

# Um Modelo de Aprendizagem Automática Para Previsão de Ausências de Maquinistas

Gonçalo Matos<sup>1</sup>, Luís Albino<sup>1</sup>, Ricardo L. Saldanha<sup>1</sup>

<sup>1</sup>SISCOG – Sistemas Cognitivos SA., Campo Grande 378 - 3º, 1700-097 Lisboa, Portugal  
e-mail: [rsaldanha@siscog.pt](mailto:rsaldanha@siscog.pt)    <https://siscog.pt>

---

## Sumário

*A capacidade de prever ausências de tripulantes é de extrema utilidade para operadores ferroviários, para apoiar a decisão dos planeadores na gestão do trabalho de tripulantes, contribuindo para a redução de despesas com compensações e horas extraordinárias e para a satisfação do cliente na medida em que se minimiza o risco de cancelamento de serviço devido à falta de tripulantes.*

*O problema em estudo consiste em desenvolver um modelo de aprendizagem automática (machine learning) capaz de prever o número de ausências para um determinado dia de operação, tendo como base um conjunto de dados históricos providenciados por um operador ferroviário da Europa do norte.*

---

**Palavras-chave:** Previsão de ausências; Aprendizagem automática; Planeamento de tripulações.

## 1. INTRODUÇÃO

Neste trabalho propomos um modelo de previsão de ausências de tripulantes de operadores ferroviários.

A capacidade de prever ausências é de extrema utilidade para apoiar a decisão dos planeadores no contexto do planeamento e gestão do trabalho de tripulantes, em particular no planeamento de turnos de reserva e na negociação de dias de folga, contribuindo assim para a redução de despesas com compensações e horas extraordinárias e para a satisfação do cliente, na medida em que se minimiza o risco de cancelamento de serviço devido à falta de tripulantes.

O problema em estudo consiste em desenvolver um modelo de aprendizagem automática (*machine learning*) capaz de prever o número de ausências para um determinado dia de operação, tendo como base um conjunto de dados históricos providenciados por um operador ferroviário, onde se incluem a duração e caracterização dos turnos planeados, dados demográficos acerca dos maquinistas, e estatísticas sobre as suas ausências ao longo do tempo.

Na perspetiva de um operador ferroviário, um horizonte temporal interessante para as previsões pode corresponder às próximas 48 horas — para efeitos de gestão em tempo real, permitindo redistribuir o trabalho planeado em reservas por forma a aumentar os níveis de reserva lá onde é mais provável haver tripulantes ausentes e vice-versa —, entre uma semana a um mês — para planeamentos de curto prazo, onde é possível aconselhar o planeador a não planear ausências ou folgas em dias em que o absentismo previsto é elevado —, ou até um ano (se o modelo o permitir) — para planeamento de longo prazo, permitindo um melhor planeamento de longo prazo de reservas e formações, e um processo mais eficiente de planeamento anual de férias.

A previsão de ausências não é um problema novo, existindo já literatura sobre o assunto, nomeadamente na previsão do absentismo em organizações [1] [2]. No sector dos transportes, conhecem-se também alguns trabalhos na área da aviação [3] [4]. Fora estes porém, e em concreto na área da ferrovia, desconhecemos a existência de outros trabalhos que, à semelhança do nosso, procurem prever ausências de maquinistas.

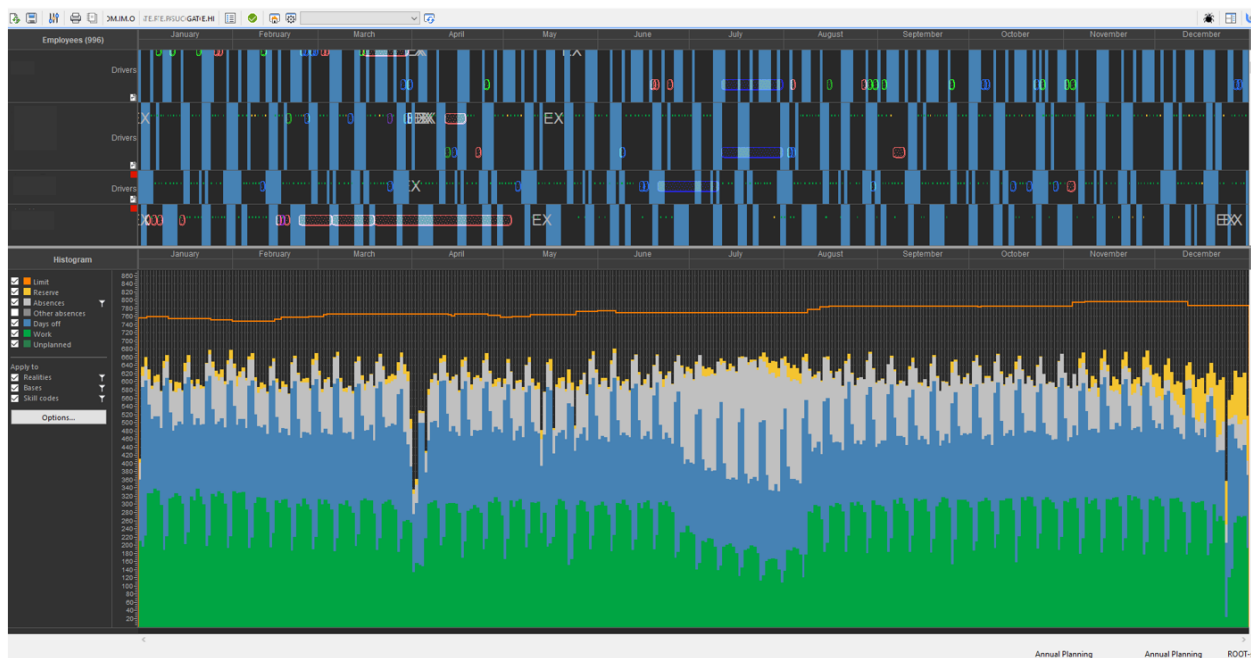


Figura 1. Perspectiva da funcionalidade Annual Planning do produto CREWS – com informação de ausências e reservas.

O presente trabalho deve ser visto como uma investigação em curso, da qual resulta um protótipo com as características que aqui se apresentam. Uma vez finalizado o protótipo, está prevista a sua integração na aplicação CREWS<sup>1</sup>, uma ferramenta que fornece apoio à decisão para o planeamento e gestão do trabalho de pessoal circulante e não circulante, atualmente usado em produção em vários operadores ferroviários e metropolitanos europeus (ver Figura 1).

## 2. METODOLOGIA

### 2.1. Descrição dos dados

Para este estudo, recorremos a um histórico de 7 anos composto por dados reais providenciados por um operador ferroviário do norte da Europa. Utilizamos os dados de absentismo de maquinistas referentes ao período de 2013 a 2018 para treinar o modelo de aprendizagem automática, e os dados relativos ao ano de 2019 como conjunto de teste, para validar o modelo e retirar conclusões acerca da sua qualidade e capacidade de extrapolação para dados nunca antes vistos. Apesar de termos também disponíveis os dados relativos aos anos de 2020 e 2021, devido à atipicidade provocada pela situação pandémica quer ao nível da procura quer ao nível da oferta de força laboral, optámos por não os utilizar no nosso estudo.

Estes dados são obtidos através dos registos de atividade gerados pela aplicação CREWS, sendo compostos por diferentes mensagens emitidas pelo sistema em determinados eventos-chave, tais como a transferência de trabalho planeado a curto prazo (*short term scheduler*) para a responsabilidade dos planeadores de tempo real (*real time dispatchers*), o registo do trabalho realizado por um tripulante no final de um turno, ou o registo de uma ausência.

<sup>1</sup> <https://www.siscog.pt/pt/produtos/#crews>

De salientar que cada um destes registos é associado a um trabalhador e a uma data específica. Assim, os dados com que alimentamos os algoritmos de aprendizagem podem representar-se numa tabela com uma linha referente a cada tripulante para um determinado dia, e contendo múltiplas colunas correspondentes a diferentes características acerca daquela mensagem emitida pelo sistema. Acresce ainda uma coluna com um valor binário que indica se aquele trabalhador esteve ausente ou não no dia correspondente àquele registo.

Os modelos de aprendizagem aprendem, portanto, a prever, para mensagens semelhantes no futuro, se o tripulante se irá ausentar ou cumprir com o trabalho planeado. Como o nosso objetivo final é estimar o número total de ausências para um determinado dia, somam-se então os resultados das previsões individuais para obter o número total de tripulantes ausentes.

Por questões de privacidade e proteção de dados, e para evitar o enviesamento dos algoritmos de aprendizagem, nenhuma das colunas presentes nos dados identifica pessoalmente o trabalhador em causa. Ao utilizador final é apresentado somente o número total de ausências estimadas, e não a previsão individualizada para cada trabalhador.

### 2.2. Preparação dos dados

Seguimos neste trabalho a metodologia KDD (*Knowledge Discovery in Databases*) para analisar os dados disponíveis, selecionar e também gerar um conjunto de características (*features*) que considerámos pertinentes tendo em conta a literatura existente sobre este problema [4], [5] e o conhecimento do domínio.

Como o nosso objetivo é criar um sistema capaz de prever ausências não planeadas, de que são exemplo aquelas motivadas por razões de doença do próprio ou apoio à família, excluímos à partida deste estudo todos os registos de outros tipos de ausências, tais como férias ou dias de formação, que serão já do conhecimento do planeador.

No decurso do processo de *feature engineering*, gerámos também algumas características (*features*) que não estavam explicitamente presentes nos dados originais, tais como as datas de feriados nacionais naquele país, a estação do ano, e o número médio de dias de ausência por ano para cada maquinista ao longo do tempo.

Nos dados que nos foram disponibilizados, os casos positivos (ausências) correspondem apenas a cerca de 10% dos registos, pelo que estes tiveram de ser *balanceados*, uma vez que em alguns algoritmos de aprendizagem automática pode ocorrer um enviesamento das previsões caso não se efetue esta correção.

Algumas das características (*features*) com domínios contínuos ou com um grande intervalo de valores possíveis, tais como a duração de um turno ou as horas de início e fim do mesmo, foram *discretizadas* num pequeno número de categorias, como curto/médio/longo e dia/noite/madrugada, respetivamente. Este é um passo não só útil para reduzir a dimensionalidade e conseqüente complexidade dos dados, como também necessário para alguns algoritmos de aprendizagem.

### 2.3. Treino dos algoritmos

Foram já explorados alguns algoritmos de aprendizagem automática, estando outros ainda em fase de exploração, nomeadamente algoritmos baseados em árvores de decisão (Random Forests, XGBoost) e Regressão Logística. De entre estes, e com base em resultados preliminares, o algoritmo de Regressão Logística mostra-se particularmente promissor.

O treino dos algoritmos efetua-se, como já se disse, num conjunto de dados de absentismo de maquinistas referentes ao período de 2013 a 2018.

## 2.4. Metodologia de teste

Para avaliar a qualidade das previsões obtidas com cada um dos algoritmos, e assim compará-los entre si, utilizou-se um conjunto de dados de teste correspondente ao ano de 2019, em tudo semelhante aos dados utilizados durante o treino.

Para simular a utilização do sistema de previsões num contexto real, modificaram-se estes dados de teste removendo ou “mascarando” os valores de algumas características que não poderiam ser conhecidas à data da previsão, por conterem conhecimento relativo ao que se passa no futuro. Por exemplo, uma das características que adicionámos aos dados, no decurso do processo de *feature engineering*, foi o número de dias decorridos desde a última ausência do maquinista. Antes de fazermos previsões sobre o futuro utilizando algum dos algoritmos, substituímos os valores reais desta característica presente nos dados de 2019 por um valor que corresponde ao último conhecido nos dados de treino (à data de 31/12/2018) incrementado do número de dias decorridos desde o início do período de previsão. Por outras palavras, face à imprevisibilidade do futuro, ao realizar uma previsão para um determinado dia, o nosso modelo assume simplesmente que o maquinista nunca terá faltado até ao dia anterior ao da previsão, incrementando-se por isso diariamente, em uma unidade, o número de dias decorridos desde a última ausência conhecida.

## 3. RESULTADOS PRELIMINARES

Seguindo a metodologia já descrita, o nosso modelo conseguiu prever as ausências para o ano de 2019. Mais concretamente, de um total de 250 095 turnos planeados, em que em 21 015 se verificaram ausências, o modelo previu 21 134 ausências, cometendo assim um erro por excesso de cerca de 0,5% no número total de ausências no ano. Se analisarmos caso a caso, para cada tripulante, se a previsão está ou não correta, a precisão do modelo ronda os 87%.

O gráfico da Figura 2 ilustra a previsão diária de ausências prevista pelo modelo Regressão Logística, para um

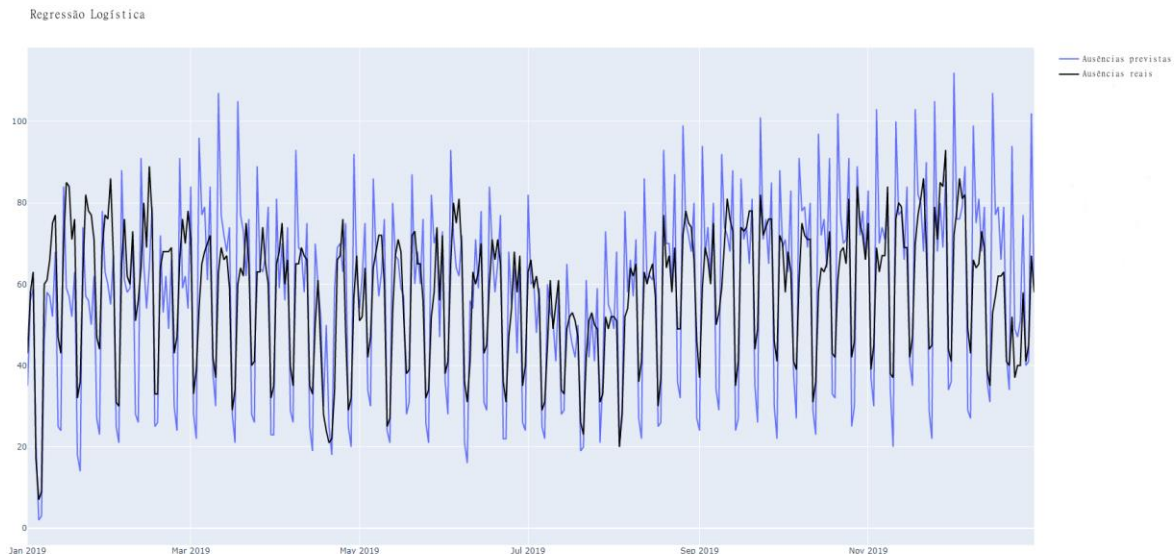


Figura 2. Desempenho do modelo Regressão Logística na previsão de ausências para um horizonte de um ano (2019). A preto os dados reais (históricos), a roxo as previsões do modelo.

horizonte temporal de um ano, e contrasta esses valores com os reais para o mesmo ano de 2019. É interessante salientar que a precisão do modelo não parece ser significativamente afetada pela distância do horizonte da previsão, mantendo-

se aproximadamente constante durante todo o período de teste, à exceção dos meses de Janeiro a Março de 2019, onde o número real de ausências foi particularmente atípico.

O Quadro 1 resume o desempenho do mesmo modelo através de um conjunto de métricas padrão conhecidas na literatura, para previsões com diferentes horizontes temporais. Uma vez que, por vezes, existem dias atípicos que são difíceis de prever e que fazem variar estas métricas ao longo do período de teste, para determinar estas métricas de forma representativa utilizou-se uma estratégia que simula a operação do modelo numa situação real, durante um ano. Assim, o modelo é re-treinado para cada dia de 2018, tendo disponível em cada iteração mais um dia de dados históricos do que na previsão anterior, e é utilizado para prever as ausências a 7 dias, 1 mês e 1 ano. As métricas que se apresentam no Quadro 1 são então calculadas sobre a totalidade das previsões agregadas, para todo o período de teste (de 1 ano), em cada um daqueles horizontes (7 dias, 1 mês, 1 ano).

*Quadro 1. Métricas de desempenho do modelo Regressão Logística para previsões a diferentes horizontes temporais.*

<b>Métrica</b>	<b>previsão a 7 dias</b>	<b>previsão a 1 mês</b>	<b>previsão a 1 ano</b>
<i>Erro médio da previsão (MFE)</i>	-5.45	-5.32	-5.27
<i>Erro absoluto médio (MAE)</i>	11.79	11.89	11.66
<i>Erro percentual absoluto médio (MAPE)</i>	25.14%	25.03%	22.27%

Mais uma vez se verifica a característica interessante deste modelo, aparentemente, não sofrer perdas significativas de desempenho quando o horizonte temporal da previsão aumenta. Observa-se até, neste quadro, uma ligeira melhoria dos resultados de algumas métricas para horizontes de previsão mais distantes, que poderá estar relacionada com flutuações naturais e imprevisíveis do número real de ausências.

## 4. INTEGRAÇÃO NO CREWS

Como já se disse na introdução, o presente estudo foca-se no desenvolvimento de um modelo de aprendizagem automática (*machine learning*), resultando num protótipo demonstrável que, uma vez finalizado, será integrado na aplicação CREWS para poder ser usado em produção pelos clientes da SISCOG, nomeadamente o operador ferroviário da Europa do Norte que forneceu os dados para o desenvolvimento do protótipo.

Não obstante, o desenvolvimento assenta na utilização de dados históricos sob a forma de registos de atividade resultantes de *mensagens emitidas pelo sistema CREWS*<sup>2</sup> em ambiente de produção.

Tal como referido, o CREWS é um produto *standard* que fornece apoio à decisão para a criação e alteração de planos operacionais otimizados, nomeadamente de pessoal circulante e local, considerando horários das viagens e planos operacionais de veículos, competências e preferências do pessoal, bem como restrições/regras laborais e operacionais [6]. Tendo sido distinguido três vezes pela Association for the Advancement of Artificial Intelligence (AAAI) (outrora American Association for Artificial Intelligence) como uma aplicação inovadora de inteligência artificial, o CREWS incorpora inovações que mereceram publicações tais como [7] (módulos de longo-prazo), [8] (módulo de curto-prazo) e [9] (módulo de gestão de operações em tempo real). A Figura 3 ilustra uma escala de pessoal planeada através do módulo de curto prazo do CREWS.

<sup>2</sup> <https://www.siscog.pt/pt/produtos/#crews>

Duties (9)	Sunday 01/22	Monday 01/23	Tuesday 01/24	Wednesday 01/25	Thursday 01/26	Friday 01/27	Saturday 01/28	Sunday 01/29	Monday 01/30	Tuesday 01/31	Wednesday 02/01	Thursday 02/02	Friday 02/03	Saturday 02/04
Plans (149)														
Driver #1	R					R				R				R
Driver #2	4889	3850	R	2383	1182	1210	R	R		8620	6501	R	554	5748
Driver #3	R		6920	6071		554	5748	6457	4043	R	697	2639	4913	R
Driver #4	4884	R	884	1644	1888	2887	802	R	R	R	3232	708	4271	4864
Driver #5	R	R	R	3232	708	4271	4584	3417	R	7243	2031	3184	3869	2271
Driver #6	3412	R	2243	2031	3184	3869	2271	R	R	R	1127	4106		390
Driver #7	R	R	R	1317	4185		308	2084	R	688	1804	1789	1811	R
Driver #8	2035	R	988	1804	1785	1011	R	R	3742	3247	R		6549	5548
Driver #9	R	3742	3247	R	6549	5548	6832	5492	4089	R	2161	1583	89	4853

Figura 3. Exemplo de uma escala planeada pelo CREWS.

Uma vez que os dados provêm deste sistema, e que o presente protótipo lhes acede diretamente através de uma ligação à base de dados que suporta aquela mesma aplicação, a integração deste modelo como um módulo no produto CREWS será muito natural e coesa. Não se trata, portanto, de um trabalho de investigação isolado, do qual resulte uma ferramenta singular e independente, mas sim de uma melhoria dos sistemas existentes e já em produção em clientes, devidamente enquadrada e integrada.

Dado que os tempos de processamento dos dados e treino dos modelos de aprendizagem são relativamente curtos (poucos minutos), mesmo para um conjunto de dados de treino com 7 anos, é viável um processo de integração no produto em que os modelos são treinados diariamente (por exemplo, todas as noites) nos sistemas do cliente, incorporando assim os novos dados reais provenientes de mais um dia de operação nas previsões efetuadas no dia seguinte.

## 5. LIMITAÇÕES E TRABALHO FUTURO

O sistema apresentado encontra-se ainda em fase de protótipo. Assim, vislumbramos como um dos desenvolvimentos próximos a sua integração no produto CREWS, conforme explicado na secção anterior.

Para além disso, a integração deste módulo de previsão de ausências no produto abrirá as portas para outras extensões e melhorias, tais como o planeamento de reservas, de férias e de formações, que poderão beneficiar assim também de desenvolvimentos futuros.

Uma vez que os resultados dos modelos de previsão não são perfeitos, sobretudo em dias atípicos que fogem do padrão habitual de ausências presente nos dados históricos, qualquer desenvolvimento dos mesmos no sentido de melhorar a qualidade das previsões será muito vantajoso para as áreas de aplicação já citadas. Nesse sentido, acreditamos que haverá margem para melhorias quer na exploração de outros algoritmos de aprendizagem ou técnicas de tratamento dos dados, quer no enriquecimento dos próprios dados, com a inclusão de novas características (*features*), tais como dados meteorológicos ou dados sobre a fadiga dos tripulantes, que possam ter alguma influência no absentismo.

## 6. CONCLUSÕES

Neste trabalho propôs-se um modelo de previsão de ausências de tripulantes de operadores ferroviários baseado em modelos de aprendizagem automática (*machine learning*) e dados reais provenientes de um sistema em produção num operador ferroviário do norte da Europa.

De entre os modelos analisados, destacou-se o modelo de Regressão Logística, com uma precisão nas previsões individuais a 1 ano a rondar, em média, os 87%. Apesar de os resultados agregados serem bastante promissores, é preciso recordar que o tipo de ausências que se está a tentar prever é, por natureza, muito imprevisível e sujeito a flutuações bruscas que prejudicam os resultados de qualquer dos modelos estudados, em determinados dias atípicos. Trata-se, portanto, de uma tarefa complexa que nenhum dos modelos até agora estudados soluciona na perfeição ou garante resultados satisfatórios para qualquer dia de operação.

Tratando-se de um projeto de investigação ainda em fase de protótipo, identificou-se desde já como trabalho futuro a integração deste no produto existente (CREWS). Esta integração do modelo num ambiente de produção poderá suportar o trabalho dos planeadores, permitindo-lhes alcançar um melhor planeamento de férias, reservas e formações.

## 7. REFERÊNCIAS

- [1] N. Lawrance, G. Petrides e M.-A. Guerry, “Predicting employee absenteeism for cost effective interventions,” *Decision Support Systems*, vol. 147, p. 113539, 2021.
- [2] S. Ali Shah, I. Uddin, F. Aziz, S. Ahmad, M. A. Al-Khasawneh e M. Sharaf, “An enhanced deep neural network for predicting workplace absenteeism,” *Complexity*, vol. 2020, 2020.
- [3] A.-H. Homaie-Shandizi, V. P. Nia, M. Gamache e B. Agard, “Flight deck crew reserve: From data to forecasting,” *Engineering Applications of Artificial Intelligence*, vol. 50, pp. 106-114, 2016.
- [4] A. H. Homaie Shandizi, Prediction of Pilot’s Absenteeism in an Airline Company, Masters thesis, École Polytechnique de Montréal, 2014.
- [5] A. Asiri e M. Abdullah, “Employees absenteeism factors based on data analysis and classification,” *Special Issue in Communication and Information Technology*, pp. 119--127, 2019.
- [6] SISCOG - Sistemas Cognitivos, S.A., “CREWS — Planeamento e Gestão de Pessoal,” 21 03 2022. [Online]. Available: <https://www.siscog.pt/pt/produtos/#crews>.
- [7] E. J. W. Abbink, L. Albino, T. Dollevoet, D. Huisman, J. Roussado e R. L. Saldanha, “Solving Large Scale Crew Scheduling Problems in Practice,” *Public Transport*, vol. 3, nº 2, p. 149–164, 2011.
- [8] J. P. Martins, E. Morgado e R. Haugen, “TPO: A System for Scheduling and Managing Train Crew in Norway,” em *Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-2003)*, Menlo Park (CA), 2003.
- [9] R. L. Saldanha, A. Frazão, J. P. Martins e E. Morgado, “Managing Operations in Real-time,” *WIT Transactions on The Built Environment, Vol 127*, pp. 521-531, 2012.