meus conhecimentos, escolhi o curso de MBA em Data Science e Analytics, da USP-Esalq.

Dei continuidade aos estudos em modelos de *Machine Learning*, com ênfase em predição e inferência estatística. A análise da base dos Microdados do ENEM (2023) me levou ao estudo do **Modelo Multinível**, considerando interações em dois níveis. O objetivo foi investigar o desempenho dos candidatos a partir do Exame Nacional do Ensino Médio, disponibilizado no acervo público do INEP.

O trabalho envolveu a formatação da base de dados, tratamento de valores nulos (notas dos candidatos) e dados ausentes, estruturando-a para aplicação no problema de modelagem. O modelo possibilitou identificar as principais variáveis que influenciam o desempenho dos candidatos nas provas de Matemática e Redação. Para isso, foi realizada uma amostragem estratificada, considerando tanto características escolares (escolas públicas e privadas) quanto o nível individual de mais de 10.000 alunos. Como resultado, obteve-se uma amostra com distribuição aproximada de Bernoulli, com médias e desvios padrão: Matemática (média = **572,29**; desvio padrão = **125,65**) e Redação (média = **692,70**; desvio padrão = **172,17**).

Entre os principais achados, verificou-se que a posse de equipamentos tecnológicos, ainda que domiciliares, exerce influência significativa no desempenho dos candidatos. O modelo indicou que todas as categorias relacionadas ao acesso a computadores mostraram-se estatisticamente significativas em comparação aos estudantes sem computador em casa. Quanto à diferença entre escolas públicas e privadas, observou-se que, entre candidatos com mais de três computadores em domicílio, as notas tenderam a se equiparar. Essa aproximação pode ser explicada pela presença de *outliers* — candidatos com notas muito acima ou abaixo da média —, especialmente concentrados em alunos de escolas privadas, que em geral apresentaram maiores desempenhos.

Esses resultados reforçam a importância das políticas de inclusão de estudantes da rede pública no ensino superior por meio do sistema de cotas, bem como da ampliação do acesso à tecnologia e às mídias digitais no processo educacional. Contudo, cabe destacar limitações da pesquisa, como a ausência de variáveis específicas sobre a estrutura das instituições escolares da rede pública.

Outras limitações permanecem em aberto e podem ser objeto de estudos futuros, sobretudo no aprofundamento do uso das bibliotecas estatísticas e na revisão de componentes do modelo ainda pouco explorados, como as **aproximações de Laplace e Newton-Raphson** para a análise da verossimilhança entre distribuições.

Sinto-me motivado e agradecido à minha família e a todos que contribuíram para minha formação, em especial àqueles que me incentivam a seguir adiante no aprofundamento teórico e prático dos modelos estatísticos aplicados a dados reais, apesar dos obstáculos encontrados ao longo dessa trajetória.

Dificuldades operacionais do projeto: A análise de um modelo dessa natureza, com base nos Microdados do ENEM, exige infraestrutura computacional compatível. O conjunto de dados do ENEM possui mais de 3 GB, demandando elevado uso de memória RAM e capacidade de processamento nas principais IDEs de análise. Para contornar essas limitações, recorri ao uso do PostgreSQL associado a outras ferramentas de suporte, além de realizar diversos testes exaustivos. Apesar das dificuldades, esse esforço possibilitou alcançar os resultados apresentados neste trabalho.