



Pós-Graduação em Ciência de Dados e Inteligência Artificial

Projeto Aplicado - Relatório Técnico

Modelo Preditivo para Gestão de Carga Horária de Colaboradores Horistas

Grupo : Maria Eduarda Pereira Vergulho; Maria Helena Sonogo Benczik; Valmor Batista Trindade.

1. ENTENDIMENTO DO NEGÓCIO

1.1. Contextualização e definição do problema

Dentro da FIESC, utilizamos como um dos recursos o regime de colaboradores remunerados por hora, denominados **horistas** neste relatório. A utilização desse perfil de profissional nas atividades diárias deve observar um conjunto de normas e diretrizes institucionais.

Entre essas diretrizes, destaca-se a responsabilidade da área de Planejamento e Controle da Produção (PCP) na gestão e no acompanhamento da ocupação dos horistas. Segundo as regras vigentes, não deve permanecer ativo na base nenhum horista sem registro de horas trabalhadas por período superior a sete meses (subutilização). Da mesma forma, não é indicado manter colaboradores que ultrapassem 120 horas mensais por mais de sete meses consecutivos (extrapolação), devendo assim, ser analisada a conversão para mensalistas.

É orientado que os colaboradores que apresentem carga horária 0 por mais de sete meses sejam desligados da instituição e os colaboradores que apresentem carga horária trabalhada mais de 120 horas por mais de 7 meses devem ser convertidos a mensalistas, ou seja, possuir um contrato de trabalho fixo com a instituição recebendo um salário mensal.

A inadequada gestão desse recurso gera impactos significativos na alocação de recursos, nos custos operacionais, no planejamento de escalas e na organização pedagógica, comprometendo a eficiência do processo como um todo.

1.2. Problema a ser resolvido

Atualmente, a análise dessas tendências é realizada de forma manual e reativa, com apoio de painéis de BI desenvolvidos para esse propósito. No entanto, não dispomos de uma avaliação preditiva que permita identificar previamente os professores que:

- poderão ultrapassar 120 horas no mês seguinte (risco de extrapolação);
- poderão apresentar 0 horas trabalhadas (risco de subutilização).

Para aprimorar a capacidade analítica e antecipar possíveis inconsistências, torna-se necessário o desenvolvimento de um modelo automatizado de previsão, capaz de sinalizar esses cenários de risco com antecedência.

1.3. Objetivo do Projeto

Desenvolver um modelo de Machine Learning alimentado pelo banco de dados de “horas trabalhadas por profissional”, com horizonte de previsão de três meses para a carga horária de cada docente. O modelo deverá:

- Gerar previsões mensais (próximos 3 meses) da carga horária por profissional, com estimativas pontuais e intervalos de confiança;
- Emitir sinalizadores de risco para:
 - **Extrapolação**: probabilidade de o docente ultrapassar 120 horas no mês seguinte;
 - **Subutilização**: probabilidade de registrar 0 horas no mês seguinte;
- Fornecer **dashboards e gráficos** que consolidem previsões, riscos, séries históricas e métricas de acurácia, oferecendo suporte visual à tomada de decisão pelo PCP e demais gestores.

2. DESCRIÇÃO DAS BASES DE DADOS UTILIZADAS

2.1. Base utilizada

O projeto utiliza a planilha consolidada com dados de atividades lançadas no sistema SGN, e extraídos do BI corporativo da FIESC.

E após o tratamento dos dados, obtivemos uma planilha com as seguintes colunas:

- Matrícula
- Colaborador
- Total de atividade
- Colunas mensais (ex: 2024-04, 2024-06...) representando total de horas por mês
- Total_Horas

Cada linha corresponde a um colaborador, com o histórico mensal de horas.

Para este projeto, a fim de realizarmos uma prova de conceito, optamos por restringir a base de dados utilizando dados da regional sudeste referentes aos últimos 12 meses.

2.2. Qualidade e estrutura

A base de dados original retirada do BI corporativo FIESC abrange todas as filiais do estado, bem como todas as atividades realizadas pelos colaboradores horistas devidamente registradas e aprovadas para pagamento.

A fim de restringir nossa análise e realizar uma prova de conceito e verificar se o método de Machine Learning escolhido atenderia a expectativa de retorno necessário para realização das análises foi escolhido trabalhar somente com os dados da regional Sudeste, que hoje é a regional que apresenta colaboradores listados extrapolando o limite imposto de 120 horas trabalhadas no mês e maior número de colaboradores registrados que não possuem atividade.

3. TRATAMENTO DOS DADOS

3.1. Transformações temporais

Os dados originais estavam distribuídos por colunas mensais. Para trabalhar com séries temporais, foi necessário:

- Converter colunas mensais em formato datetime
- Reestruturar para formato long (estrutura de série temporal)
- Ordenar cada série por professor e por mês

Essa formatação permitiu uso de modelos que dependem de ordem cronológica.

3.2. Tratamento de dados faltantes

A ausência de valores foi tratada segundo critérios:

- Meses inexistentes no arquivo → considerados 0 horas (periodicidade válida)
- Histórico insuficiente → mantido para evitar distorção artificial
- Buracos em meses intermediários → preenchidos com 0, pois a ausência representa inatividade real.
- Datas em formato ano-mês → convertidas para série temporal
- As horas são somas de carga horária diurna + noturna → consistente
- Não foram identificados valores negativos.

3.3. Tratamento de outliers

Diferentemente de bases públicas ou dados de sensores, a base utilizada neste projeto não é um banco de dados aberto ou oriundo de medições externas. Trata-se de um registro administrativo controlado, no qual a carga horária mensal de um colaborador necessariamente se encontra dentro de um intervalo operacional delimitado pela própria política institucional.

Ao analisar a distribuição de horas mensais por professor, observou-se que todos os valores se situam entre 0 e 200 horas, faixa que corresponde:

- aos limites contratuais vigentes,
- às regras internas de alocação docente,
- aos padrões de atividades acadêmicas típicas.

Portanto:

- Não existem outliers estatísticos,
- Não existem valores atípicos artificiais,
- Não existe necessidade de remoção ou tratamento de valores extremos, porque todo o conjunto de dados já se encontra naturalmente limitado e validado pela instituição que o gera.

Dessa forma, a etapa de detecção e correção de outliers foi documentada, mas não exigiu ações adicionais, uma vez que os dados:

- estão dentro dos limites aceitáveis,
- seguem coerência operacional,
- não apresentam valores anômalos que necessitem intervenção.

Essa característica é comum em bases administrativas de RH, nas quais o processo de geração dos registros impede automaticamente a ocorrência de valores fora do domínio esperado.

3.4. Engenharia de atributos (features criadas)

Esta etapa é fundamental em modelagem de séries temporais com ML, pois transforma dados brutos em informações úteis.

Features criadas:

1. **Lag Features** (valores passados) - permitem ao modelo aprender dependência temporal.
 - lag_1 → horas do mês anterior
 - lag_2 → duas competências anteriores
 - lag_3 → três competências anteriores
2. **Rolling Statistics (tendências)**
 - media_movel_3m → suaviza ruídos e captura comportamento médio
 - desvio_3m → mede instabilidade da carga horária
 - tendencia → taxa de variação da média móvel
3. **Variação percentual mês a mês**
 - $variacao_mensal = (horas_atual - horas_anterior) / horas_anterior$
4. **Codificação de mês** - Essas variáveis ajudam o modelo a captar sazonalidade.
 - mes_num → 1 a 12
 - ano → valor numérico

3.5. Organização final

Após criação das features os dados foram escalados apenas quando necessário (modelos baseados em árvores não exigem) e a divisão entre treino e teste foi temporal, preservando ordem cronológica.

4. MACHINE LEARNING - TÉCNICA ESCOLHIDA

4.1. Técnica escolhida

Optou-se pela utilização do modelo Random Forest Regressor porque ele apresenta um desempenho consistente em séries temporais curtas, característica importante para este estudo visto que a fim de fazermos uma prova de conceito do modelo limitamos a base de dados. Além disso, o modelo lida de forma eficaz com dados ruidosos, reduzindo o impacto de variações inesperadas e valores atípicos graças à sua estrutura baseada em múltiplas árvores. Outro ponto relevante é que o Random Forest não exige normalização das variáveis, o que simplifica o processo de preparação dos dados sem comprometer a qualidade das previsões. Também se destaca por sua robustez em bases pequenas, conseguindo generalizar bem mesmo com um volume reduzido de amostras.

Embora o XGBoost tenha sido considerado como alternativa, sua aplicação não foi necessária nesta etapa, uma vez que o Random Forest atendeu plenamente aos requisitos e objetivos da fase atual do projeto.

4.2. Processo de Treinamento

A separação entre os conjuntos de treino e teste foi realizada respeitando a ordem cronológica da série temporal, garantindo que o modelo fosse avaliado apenas com dados futuros em relação ao período de treinamento, preservando assim a integridade temporal da análise. O modelo foi treinado utilizando *features* de defasagem (lag) e variáveis de tendência, o que permitiu capturar padrões históricos relevantes e movimentos estruturais da série ao longo do tempo. Para estimar os valores futuros, aplicou-se uma abordagem de simulação autorregressiva, na qual as previsões geradas para um período são utilizadas como entrada para estimar os períodos seguintes, possibilitando projetar com consistência os três meses futuros.

4.3. Interpretação dos resultados

As análises de tendência evidenciaram que professores que apresentam aumento consistente de carga horária ao longo do histórico tendem a projetar valores superiores a 120 horas nos meses futuros, o que reforça a coerência do modelo em identificar padrões de crescimento. Por outro lado, professores cuja carga demonstra trajetória de queda aparecem com projeções próximas de 0 hora, indicando redução progressiva e alinhada ao comportamento observado

nos dados. De modo geral, o modelo demonstrou boa capacidade de capturar oscilações e variações naturais da carga horária, respondendo de forma adequada às mudanças estruturais presentes na série temporal.

5. AVALIAÇÃO DOS RESULTADOS

5.1. Métricas analisadas

Como o problema é regressão com interpretação binária posterior, usamos:

- **RMSE** para erro das previsões:
 - Valor típico esperado: entre 8h e 20h por mês (varia conforme a base).
- **Acurácia dos riscos binários (0 ou 1):**
 - Indicador de sensibilidade e especificidade das categorias risco.

Para classificação derivada:

- **Acurácia** na detecção de risco >120h e <10h
- **Sensibilidade** para risco de excesso
- **Especificidade** para risco de subutilização

5.2. Limitações

Algumas limitações observadas no estudo dizem respeito à disponibilidade e à completude dos dados. Em primeiro lugar, determinados colaboradores possuem poucos meses registrados, seja por terem ingressado recentemente na instituição ou por terem passado longos períodos afastados, o que reduz a consistência histórica necessária para modelos preditivos mais robustos.

Além disso, eventos externos como férias e afastamentos não são refletidos nos registros analisados, pois essas ocorrências não representam horas produtivas e, portanto, não são contabilizadas para pagamento; tais informações são registradas exclusivamente no sistema do RH, que não possui integração com o SGN, resultando em lacunas importantes na série temporal.

Por fim, mudanças de contrato — como a conversão de um professor horista para mensalista — não foram incorporadas como *features*, o que pode impactar a interpretação e a continuidade das séries individuais.

5.3. Possíveis melhorias

Nesta fase do estudo, optou-se por trabalhar exclusivamente com a regional Sudeste com o objetivo de estabelecer uma prova de conceito. Essa abordagem permitiu validar a metodologia em um ambiente controlado antes de realizar uma expansão mais ampla. No entanto, para obter e expandir a análise a ampliação para todas as regionais é recomendada, possibilitando uma avaliação completa e representativa.

Outro ponto é a possibilidade de incorporar dados administrativos, tais como tipo de contrato, turno e campus. Todas essas informações já estão disponíveis na base e podem enriquecer substancialmente o modelo, permitindo análises mais aprofundadas e maior precisão nas projeções. Da mesma forma, a utilização de modelos híbridos, combinando Random Forest e Prophet, representa uma oportunidade para capturar diferentes dimensões da dinâmica temporal, ampliando a capacidade preditiva.

O treinamento de modelos organizados por clusters de instrutores horistas também é uma estratégia a ser considerada, visto que o corpo profissional é bastante heterogêneo, abrangendo áreas como educação, consultoria empresarial, saúde e atividades físicas. A segmentação por perfis de atuação tende a gerar modelos mais especializados e sensíveis às particularidades de cada grupo.

Por fim, a inclusão de indicadores acadêmicos, como número de turmas e disciplinas associadas, pode agregar contexto operacional relevante ao processo de previsão, contribuindo para aumentar a robustez do estudo e fornecer subsídios mais completos para a tomada de decisão.