POLITÉCNICO DE LEIRIA | ESCOLA SUPERIOR DE EDUCAÇÃO E CIÊNCIAS SOCIAIS

Sistemas de Gestão Escolar e Previsão de Insucesso: o caso de uma Escola Secundária do Distrito de Leiria

Dissertação de Mestrado

Henrique Jorge Leal Sales Fidalgo

Trabalho realizado sob a orientação de

Doutor Filipe Santos

Leiria, Março, 2024

Mestrado em Utilização Pedagógica das TIC

ESCOLA SUPERIOR DE EDUCAÇÃO E CIÊNCIAS SOCIAIS

INSTITUTO POLITÉCNICO DE LEIRIA

AGRADECIMENTOS

Ao professor doutor Filipe Santos, meu orientador, pela sua capacidade de partilhar conhecimento, entusiasmar e fazer despertar o interesse pela ciência de dados que tanto tem para crescer nas escolas.

À minha família pelo apoio e sobretudo pelo tempo em comum que para o desenvolvimento deste trabalho foi necessário abdicar.

RESUMO

Este estudo tem como objetivo analisar o potencial dos dados digitais de uma escola na aplicação de técnicas de análise provenientes da ciência de dados para identificar precocemente, possíveis casos de insucesso escolar. Ao explorar esta abordagem, espera-se entender se os dados registados no *software* de gestão de uma escola têm qualidade para poder produzir padrões e correlações significativas que possam ser utilizados para intervenções precoces e personalizadas, com vista a melhorar os resultados académicos e promover o sucesso dos alunos.

Os objetivos serão avaliar a qualidade da informação dos dados registado no *software* de gestão escolar de uma escola e analisar a possibilidade de correlação entre diversas variáveis contidas nesses dados com o sucesso escolar.

Para a realização deste estudo, foram recolhidos dados do *software* de gestão escolar, os quais foram posteriormente sujeitos a um processo de tratamento para identificar o seu potencial na identificação de eventuais padrões. Adicionalmente, foi conduzida uma entrevista com a responsável pela inserção dos dados relativos aos alunos no *software* de gestão escolar, com o propósito de analisar potenciais falhas nesse procedimento.

Os resultados e conclusões deste estudo identificam que o uso dos dados digitais em algoritmos de inteligência artificial depende não apenas da qualidade e da uniformização da introdução da informação nos registos digitais, mas também de um grande volume de dados. No futuro, as técnicas de análise existentes na ciência de dados podem contribuir significativamente para auxiliar na tomada de decisões relacionadas ao combate do insucesso escolar.

Palavras chave

Ciência de dados, Sucesso Escolar, Inteligência Artificial

ABSTRACT

This study aims to analyze the potential of digital data from a school in applying data science analysis techniques to identify potential cases of academic underachievement early on. By exploring this approach, the goal is to understand if the data recorded in a school's management *software* is of sufficient quality to produce significant patterns and correlations that can be used for early and personalized interventions to improve academic outcomes and promote student success.

The objectives are to assess the quality of the data recorded in the school's management *software* and to analyze the potential correlation between various variables contained in this data and academic success.

To conduct this study, data was collected from the school's management *software*, which was subsequently subjected to a processing treatment process to identify its potential in identifying any patterns. Additionally, an interview was conducted with the head of the administrative team responsible for inputting data into school management *software*, with the purpose of analyzing potential shortcomings in this process.

The results and conclusions of this study identify that the use of digital data in artificial intelligence algorithms depends not only on the quality and standardization of data input into digital records but also on a large volume of data. In the future, data science analysis techniques can significantly contribute to assisting in decision-making related to addressing academic underachievement.

Keywords

Data Science, Academic Success, Artificial Intelligence

ÍNDICE GERAL

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE TABELAS

ABREVIATURAS

EE – Encarregado de Educação

EDM - *Educational Data Mining*

IA – Inteligência Artificial

TIC - Tecnologias de Informação e Comunicação

INTRODUÇÃO

O objetivo máximo da direção de uma escola é o sucesso escolar dos seus alunos, para atingir esse objetivo, no início do ano letivo, são definidas estratégias pedagógicas a aplicar ao longo do ano letivo. Estratégias essas que passam pela criação de apoios educativos, criar clubes e projetos sendo atribuídas horas letivas do crédito horário definido pela tutela para cada escola.

Atualmente as decisões do crédito horário a atribuir são pouco fundamentadas, sendo na maioria das vezes uma prossecução do que foi efetuado no ano letivo anterior. Para que estas decisões pudessem ser "cirurgicamente" aplicadas seria necessário à direção escolar saber de antemão a que alunos se poderiam aplicar horas do crédito horário em apoios ou clubes que pudessem aumentar o seu sucesso escolar. A decisão da distribuição de crédito horário tendo em vista o sucesso escolar é difícil de realizar no início do ano letivo, ela seria bem mais simples de aplicar no final do ano letivo quando se observam as pautas de classificação, mas obviamente impossível de aplicar e sem qualquer efeito nessa altura.

As escolas estão hoje equipadas com ferramentas informáticas de gestão que foram desenvolvidas e aprimoradas ao longo de anos de desenvolvimento com o contributo e know-how dos diversos atores da comunidade escolar. Estas ferramentas são autênticos sistemas operativos da própria escola, entre os mais conhecidos destacam-se o INOVAR, JPM, SIGE e E360.

Todas estas plataformas geram um volume imenso de dados, que poderiam ser analisados e relacionados e auxiliar no processo de tomada de decisão da direção de uma escola, nas diversas vertentes de combate ao insucesso escolar.

A Inteligência Artificial (IA) na educação é um tópico inúmeras vezes abordado, mais na sua contribuição para a vertente pedagógica do que na vertente de tomada de decisão.

> "Os sistemas de IA estão atualmente a ajudar alguns educadores a identificar necessidades de aprendizagem específicas, proporcionando aos alunos experiências de aprendizagem personalizadas e ajudando algumas escolas a tomar melhores decisões, para que possam utilizar mais eficazmente os recursos

pedagógicos de que dispõem.” (European Commission, Directorate-General for Education, Youth, Sport and Culture, 2022)

Segundo a União Europeia (2022) “Os dados dos sistemas de informação dos alunos podem também ser utilizados no planeamento de recursos e cursos, bem como para prever o abandono escolar e as necessidades de orientação.” Assim, nas reflexões sobre este tema surgiu a questão, poderá a inteligência artificial contribuir para fundamentar a tomada de decisão da direção de uma escola tendo em vista melhorar o sucesso escolar dos seus alunos?

Na sequência da reflexão para o desenvolvimento desta investigação coloca-se a seguinte questão de investigação: “Os dados digitais gerados pelos Sistemas de Gestão Escolar de uma escola podem prever o insucesso escolar?”

Para tentar dar resposta a esta questão definiram-se como objetivos da investigação os seguintes:

- Avaliar a qualidade da informação digital para previsão do insucesso escolar
- Analisar a correlação entre certas variáveis e o insucesso escolar;
- Criar e avaliar o modelo preditivo de insucesso escolar

REVISÃO DA LITERATURA

BIG DATA

A nossa vida quotidiana vai alimentado diversas bases de dados amplamente usadas e dispersas em rede e online o *Big Data* surge como consequência disso. Tome-se como exemplo uma ida ao hipermercado, no processo de compra o cliente usa o cartão de fidelidade para aceder aos descontos e os itens da sua compra ficam registados numa base de dados. Estes dados poderão ser fornecidos a parceiros do hipermercado e vice-versa, criando aqui uma rede de informação sobre o cliente que eficientemente é utilizada por um sistema informático (muitas vezes recorrendo a **sistemas de inteligência artificial**) para tentar vender artigos direcionados ao perfil de cada cliente. Isto é replicado por quase todos os campos da nossa vida moderna, tal como refere Hershkovitz & Alexandron (2020). Neste contexto, caracterizado por uma vasta quantidade e diversidade de fontes de dados, emerge uma nova área de estudo, conhecida como *Big Data*. Esta disciplina, embora não se restrinja apenas a questões tecnológicas, tem sido impulsionada pelos avanços nos projetos tecnológicos.

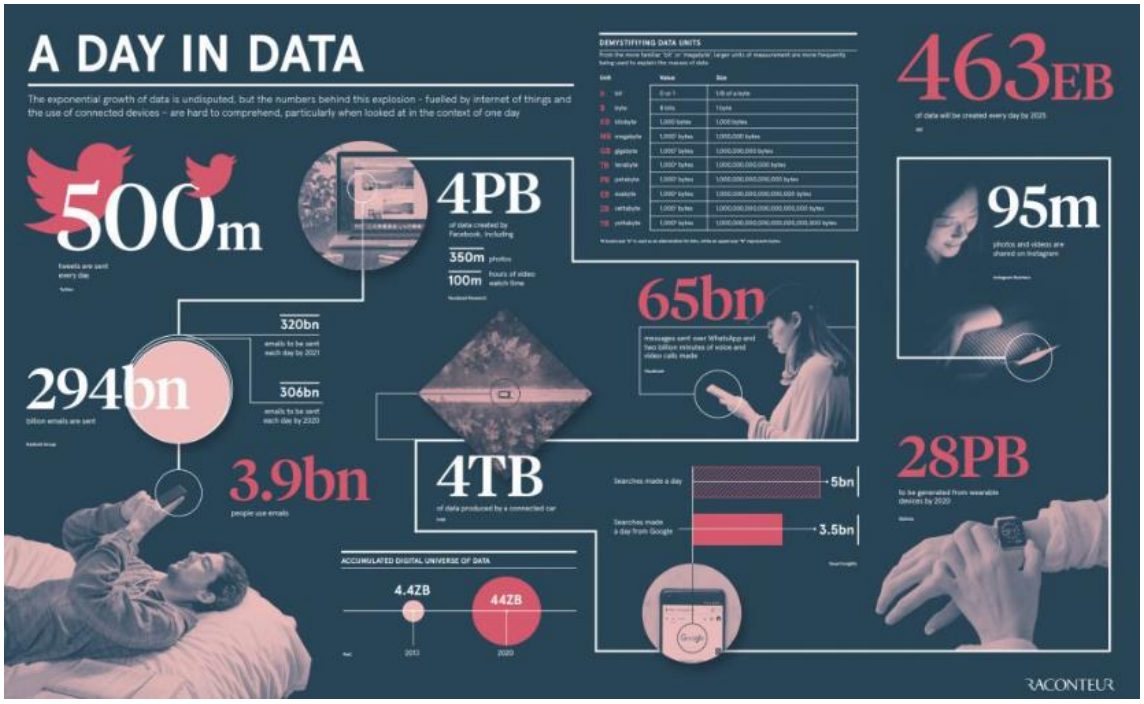


Figura 1- Volume de dados diários registado no uso de diversas plataformas

(Fonte: https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world/ consultado em 02-03-2024)

De acordo com Hershkovitz & Alexandron (2020) o *Big Data* obedece a cinco características conhecidas como os cinco V's, o volume, a velocidade, a variedade, a veracidade/precisão e o valor.

Com o avanço da tecnologia, o quotidiano inundou-se de uma vasta quantidade de dados e informações, segundo Ribeiro (2014, p. 98), o aumento no volume de dados e informações é ilustrado pelo uso crescente de dispositivos móveis, sensores industriais e biomédicos, fotos, vídeos, e-mails, redes sociais, comércio eletrónico, entre outros. Este crescimento é alimentado pela crescente presença de tecnologias como câmaras de vigilância, medidores inteligentes, GPS, aplicações de mensagens, e várias outras que facilitam a mobilidade urbana.

O uso desses dados de utilizadores, replicados por milhares ou até milhões de pessoas resultam num **volume** de informação desmesurado que só poderá ser analisado com recurso a *software* específico e hardware com grande poder de processamento.

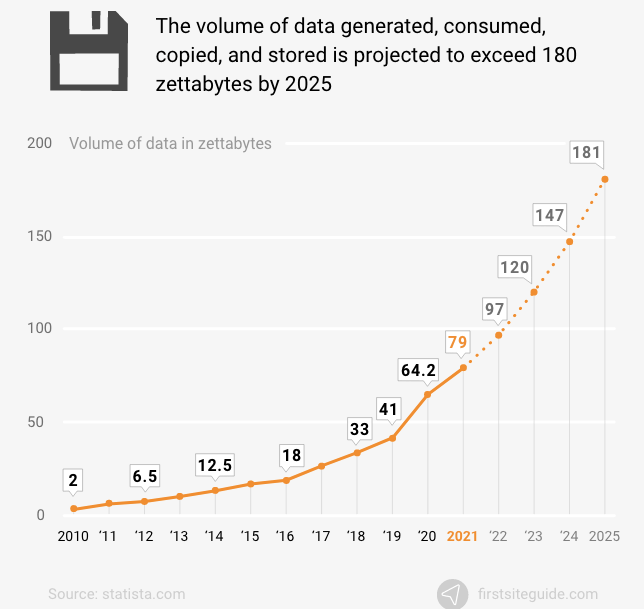


Figura 2 - Volume de dados gerados mundialmente em zettabytes

Com o avanço das tecnologias de transmissão de informação, como redes de fibra ótica, 5G e comunicações em tempo real para controlo de processos observa-se uma significativa melhoria na **velocidade** de troca de dados e informações (Pereira, 2014). De acordo com Florissi (2012) citado por Pereira(2014), esse aumento na velocidade tende a

continuar, uma vez que o desenvolvimento da tecnologia de processadores e dos hardwares de armazenamento, como discos rígidos e memórias flash, duplica seu poder a cada período de dois anos.

Outro ponto a considerar é a **variedade** de dados disponíveis. A abundância de informações provenientes de diversos meios resulta numa sobrecarga de dados e informação disponíveis para a sociedade (Ribeiro, 2014, p. 97).

Os registos poderão ser de diversos tipos, só assim se consegue correlacionar registos que muitas vezes se julgavam não estar relacionados, no entanto uma grande variedade de variáveis poderão trazer problemas estatísticos. Este problema é conhecido na ciência de dados como a maldição da dimensionalidade ("curse of dimensionality") tal como Awan (2023) defende "À medida que adicionamos mais dimensões ao nosso conjunto de dados, o volume do espaço aumenta exponencialmente. Isso significa que os dados tornam-se esparsos." Trazendo problemas como:

- Aumento da computação, mais dados significam maiores recursos computacionais para processamento desses dados;
- Sobreajuste, tornando-se as dimensões maiores necessariamente os modelos tornam-se mais complexos ajustando-se ao ruído em vez do padrão subadjacente
- Perda de significado das distâncias: Em espaços de alta dimensão, a diferença nas distâncias entre os pontos de dados tende a diminuir, tornando as medidas de distância menos discriminativas. Isso pode afetar algoritmos que dependem fortemente de medidas de proximidade.
- Desafios de visualização: Visualizar dados em dimensões elevadas é difícil, senão impossível. Como resultado, a análise exploratória dos dados torna-se mais desafiadora, pois não podemos contar com técnicas visuais para entender a estrutura dos dados.

A quarta caraterística é a **veracidade** ou **precisão**, ou seja, até que ponto os dados recolhidos representam de maneira confiável o que foram projetados para medir.

Um aspeto crucial a considerar é o valor dos dados reunidos, é imperativo que os quatro V's anteriores, volume, variedade, velocidade e veracidade possam gerar valor. O foco do *Big Data* deve ser direcionado para a criação de valor, ou seja, a grande quantidade de

dados deve ser analisada e convertida em informações aplicáveis a tendências, previsões que auxiliam a tomada de decisão (Pereira, 2014), através da análise preditiva.

DATA MINING

A mineração de dados (*Data Mining*) permite identificar insights valiosos e informações ocultas nos dados.

> "é o processo de explorar uma grande coleção de dados em busca de padrões consistentes, ou relações entre uma certa quantidade de dados. Uma vez identificadas as relações, elas precisam ser validadas de acordo com os parâmetros das buscas. O objetivo principal do *Data Mining* é criar novos conjuntos de dados, de modo a identificar novas tendências, comportamentos de massa, buscas massivas em um determinado período e assuntos relacionados a elas visualizados pelos consumidores" (O que é Data Mining? – Tecnoblog, s.d).

O *Data Mining* é uma técnica inserida em uma categoria de ferramentas analíticas que examinam grandes volumes de dados em busca de padrões ou agrupamentos subjacentes. Seu objetivo é identificar implicitamente tendências ou relações significativas entre os dados, fornecendo insights valiosos para tomada de decisões. (Pereira, 2014)

As técnicas de *Data Mining* não se limitam a interpretar os dados armazenados; o seu objetivo é extrair conclusões a partir de correlações nas informações não explícitas. Essas técnicas são projetadas para lidar com grandes volumes de dados, com o propósito de descobrir padrões úteis e recentes que poderiam passar despercebidos. (Pereira, 2014)

O *Data Mining* revela conhecimento implícito ou informações preditivas presentes em um *Data Warehouse* (repositório central de informação) ou em outros tipos de bases de dados, sem a necessidade de consultas específicas ou solicitações direcionadas. Esse processo emprega técnicas avançadas, como Redes Neurais, que têm a capacidade de aprender com o ambiente e aprimorar seu desempenho ao longo do tempo, além de técnicas heurísticas para resolver problemas quando a solução não é conhecida previamente e deteção de desvios por meio de regras (Júnior, 2010).

Na figura 3 podemos ver um grafismo de comparação entre *BigData* e *DataMining*.

Figura 3 - Data Mining vs Big Data

ANÁLISE PREDITIVA

A análise preditiva concentra-se principalmente em prever eventos futuros com base em dados históricos e padrões identificados, com o objetivo de orientar a tomada de decisão atual de forma informada e estratégica. Para operacionalizar este processo, são utilizadas diversas ferramentas ou algoritmos, que permitem compreender os dados existentes através da análise de grandes volumes de dados e identificar padrões, tendências e relações ocultas e dessa forma gerar regras de previsão. Santos, H. G. (2018)

A esta coleção de técnicas e algoritmos é conhecida como **modelos preditivos**, a sua aplicação pode ser amplamente diversificada, abrangendo áreas que vão desde finanças e marketing, saúde, comércio entre outros.

Importa referir que o trabalho com análise de dados deu origem a um novo perfil profissional, conhecido como Cientista de Dados (*Data Scientist*), "A característica principal deste profissional é ter a capacidade de aplicar ferramentas analíticas e algoritmos para gerar previsões sobre produtos, serviços, e comportamento de indivíduos" (Pereira,2014) segundo o mesmo autor "este perfil deve ter forte conhecimento em disciplinas como a matemática e a estatística, com treino avançado em estratégias para tratamento de grandes conjuntos de dados, fazendo uso de modelos matemáticos, formulação de hipóteses e técnicas de regressão."

O cientista de dados conduz o seu trabalho seguindo um fluxo de trabalho que geralmente passa por várias etapas que orientam todo o processo a ser desenvolvido.

> "O fluxo de trabalho em ciência de dados define as fases (ou passos) num projeto de ciência de dados. Utilizar um fluxo de trabalho em ciência de dados bem definido é útil pois fornece uma forma simples de lembrar a todos os membros da equipa de ciência de dados do trabalho a ser feito num projeto de ciência de dados." (What is a Data Science Workflow?, s.d.).

As etapas deverão ser bem definidas, desde a formulação do problema até à comunicação dos resultados e *insights* alcançados. Este processo meticuloso e estruturado é fundamental para garantir que os objetivos do projeto sejam claramente compreendidos e que as soluções propostas sejam eficazes e relevantes. Ver figura 4.

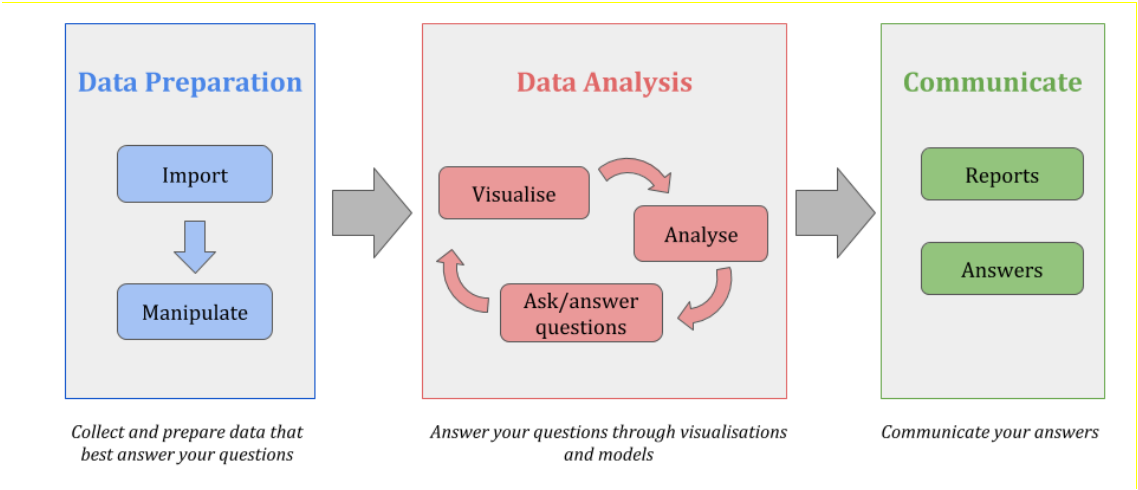


Figura 4 – Exemplo de Workflow para análise de dados

De acordo com Oliveira (2013) e Tavares (2014), o processo começa com a **preparação dos dados** usando técnicas estatísticas para separar e combinar conjuntos. Além disso, conforme os autores, também é possível utilizar técnicas de categorização, limpeza e transformação dos dados, incluindo a consideração da proveniência dos dados, para auxiliar no processo de categorização.

A transformação dos dados é conhecida como Engenharia de Recursos (***Feature Engineering*)** e envolve o processo de aplicar o conhecimento de domínio dos dados, criando recursos que ajudam algoritmos de *machine learning* a aprender melhor. Em geral, isso ocorre após a recolha e limpeza dos dados, e antes do treino dos modelos. (Alteryx, s.d.)

No final desta fase, é possível chegar à definição e preparação de que serão úteis na construção do grande conjunto de dados. (Pereira, 2014)

Após a extração dos dados, avança-se para a **fase de análise**, onde são aplicadas metodologias para avaliar a qualidade dos dados. A qualidade dos dados (*Data Quality)* refere-se à condição dos dados em termos de precisão, integridade, consistência, confiabilidade e atualização. "A qualidade dos dados é uma visão ou uma avaliação da adequação dos dados para servir a seu propósito em um determinado contexto". Mahanti (2019).

A qualidade de dados obedece a uma estrutura que define e avalia a qualidade dos dados, a maioria dos autores defende uma estrutura com seis dimensões: (As 6 Dimensões da Qualidade de Dados (*Data Quality*) - Data Science Academy, s.d.)

Segundo o mesmo autor as dimensões são as seguintes:

- Precisão: É a dimensão relativa ao que os dados representam, refletindo a realidade e sem erros.
- Completude: Esta dimensão indica se os dados necessários a determinado estudo estão presentes para que tal não afete as conclusões.
- Consistência: Refere-se à uniformidade dos dados tanto ao longo do tempo como à sua utilização em diferentes conjuntos de dados.
- Confiabilidade: Concerne à confiança da fonte de onde os dados são provenientes, fontes de maior confiança tendem a fornecer dados mais precisos e válidos.
- Relevância: Os dados devem ser relevantes para os objetivos da análise.

Para garantir a qualidade dos dados, os cientistas de dados e profissionais envolvidos na análise de dados geralmente realizam-se atividades de **limpeza e preparação de dados**. Isso envolve a identificação e correção de erros, a remoção de dados duplicados ou irrelevantes, e a verificação da consistência dos dados.

A garantia da qualidade dos dados é uma etapa fundamental em qualquer projeto de ciência de dados, uma vez que a precisão e confiabilidade dos resultados dependem diretamente da qualidade dos dados utilizados.

QUANTIDADE E QUALIDADE DOS DADOS

Sem uma **quantidade** substancial de dados, não é possível realizar Análise Preditiva de forma eficaz. Para que os modelos preditivos possam aprender e generalizar adequadamente, é essencial contar com milhares de registos. Caso não hajam dados suficientes para o treino, um modelo pode não ser capaz de aprender. Isso significa que ele irá absorver apenas as informações presentes nos dados fornecidos durante o treino, tornando-se incapaz de aplicar esse conhecimento a novos conjuntos de dados e, consequentemente, incapaz de fazer previsões precisas.

Segundo Pereira (2014) outro ponto importante é a **qualidade** dos dados. A precisão e integridade dos dados têm um impacto direto na qualidade do modelo resultante.

Por último, na fase da "**apresentação**", os dados analisados poderão ser utilizados tanto por humanos como por sistemas computorizados ou de inteligência artificial.

> "Ser acessível por humanos, é necessário um sistema de mediação que apresente os dados de forma amigável. Isso geralmente é feito com a ajuda de um painel educacional, que torna os dados sobre os alunos acessíveis a várias partes interessadas (por exemplo, alunos, professores, gestão escolar, etc.) de maneira eficaz." (Verbert et al., 2013)

As pessoas que precisam tomar decisões com base em dados devem, é claro, entender os dados apresentados, bem como seu uso. Uma das formas de apresentar resultados é através de ***dashboards***. Um *dashboard* consiste num painel visual que exibe informações, métricas e indicadores relevantes, com o objetivo de mostrar dados cruciais para a estratégia e tomada de decisão e para o cumprimento dos objetivos organizacionais. Este recurso gráfico possibilita o acompanhamento visual de dados. Na prática, o *dashboard* orienta a atenção para os dados, que atuam como verdadeiros indicadores do sucesso ou fracasso na performance. (Patel, s.d.)

Figura 5 - Exemplo de dashboard INOVAR

No que concerne à apresentação/tratamento de dados por sistemas computorizados, depois de recolher um conjunto de dados e realizar a seleção de atributos, um modelo preditivo pode ser construído a partir dos dados e processado através de um algoritmo. Nos termos mais gerais, o objetivo de um modelo preditivo é fazer uma previsão, dada alguma informação relacionada conhecida. (Brooks, 2017)

Tal como defendido por diversos autores (Hastie et al., 2008) e também por (Domingos, 2012) para o sucesso dos modelos preditivos, a divisão dos dados para treinar e validar o modelo é de extrema importância. O treino do modelo permite que este "aprenda" os padrões nos dados e se ajuste adequadamente às características observadas. Para a validação do modelo ambos os autores, salientam que essa etapa é crucial para verificar como o modelo se comporta em dados não vistos durante o treino, evitando possíveis problemas de um modelo excessivamente ajustado aos dados de treino e avaliando sua capacidade de generalização.

Tal como também referido por Saha e Raykar (2015) é amplamente conhecido que o desempenho do modelo nos dados utilizados para treinar o modelo (conjunto de treino) é uma estimativa excessivamente otimista do desempenho nos dados não vistos. Assim é comum reservar uma parte dos dados para avaliar o desempenho do modelo.

> "Quando estamos numa situação de dados abundantes, é dividir os dados em três partes: conjunto de treino, conjunto de validação e conjunto de teste.

> O conjunto de treino é utilizado para ajustar o modelo, ou seja, estimar os parâmetros do modelo. O conjunto de validação é utilizado para seleção do modelo, isto é, utilizamos o desempenho do modelo no conjunto de validação para selecionar entre vários modelos concorrentes" (Raykar & Saha, 2015, p. 3)

Neste estudo foi decidido reservar um terço dos dados para teste do modelo.

ERROS MAIS COMUNS NA ANÁLISE PREDITIVA

Segundo Wolf (2003), são quatro os motivos mais comuns da falha em análises preditivas: em primeiro lugar, quando as bases de dados não possuem qualidade suficiente para servir de suporte aos modelos; em segundo lugar, quando os modelos não são desenvolvidos adequadamente ou apresentam falhas; em terceiro lugar, quando elementos ou variáveis que podem influenciar o modelo não são considerados na sua construção; e por fim, quando não há supervisão do funcionamento da análise preditiva e atualização dos modelos para se adaptarem a mudanças ao longo do tempo.

A CIÊNCIA DOS DADOS NA EDUCAÇÃO

A CIÊNCIA DE DADOS COMO FERRAMENTA NA GESTÃO ESCOLAR

Considerados como o "novo ouro", os dados têm o potencial de enriquecer empresas e organizações. Neste sentido, a ciência de dados surge como uma ferramenta valiosa, mostrando que as informações disponíveis podem ser exploradas para obter *insights* relevantes, adaptando-se tanto ao contexto escolar como universitário. (Wolff, 2023)

Assim, a utilização inteligente dos dados pode desempenhar um papel importante na gestão e tomada de decisões nas instituições de ensino. Através da análise dos dados disponíveis, é possível identificar padrões, antecipar necessidades e ajustar estratégias para melhorar a eficiência e o desempenho académico, contribuindo para o sucesso dos alunos e para o desenvolvimento institucional.

Os gestores escolares, tal como os gestores de qualquer tipo de negócio, dispõem atualmente de uma vasta quantidade de dados e recursos tecnológicos avançados para monitorar, avaliar e analisar o progresso das suas instituições. No entanto, grande parte destes recursos ainda não é totalmente integrada nas rotinas de tomada de decisão, não potenciando eficazmente as escolhas dos decisores. Muitas decisões continuam a ser tomadas com base na intuição e na negociação, enquanto os dados são frequentemente utilizados como um último recurso, mais como uma ferramenta de argumentação do que como um suporte efetivo à decisão. (Wolff, 2023)

Os gestores escolares tendem a possuir características analíticas, no entanto, enfrentam dificuldades em transmitir essas habilidades a outros intervenientes no processo decisório, resultando muitas vezes numa gestão centralizada ou com silos de informação isolados. A adoção de uma cultura analítica requer uma mudança de hábitos e rotinas, onde os fluxos de informação assumem um papel central na transformação do próprio negócio. Estes recursos promovem uma tomada de decisão fundamentada em evidências e fomentam uma gestão descentralizada, contribuindo assim para uma maior eficiência e eficácia na gestão escolar.

A ciência de dados poderá ser usada para identificar padrões de abandono escolar ou para avaliar a eficácia de um currículo; uma vez que a recolha e análise de dados geralmente é feita em níveis mais altos (principalmente devido à infraestrutura, orçamento e questões relacionadas ao conhecimento profissional), essa questão é de grande importância para os formuladores de políticas educacionais (European Commission et al., 2016).

Diversos estudos demostram a utilidade de modelos preditivos no abandono escolar é o caso de (Sivakumar, 2016), ou de (Brooks et al., 2015).

A CIÊNCIA DE DADOS COMO FERRAMENTA NO PROCESSO APRENDIZAGEM

Segundo Cunha (2023), o surgimento de recursos tecnológicos na educação está em ascensão, ampliando significativamente as possibilidades de aproveitamento dessas tecnologias para beneficiar o processo de aprendizagem. A gestão da aprendizagem direcionada por dados educacionais tem ganhado crescente relevância nos últimos anos (KOVANOVIĆ et al., 2019). Esta prática consiste na recolha e análise de dados educacionais para compreender o desempenho dos alunos e identificar oportunidades de melhoria no processo de ensino.

A gestão da aprendizagem terá de adaptar-se a uma bordagem onde os dados digitais terão um papel importante.

> "orientada por dados educacionais, é fundamental porque coloca o estudante no centro do processo, tornando-o responsável pelo seu próprio progresso e desenvolvimento. Isso aumenta a motivação e o engajamento dos alunos, além de melhorar a qualidade do ensino." (Cunha,2023)

Neste contexto surge uma nova área de pesquisa conhecida como *Educational Data Mining* (EDM).

Segundo Cunha (2023), a EDM envolve a recolha de dados tanto em ambientes de aprendizagem formais como não formais, a partir de diversas fontes, com o objetivo de processar, organizar, interpretar e analisar esses dados, visando gerar novos conhecimentos e descobertas. Este processo estabelece padrões e métricas a serem alcançados. A fonte de dados que alimenta a EDM é obtida através das mais diversas plataformas de ensino, no entanto o mesmo autor citando BAKER (2011) refere que "Um grande problema, no entanto, é que com tantas fontes de dados diferentes em EDM, existe uma falta de padronização na maneira como os dados são recolhidos e armazenados, que, por si só, constitui um dos desafios da área"

A ciência de dados poderá estudar problemas tradicionais do processo de aprendizagem com métodos avançados com recurso a *software*s educacionais inteligentes em que "a arquitetura do sistema desses *software*s possui algoritmos para identificar padrões de comportamento dos estudantes, o que permite realizar intervenções precoces" (Cunha,

2023, p.11). Por exemplo, poderão ser construídos modelos matemáticos que descrevem com bastante precisão o progresso dos alunos em uma determinada questão e sugerem formas que os professores podem usar para auxiliar cada aluno. Esses modelos podem ser implementados através de diferentes tecnologias, incluindo *dashboards* e inteligência artificial.

Os *dashboards* podem ser utilizados para apresentar visualmente os dados relevantes aos professores, "ele é projetado para fornecer uma visão geral dos dados de aprendizagem e ajudar os professores a tomar decisões e a realizar interferências no processo de ensino." (Cunha, 2023. p.12).

Na perspetiva do aluno, trata-se de uma ferramenta tecnológica que possibilita visualizar o seu próprio desempenho, permitindo comparações que levam a reflexões para análise e tomada de decisão em colaboração com o professor. Este recurso pode estimular a perceção da necessidade de alterar comportamentos de estudo.

Na perspetiva do professor poderá auxiliar na identificação dos alunos que necessitam de maior apoio em áreas específicas, permitindo adaptar o conteúdo das aulas às necessidades individuais de cada estudante.

No que concerne à inteligência artificial entre diversas aplicações é vista como uma oportunidade no setor da Educação como "a possibilidade de gerar percursos de aprendizagem individualizados, a garantia de acesso universal para todos os estudantes, a automatização de tarefas administrativas ou a possibilidade de garantir um serviço de mentoria fora da sala de aula e em todos os momentos." (O ano da Inteligência Artificial. Qual o impacto na Educação?, 2023)

Ambas as abordagens têm o objetivo de fornecer aos educadores *insights* valiosos e recomendações acionáveis para promover o sucesso dos alunos de forma mais eficaz e personalizada. Através de várias fontes de dados, como material, sistemas de gestão escolar, ou sistemas em que os alunos relatam como se sentem enquanto aprendem, poderá prever-se quais dos alunos que terão dificuldades e intervir a tempo.

DESAFIOS PARA A CIÊNCIA DE DADOS NA EDUCAÇÃO

Ambientes de aprendizagem eficazes combinam corretamente os materiais de aprendizagem com o conhecimento pedagógico e o ambiente físico. Para que a ciência de dados seja utilizada, devem os dados ser registados massivamente; ou seja, ambientes de aprendizagem digital precisam ser usados com mais frequência. Isso, por sua vez, exige repensar os ambientes pedagógicos e físicos. Assim, os professores terão de adquirir novas competências, incluindo a capacidade de interpretar representações de dados e, posteriormente, tomar medidas de promoção da aprendizagem. (Hershkovitz & Alexandron, 2019, p.8)

Se a intenção é que a ciência de dados seja uma ferramenta de auxílio no processo ensino aprendizagem implica que os sistemas de aprendizagem baseados em computador devam ser usados com maior frequência para que o registo de dados possa incentivar a adaptação de sistemas de aprendizagem e a reciclagem de professores no uso desses sistemas. (Collins & Halverson, 2018).

Outro desafio reside no que acontece nos bastidores dos algoritmos baseados em dados. Por exemplo, para que um algoritmo possa ajudar os alunos de um curso online de álgebra, é necessário que ele compreenda a estrutura do conhecimento de álgebra, de forma a corresponder perguntas com essa estrutura e a identificar os tópicos que podem auxiliar na progressão para outros temas. Esta é a única maneira pela qual o sistema, ao identificar os pontos fracos e fortes dos alunos em questões específicas, poderá traçar o percurso ideal para cada aluno. Este desafio está intimamente ligado à Inteligência Artificial, que lida com a modelação e representação do conhecimento, bem como com o desenvolvimento de processos de tomada de decisão "inteligentes". (Hershkovitz & Alexandron, 2019, p. 13-14)

Para além da infraestrutura, a tecnologia essencial para impulsionar a revolução da ciência de dados na educação, o uso dos dados traz consigo os seus próprios obstáculos. Em particular, a ausência de padronização é um dos principais desafios que torna este processo bastante complexo.

Naturalmente, estes desafios acarretam sérias implicações económicas, uma vez que exigem significativos investimentos financeiros. Considerando que a maior parte do

investimento em educação e aprendizagem é proveniente de fontes públicas e que a estrutura de custos muitas vezes é inflexível, encontrar modelos económicos viáveis para introdução da ciência de dados em educação torna-se um desafio significativo.

Finalmente, o uso de ciência de dados está associado a questões éticas significativas relacionadas ao uso de dados, sobretudo com a aplicação do regulamento geral sobre a proteção de dados (RGPD). Em primeiro lugar, os alunos das escolas são menores, o que é relevante quando consideramos a sua capacidade para consentir com a recolha dos seus dados. Em segundo lugar, os sistemas de aprendizagem são frequentemente utilizados como parte dos currículos obrigatórios, levantando a questão do que fazer quando os alunos se recusam a permitir a recolha dos seus dados. Isso requer um conjunto de regras claras e aplicáveis, apoiadas por tecnologia e metodologia adequadas, para garantir que os dados recolhidos sejam utilizados de forma apropriada e exclusivamente para melhorar a educação e a aprendizagem. (Hershkovitz & Alexandron, 2019, p. 13-14)

Esta temática será desenvolvida no capítulo seguinte.

CONSIDERAÇÕES ÉTICAS DO USO DE DADOS NA EDUCAÇÃO

No dia 25 de outubro de 2022, a Comissão Europeia divulgou as Orientações Éticas destinadas a professores sobre a Utilização de Inteligência Artificial (IA) e dados no contexto do ensino e da aprendizagem.

As Diretrizes Éticas foram elaboradas por um Grupo de Especialistas da Comissão como parte integrante do Plano de Ação para a Educação Digital (2021-2027).

Conforme os sistemas de IA se desenvolvem o uso de dados torna-se mais disseminado, é crucial aprofundar a compreensão do impacto deles na educação. O rápido crescimento de sua utilização requer, uma exigência mais profunda na utilização da IA e consequentemente no uso de dados, mas também habilitação para os usar de forma positiva, crítica e ética.

> "É evidente que temos de assegurar que os professores e educadores compreendem o potencial da IA e dos megadados no domínio da educação, estando simultaneamente conscientes dos riscos associados." (European Commission, Gabriel, Marya, 2022, p. 6)

A recolha e análise de dados sobre os alunos pela escola para um melhor planeamento na afetação de recursos ou nos procedimentos como criar turmas, horários e identificação dos alunos que necessitam de apoios complementares à sua aprendizagem, poderá ser uma tarefa delegada a um sistema de IA. No entanto há que considerar que estes sistemas não são infalíveis, há que considerar o risco de erros de análise na sua utilização, até porque na educação existem diversos fatores, muitas vezes externos à escola, que podem influenciar o sucesso ou insucesso escolar.

> "A utilização da IA na educação é tal como a génese da própria IA um processo de aprendizagem contínua tal como defende o grupo de peritos da EU "As orientações éticas sobre a utilização de IA e de dados no ensino e na aprendizagem são um processo gradual de deliberação e aprendizagem

> contínuas.” (European Commission, Directorate-General for Education, Youth, Sport and Culture, 2022, p. 17)

Um dos princípios éticos envolve a salvaguarda da privacidade e proteção dos dados, garantindo a preservação da identidade do utilizador face às informações de identificação pessoal.

A segurança dos dados é crucial para garantir a privacidade, requerendo medidas robustas para prevenir acesso não autorizado ou violações de dados. Restrições de acesso são fundamentais, limitando quem pode visualizar e utilizar os dados.

Manter o anonimato do utilizador é outra preocupação crucial, especialmente em conjuntos de dados extensos nos quais as pessoas ainda podem ser identificadas devido a combinações únicas de características. A capacidade de tornar um utilizador não identificável em conjuntos de dados anonimizados, frequentemente através de técnicas como mascaramento de dados ou pseudonimização, é uma parte fundamental da segurança dos dados.

DADOS COM POTENCIAL DE CORRELAÇÃO DE PREVISÃO DO INSUCESSO ESCOLAR

CONTEXTO SOCIOECONÓMICO

Uma maior qualidade do ensino traduz-se mais tarde num bom desempenho académico e consequentemente num bom desempenho profissional, pelo que a preocupação passa agora por garantir a maior qualidade. (Silva, 2019)

O **contexto socioeconómico do agregado familiar** é certamente um denominador comum do desempenho escolar, pais com maior **nível de escolaridade** ou com rendimentos mais elevados oferecem uma quantidade e variedade de recursos ao seu dispor, provendo o seu desempenho cognitivo, tal como defende Carneiro "*families with better educated fathers provide better home and school environments for their children*" (Carneiro, 2008). A envolvência do agregado familiar na vida escolar é um fator determinante do sucesso escolar, tal como indica o PISA 2022:

> "Em todos os países/economias com dados disponíveis, os alunos que beneficiaram de um maior apoio das suas famílias referiram um maior sentimento de pertença à escola e de satisfação com a vida, bem como uma maior confiança na sua capacidade de aprendizagem autónoma." (PISA, 2022, p.93)

Os estudos de Carneiro (2008) e também de Silva (2019) apontam que pais de maior escolaridade ou maior rendimento não procuram escolas com maior número de recursos mas sim matriculam seus filhos em escolas com colegas semelhantes (crianças de escolas com alto nível de escolaridade) e que o seu desempenho é em parte determinado pelas características individuais e familiares.

NÍVEL DE ESCOLARIDADE DOS PAIS

Estando comprovado que a influência do contexto familiar é um fator determinante para o sucesso escolar, importa referir que as diferenças económicas de cada agregado familiar

têm diferentes ponderações nesse sucesso. Azevedo (2011) defende que o desempenho está diretamente relacionado com o **nível escolar dos elementos da família**, famílias com maior nível escolar têm educandos com maior sucesso.

Silva (2019) cita no seu estudo diversos autores que chegaram a esta mesma conclusão:

> "Carneiro (2008) procurou aplicar a metodologia do relatório Coleman ao contexto português, a partir dos dados do PISA de 2000. Mais uma vez as conclusões vão no mesmo sentido: o contexto familiar é fundamental e decisivo para o desempenho escolar do aluno." (Silva, 2019)

Esta correlação contexto familiar/desempenho escolar está provavelmente relacionado com as fases iniciais do desenvolvimento do cérebro humano, segundo Mustard (2010), afetam os estágios de desenvolvimento posteriores, assim o contexto familiar, onde está inserida a criança, é de extrema importância para o desenvolvimento de competências/habilidades que serão a base de desenvolvimento dos estágios de desenvolvimento cognitivo posteriores. Um maior estímulo no desenvolvimento das capacidades da criança em contexto familiar trará num futuro melhores resultados no seu sucesso escolar. O contrário acontecerá em ambientes pouco favoráveis, com menor estímulo das capacidades essências, como são geralmente os contextos familiares de famílias com baixos rendimentos. (Carneiro, 2008).

O estímulo do desenvolvimento das capacidades da criança faz-se muitas vezes com a participação em diversas atividades, ora a capacidade de fornecer um maior número de atividades/estímulos está relacionada com a capacidade económica do agregado familiar, um agregado com maior robustez económica. Por norma agregados familiares com maior capacidade económica são compostos por pais com maior grau académicos, segundo Silva (2019) os alunos com pais com maior grau académico têm melhor sucesso escolar.

ESTRUTURA DO AGREGADO FAMILIAR

No que concerne ao contributo que a escola pode ter no desempenho escolar, não há um consenso na literatura, podemos encontrar autores que defendem que a escola tem um papel importante no desempenho escolar dos alunos como defende Pereira e Reis (2012), assim como podemos encontrar estudos que indicam a nulidade do valor acrescentado por parte das escolas tal como demonstra Sousa (2016).

Não desvalorizando o papel da escola, a maioria dos autores defende que é o contexto familiar que faz a diferença no sucesso escolar sendo esse fator defendido por diversos autores. Azevedo (2011) demonstra que o desempenho escolar está diretamente relacionado com a **estrutura do agregado familiar**, assim como Pereira, M (2010) que menciona que o contexto familiar é um dos fatores que gera maior desigualdade nos resultados dos alunos. O PISA 2022 vem mostrar o mesmo afirmando o relatório que tarefas simples como jantar com a família poderá ter implicância no desempenho escolar:

> *"*Em Portugal, os alunos que afirmaram jantar com os seus pais ou com alguém da família todos ou quase todos os dias, pontuaram, em média, 483 pontos a matemática, enquanto que os alunos que afirmaram fazê-lo apenas uma ou duas vezes por ano pontuaram 395 pontos a matemática*"* (PISA, 2022, p. 93)

O contexto familiar poderá também influenciar um outro fator determinante para o sucesso escolar, a assiduidade. Segundo a ONU crianças que vivam em lugares mais pobres têm três vezes menos probabilidade de ir às aulas do que crianças que vivam em locais mais favorecidos (ONU, 2013).

Silva, (2019) citando , Gennetian et al. (2018), defende que existe uma correlação positiva entre os níveis de rendimento familiar e a frequência escolar dos alunos em todos os níveis de escolaridade e refere ainda que no contexto familiar a **influência da mãe** é marcante no desempenho da criança:

> "a escolarização da mãe influencia de forma diferente o desempenho da criança, comparativamente à educação do pai. Existe evidência no sentido da educação da

> mãe desempenhar um papel muito importante para a educação de seus filhos, sendo este efeito superior ao da educação paterna" (Silva, 2019, p. 11)

O papel do pai tem um impacto mais tardio segundo Silva (2009) " o impacto que os pais têm sobre o desenvolvimento cognitivo dos seus filhos tende a aumentar à medida que as crianças crescem."

Existem numerosas investigações que têm abordado a questão do impacto da educação parental no rendimento escolar dos filhos. A literatura existente sugere, de forma geral, uma correlação positiva entre o **nível escolar dos pais** e o investimento no desenvolvimento do capital humano dos filhos, resultando em uma melhoria no seu desempenho académico, aspeto que é alvo de análise no presente estudo. Este efeito, acompanhado por outras influências a nível escolar, individual e familiar, destaca a importância de controlar diversas dimensões ao conduzir a investigação.

ACESSO A TECNOLOGIAS DE INFORMAÇÃO E COMUNICAÇÃO

Nos últimos anos, tem-se observado uma crescente integração de computadores nas escolas e nas salas de aula. Tal como refere Domingues (2017) "a literatura científica e pedagógica disponível sobre este tema, através de estudos conduzidos por autores como Ramos (2007) e Santos (2006), proporciona uma análise das potencialidades das tecnologias e dos diferentes programas utilizados para aprimorar a qualidade do processo de ensino-aprendizagem".

Para outros autores como Machin et al. (2006) citado por Domingues, (2017), o impacto da utilização das novas tecnologias é positivo, na medida em que estas facultam melhorias no aproveitamento dos alunos.

A Internet representa uma vasta base de dados e conhecimento para a educação, onde toda a informação está acessível com apenas um simples clique.

A utilização da Web como ferramenta pedagógica apresenta a vantagem de motivar os alunos para alcançarem a excelência, dinamizando o conteúdo das suas aprendizagens e promovendo a autonomia e a criatividade, elementos essenciais para a sua formação. Gonçalves (2012)

Embora a utilização da Web como ferramenta pedagógica tenha a vantagem de motivar os alunos para alcançarem a excelência, dinamizando o conteúdo das suas aprendizagens e promovendo a autonomia e a criatividade essenciais à sua formação, é importante ressaltar que não há aprendizagem sem organização e controlo.

Os alunos muitas vezes possuem limitadas capacidades e conhecimentos em metodologia de pesquisa, tornando essencial uma orientação cuidadosa por parte do professor. Encorajar uma pesquisa livre, sem qualquer orientação, numa aula, especialmente com alunos inexperientes, pode acarretar mais desvantagens do que benefícios.

Gonçalves (2012), citanto Carvalho (2006) refere que é fundamental orientar os alunos na avaliação da informação encontrada, auxiliando-os a identificar critérios que os guiem nesse processo. Gonçalves (2012)

METODOLOGIA

PERGUNTA DE PARTIDA E OBJETIVOS DE INVESTIGAÇÃO

Para iniciar a primeira etapa desta investigação torna-se necessário focar num problema, assim o foco da investigação será o "insucesso escolar". A questão de investigação proposta é: "Os dados digitais gerados pelos Sistemas de Gestão Escolar de uma escola podem prever o insucesso escolar?"

Os objetivos de investigação delineados são os seguintes:

1. Avaliar a qualidade dos dados provenientes do *software* de gestão escolar de uma determinada escola.
2. Analisar a correlação entre certas variáveis e o sucesso ou insucesso escolar dos alunos.
3. Desenvolver e avaliar um modelo de previsão do sucesso escolar com base nos dados disponíveis.

TIPO DE ESTUDO

Um estudo de caso é uma abordagem de pesquisa que se concentra em investigar um caso específico, geralmente num determinado contexto através da recolha e análise de dados que poderão ser qualitativos, como entrevistas e observações, ou quantitativas, como estatísticas e métricas. O objetivo principal de um estudo de caso é entender e identificar padrões emergentes e *insights* significativos.

DESCRIÇÃO DO CASO

O agrupamento de escolas em estudo é composto por diversas escolas que abrangem todos os ciclos de ensino desde o pré-escolar ao secundário. À data de elaboração deste estudo conta com 1835 alunos, repartidos da seguinte forma:

- Ensino secundário – 656
- Ensino profissional - 387
- 3º ciclo - 293

- 2º ciclo - 68
- 1º ciclo - 281,
- pré-escolar - 150

A informatização das escolas fez com que os registos deixassem os livros de pontos ou atas em arquivos mortos e passassem a ficar registados em disco rígidos alocados em servidores, muitas vezes distantes geograficamente da escola. Grande parte do trabalho do professor, diretor de turma, direção, e restantes elementos da orgânica de uma escola são a registar ou consultar registos nos *software*s de gestão escolar. Desta forma, visa-se compreender se estes dados possuem qualidade preditiva do insucesso escolar.

Para elaboração deste trabalho de análise de dados e com o objetivo de responder ao primeiro objetivo de investigação, adotou-se o *workflow* da figura 4, já identificado na introdução teórica, que se divide em três fases principais (*Data Preparation, Data Analysis e Communicate*). As três fases são comuns para os cientistas de dados.

FASE DO *DATA PREPARATION*

A fase de *Data Preparation* começa com o ***import*** (importação dos dados). Assim, torna-se pertinente identificar que dados importar.

A literatura menciona que as variáveis com potencial de correlação com o insucesso escolar são o **nível académico dos Encarregados de Educação** e dos elementos do seu **agregado familiar**. (Azevedo, 2011, citado por Silva, 2019) "Este desempenho tende a ser maior quanto maior for a percentagem de indivíduos na família com elevados níveis de educação." (Silva, 2019) Estes dados estão registados na plataforma de gestão escolar no formulário do processo escolar do aluno, como por exemplo: resultados académicos, assiduidade, acesso a tecnologia, profissão dos Encarregados de Educação e formação académicas dos pais.

Silva (2019) refere ainda que o **impacto da mãe** como Encarregada de Educação é marcante no desempenho da criança.

Embora não haja acesso direto ao *status* socioeconómico, a profissão dos pais poderá ser um indicador indireto, dada a sua forte associação com o *status* económico das famílias. Esta abordagem permite inferir, de forma indireta, informações relevantes para a análise.

O *software* que nos dá acesso às variáveis acima mencionadas é o *software* de gestão escolar INOVAR. O INOVAR é um dos *softwares* de gestão escolar mais conhecidos disponibiliza toda a informação e procedimento decorrente da atividade letiva como a caracterização da turma, planificações, sumários, faltas, registo de ocorrências, critérios de avaliação, notas, sínteses descritivas, pautas, análise estatística, sinalizações no âmbito do DL 54/2018, entre outros. A plataforma permite ainda manter pais e encarregados de educação a par da vida escolar dos seus educandos. A par da vertente letiva os módulos relacionados com as áreas administrativas e arquivísticas, contém matrículas e ofertas formativas, permite seriar candidaturas do Portal das Matrículas, gere fichas e registos biográficos de alunos, emite documentação certificadora, é veículo de transmissão de dados solicitados pela tutela, é por isso um verdadeiro sistema operativo de uma escola.

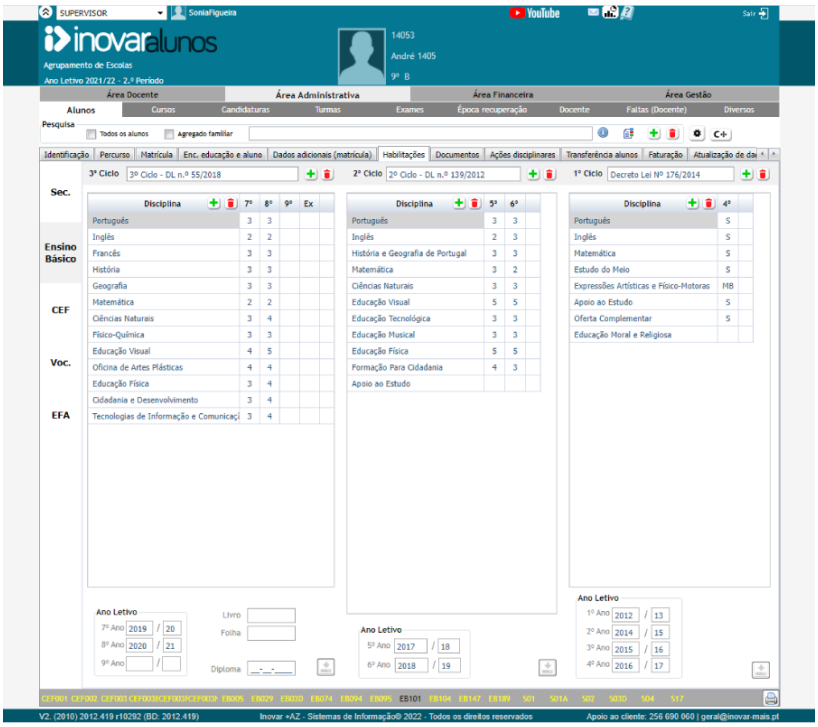


Figura 6 - Ambiente de trabalho do INOVAR

Para iniciar esta investigação, procedeu-se à exportação de múltiplos relatórios acessíveis no *software* INOVAR para ficheiros Excel. Posteriormente, unificou-se todos os dados numa folha Excel, utilizando o número de processo do aluno como elo entre os registos.

Processo	Naturalidade	Sexo	Tipo ident.	Validade	Nacionalidade	Freguesia naturalidade	Concelho naturalidade	Distrito naturalidade	Cód. postal 4	Cód. postal 3	Cód. postal localidade	Fregues
27811	Portugal	M	CC	20-11-2022	Portugal	A dos Francos	Caldas da Rainha	Leiria	2500	037	A DOS FRANCOS	A dos Francos
27406	Portugal	F	CC	15-02-2023	Portugal	A dos Francos	Caldas da Rainha	Leiria	2500	053	A DOS FRANCOS	A dos Francos
27382	Portugal	M	CC	28-02-2023	Portugal	A dos Francos	Caldas da Rainha	Leiria	2500	018	A DOS FRANCOS	A dos Francos
27380	Portugal	F	CC	26-07-2023	Portugal	União das freguesias de Caldas da Rainha - Nossa Senhora do Pópulo, Coto e São Gregório	Caldas da Rainha	Leiria	2500	051	A DOS FRANCOS	A dos Francos
26972	Portugal	F	CC	07-03-2022	Portugal	União das freguesias de Caldas da Rainha - Nossa Senhora do Pópulo, Coto e São Gregório	Caldas da Rainha	Leiria	2500	043	A DOS FRANCOS	A dos Francos
26954	Portugal	M	CC	24-10-2022	Portugal	A dos Francos	Caldas da Rainha	Leiria	2500	043	A DOS FRANCOS	A dos Francos
27384	Portugal	F	CC	25-01-2023	Portugal	União das freguesias de Caldas da Rainha - Nossa Senhora do Pópulo, Coto e São Gregório	Caldas da Rainha	Leiria	2500	028	A DOS FRANCOS	A dos Francos
27383	Portugal	F	Pass		Portugal	A dos Francos	Caldas da Rainha	Leiria	2500	014	A DOS FRANCOS	A dos Francos
27409	Portugal	F	CC	28-08-2023	Portugal	Alvorninha	Caldas da Rainha	Leiria	2500	368	ALVORNINHA	Alvorninha
27391	Portugal	M	CC	12-11-2023	Portugal	União das freguesias de Caldas da Rainha - Nossa Senhora do Pópulo, Coto e São Gregório	Caldas da Rainha	Leiria	2500	389	ALVORNINHA	Alvorninha
27410	Portugal	M	CC	13-11-2023	Portugal	União das freguesias de Caldas da Rainha - Nossa Senhora do Pópulo, Coto e São Gregório	Caldas da Rainha	Leiria	2500	348	ALVORNINHA	Alvorninha
27948	Portugal	F	CC	08-11-2024	Portugal	União das freguesias de Caldas da Rainha - Nossa Senhora do Pópulo, Coto e São Gregório	Caldas da Rainha	Leiria	2500	368	ALVORNINHA	Alvorninha
27877	Portugal	M	CC	11-01-2023	Portugal	União das freguesias de Caldas da Rainha - Nossa Senhora do Pópulo, Coto e São Gregório	Caldas da Rainha	Leiria	2500	796	SANTA CATARINA CLD	Santa Catarina
27797	Portugal	M	CC	22-06-2023	Portugal	Santa Catarina	Caldas da Rainha	Leiria	2500	791	SANTA CATARINA CLD	Santa Catarina
27834	Portugal	M	CC	16-07-2023	Portugal	Santa Catarina	Caldas da Rainha	Leiria				Santa Catarina

Figura 7 - Exemplo de relatório exportado do INOVAR

Após a exportação, tornou-se necessário consolidar os dados numa única folha de Excel, a fim de estabelecer a correlação entre as informações do registo biográfico do aluno e os dados académicos, como a **assiduidade**, os resultados académicos, entre outros.

No caso da assiduidade esperava-se que a assiduidade nas aulas e o tempo dedicado ao estudo tivessem um impacto positivo no sucesso escolar. Esses fatores estão intrinsecamente relacionados com o sucesso escolar. (Fernandes, 2015)

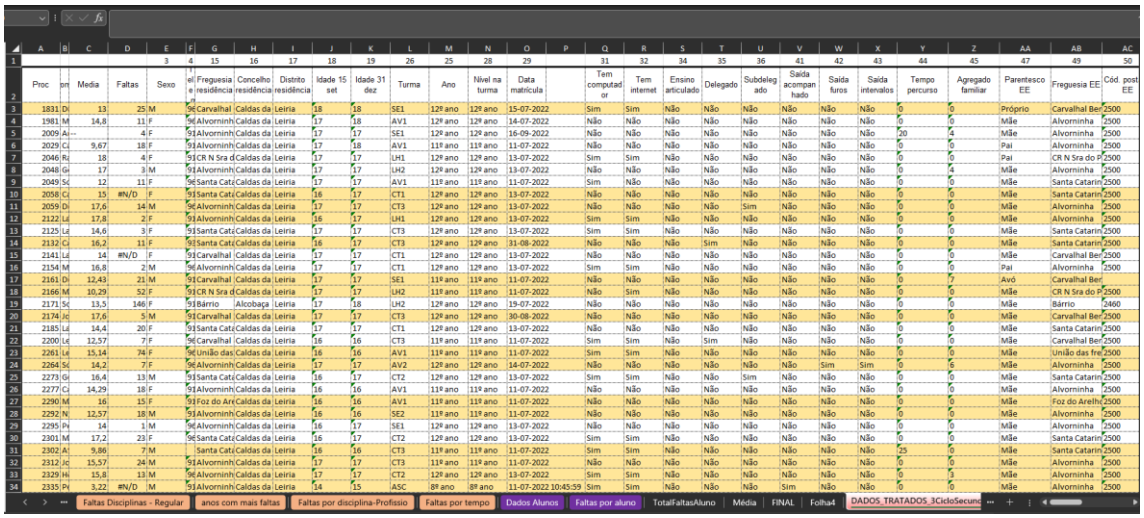
De seguida, na fase da *Data Preparation*, vem o passo da **manipulação dos dados**. Aqui foi necessário proceder à computação de alguns dados (*feature engineering*), isto é, selecionar, preparar ou transformar as variáveis de um conjunto de dados para melhorar o desempenho e a eficácia de modelos de aprendizagem automática. Neste caso foi necessário preparar a variável assiduidade total dos alunos, uma vez que essa informação não estava registada nos dados brutos (*raw*).

A plataforma de *software* permite o registo de várias categorias de faltas, incluindo faltas de presença, divididas em faltas injustificadas, justificadas e de pontualidade, bem como faltas disciplinares, faltas relacionadas com material e faltas de trabalho de casa. Apenas foram consideradas falta de assiduidade aquelas em que o aluno não esteve presente em sala de aula.

A literatura identifica como variáveis preditivas do insucesso escolar: a formação académica do Encarregado de Educação, o nível de escolaridade dos pais ou Encarregado de Educação e assiduidade, foram essas as variáveis tidas em conta na consolidação e relação dos dados na folha principal.

No entanto, visto que algumas variáveis, como a assiduidade, só ficam disponíveis após o primeiro período, este modelo seria utilizado para prever o insucesso com base no desempenho do aluno durante o primeiro período.

Por motivos de proteção de dados as variáveis relacionadas com a identificação do aluno, Encarregado de Educação e pais foram ocultadas tendo sido feita apenas relação com o número de processo do aluno.



Proc	Media	Faltas	Sexo	Freguesia residência	Concelho residência	Distrito residência	Idade 15 set	Idade 31 dez	Turma	Ano	Nível na turma	Data matrícula	Tem computador	Tem internet	Ensino articulado	Delegado	Subdelegado	Saída acompanhado	Saída furos	Saída intervalos	Tempo percurso	Agregado familiar	Parentesco EE	Freguesia EE	Cód. post EE
1831	13	25	M	Carvalhal	Caldas da	Leiria	18	18	SE1	12º ano	12º ano	15-07-2022	Sim	Sim	Não	Não	Não	Não	Não	Não	0	0	Próprio	Carvalhal Ber	2500
1981	14,8	11	F	Alvorninh	Caldas da	Leiria	17	18	AV1	12º ano	12º ano	14-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Alvorninha	2500
2009	A---	4	F	Alvorninh	Caldas da	Leiria	17	17	SE1	12º ano	12º ano	16-09-2022	Não	Não	Não	Não	Não	Não	Não	Não	20	4	Mãe	Alvorninha	2500
2029	9,67	18	F	Alvorninh	Caldas da	Leiria	17	18	AV1	11º ano	11º ano	11-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Pai	Alvorninha	2500
2046	18	4	F	CR N Sra d	Caldas da	Leiria	17	17	LH1	12º ano	12º ano	13-07-2022	Sim	Sim	Não	Não	Não	Não	Não	Não	0	0	Pai	CR N Sra do P	2500
2048	17	3	M	Alvorninh	Caldas da	Leiria	17	17	LH2	12º ano	12º ano	13-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	4	Mãe	Alvorninha	2500
2049	12	11	F	Santa Cat	Caldas da	Leiria	17	17	AV1	11º ano	11º ano	11-07-2022	Sim	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Santa Catarin	2500
2058	15	#N/D	F	Santa Cat	Caldas da	Leiria	16	17	CT1	12º ano	12º ano	13-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Santa Catarin	2500
2059	17,6	14	M	Alvorninh	Caldas da	Leiria	17	17	CT3	12º ano	12º ano	13-07-2022	Não	Não	Não	Não	Sim	Não	Não	Não	0	0	Mãe	Alvorninha	2500
2122	17,8	2	F	Alvorninh	Caldas da	Leiria	16	17	LH1	12º ano	12º ano	13-07-2022	Sim	Sim	Não	Não	Não	Não	Não	Não	0	0	Mãe	Alvorninha	2500
2125	14,6	3	F	Santa Cat	Caldas da	Leiria	17	17	CT3	12º ano	12º ano	13-07-2022	Sim	Sim	Não	Não	Não	Não	Não	Não	0	0	Mãe	Santa Catarin	2500
2132	16,2	11	F	Santa Cat	Caldas da	Leiria	16	17	CT3	12º ano	12º ano	31-08-2022	Não	Não	Sim	Não	Não	Não	Não	Não	0	0	Mãe	Santa Catarin	2500
2141	14	#N/D	F	Carvalhal	Caldas da	Leiria	17	17	CT1	12º ano	12º ano	13-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Carvalhal Ber	2500
2154	16,8	2	M	Alvorninh	Caldas da	Leiria	17	17	CT1	12º ano	12º ano	13-07-2022	Sim	Sim	Não	Não	Não	Não	Não	Não	0	0	Pai	Alvorninha	2500
2161	12,43	21	M	Carvalhal	Caldas da	Leiria	17	17	SE1	11º ano	11º ano	11-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	7	Avó	Carvalhal Ber	
2166	10,29	52	F	CR N Sra d	Caldas da	Leiria	17	17	LH2	11º ano	11º ano	11-07-2022	Não	Sim	Não	Não	Não	Não	Não	Não	0	0	Mãe	CR N Sra do P	2500
2171	13,5	146	F	Bárrio	Alcobaça	Leiria	17	18	LH2	12º ano	12º ano	19-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Bárrio	2460
2174	17,6	5	M	Carvalhal	Caldas da	Leiria	17	17	CT3	12º ano	12º ano	30-08-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Carvalhal Ber	2500
2185	14,4	20	F	Santa Cat	Caldas da	Leiria	17	17	CT1	12º ano	12º ano	13-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Santa Catarin	2500
2200	12,57	7	F	Carvalhal	Caldas da	Leiria	16	16	CT3	11º ano	11º ano	11-07-2022	Sim	Sim	Não	Sim	Não	Não	Não	Não	0	0	Mãe	Carvalhal Ber	2500
2261	15,14	74	F	União das	Caldas da	Leiria	16	16	AV1	11º ano	11º ano	11-07-2022	Sim	Sim	Não	Não	Não	Não	Não	Não	0	0	Mãe	União das fre	2500
2264	14,2	7	F	Alvorninh	Caldas da	Leiria	17	17	AV2	12º ano	12º ano	14-07-2022	Não	Não	Não	Não	Não	Não	Sim	Sim	0	6	Mãe	Alvorninha	2500
2273	16,4	13	M	Santa Cat	Caldas da	Leiria	16	17	CT2	12º ano	12º ano	13-07-2022	Sim	Sim	Não	Não	Sim	Não	Não	Não	0	0	Mãe	Santa Catarin	2500
2277	14,29	18	F	Alvorninh	Caldas da	Leiria	16	16	AV1	11º ano	11º ano	11-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Alvorninha	2500
2290	16	15	F	Foz do Arc	Caldas da	Leiria	16	16	AV1	11º ano	11º ano	11-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Foz do Arelho	2500
2292	12,57	18	M	Alvorninh	Caldas da	Leiria	16	16	SE2	11º ano	11º ano	11-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Alvorninha	2500
2295	14	1	M	Alvorninh	Caldas da	Leiria	16	17	SE1	12º ano	12º ano	13-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Alvorninha	2500
2301	17,2	23	F	Santa Cat	Caldas da	Leiria	16	17	CT2	12º ano	12º ano	13-07-2022	Sim	Sim	Não	Não	Não	Não	Não	Não	0	0	Mãe	Santa Catarin	2500
2302	9,86	7	M	Santa Cat	Caldas da	Leiria	16	16	CT3	11º ano	11º ano	11-07-2022	Sim	Sim	Não	Não	Não	Não	Não	Não	25	0	Mãe	Santa Catarin	2500
2312	15,57	24	M	Alvorninh	Caldas da	Leiria	17	17	CT3	11º ano	11º ano	11-07-2022	Não	Não	Não	Não	Não	Não	Não	Não	0	0	Mãe	Alvorninha	2500
2329	15,8	13	M	Alvorninh	Caldas da	Leiria	17	17	CT2	12º ano	12º ano	13-07-2022	Sim	Sim	Não	Não	Não	Não	Não	Não	0	3	Mãe	Alvorninha	2500
2335	3,22	#N/D	M	Alvorninh	Caldas da	Leiria	14	15	ASC	8º ano	8º ano	11-07-2022 10:45:59	Sim	Sim	Não	Não	Não	Sim	Não	Não	0	0	Mãe	Alvorninha	2500

Figura 8 – Dados dos diversos relatórios agregados

Dado o objetivo de analisar o valor preditivo da formação académica do agregado familiar no sucesso escolar do aluno, foi necessário criar uma variável derivada da formação académica de cada um dos pais, uma vez que essa informação, embora não constasse nos dados em bruto, era possível determinar. Isso implicou a realização da fase de *feature engineering*. Nos dados em estado "*raw*" (dados em bruto), o registo da formação académica dos pais era individual. Optou-se por atribuir a seguinte pontuação por nível de formação:

- Ensino Básico – 1 ponto
- Ensino Secundário – 10 pontos
- Ensino Superior – 30 pontos

Esta nova variável possibilitou a aplicação de técnicas de análise estatística simples no Excel, permitindo identificar o nível de formação académica do agregado familiar. Os agregados com pontuação abaixo de 10 pontos possuíam o ensino básico, os agregados com pontuação entre 10 e 30 pontos possuíam ensino secundário e os restantes agregados familiares ensino superior.

Exemplo:

- Pai com primeiro ciclo – 1 ponto
- Mãe com secundário – 10 pontos
- **Total:** 11 pontos – Agregado familiar com nível académico secundário

- Pai com Licenciatura – 30 pontos
- Mãe com primeiro ciclo – 1 pontos
- **Total:** 31 pontos - Agregado familiar com nível académico superior

Como resultado, foi criada uma coluna (soma das classificações académicas) na tabela com essa informação, conforme demonstrado na figura seguinte:

	A	C	D	E	F	G	H	I	J	K	L	M
1	Proc	Media	Média Arredondada	Faltas	Sexo	Profissão EE	Form. académica EE	Form. académica Pai	Class Form. Académica Pai	Form. académica Mãe	Class Form. Académica Mãe	Soma Class Académica
2	1831	11,75	12	25	M	Profissão Desconhecida	Formação Desconhecida	Básico (3º ciclo)	1	Básico (3º ciclo)	1	2
3	1981	15,4	15	11	F	Empregado de biblioteca	Licenciatura	Secundário	10	Licenciatura	30	40
4	2029	10,67	11	18	F	Mecânico e reparador de	Básico (3º ciclo)	Básico (3º ciclo)	1	Básico (3º ciclo)	1	2
5	2046	19,6	20	4	F	Técnico operador das tec	Básico (3º ciclo)	Básico (3º ciclo)	1	Básico (2º ciclo)	1	2
6	2048	17,4	17	3	M	Sapateiros e similares	Básico (2º ciclo)	Básico (2º ciclo)	1	Básico (2º ciclo)	1	2
7	2049	13	13	11	F	Outros trabalhadores da	Básico (1º ciclo)	Básico (2º ciclo)	1	Básico (1º ciclo)	1	2
8	2058	15,8	16	0	F	Operadores de caixa e ve	Básico (3º ciclo)	Formação Desconh	0	Básico (3º ciclo)	1	1
9	2059	18,2	18	14	M	Contabilista, auditor, revi	Licenciatura	Bacharelato	30	Licenciatura	30	60
10	2122	18,8	19	2	F	Auxiliar de cuidados de c	Secundário	Secundário	10	Secundário	10	20
11	2125	15,2	15	3	F	Agricultor e trabalhador q	Básico (2º ciclo)	Básico (2º ciclo)	1	Básico (2º ciclo)	1	2
12	2132	16,6	17	11	F	Outros profissionais de n	Secundário	Licenciatura	30	Secundário	10	40
13	2141	15,4	15	0	F	Engenheiro químico	Bacharelato	Básico (3º ciclo)	1	Bacharelato	30	31
14	2154	17,6	18	2	M	Profissão Desconhecida	Secundário	Secundário	10	Secundário	10	20
15	2161	12,71	13	21	M	Profissão Desconhecida	Formação Desconhecida	Formação Desconh	0	Básico (1º ciclo)	1	1
16	2166	10,43	10	52	F	Educador de infância	Licenciatura	Formação Desconh	0	Licenciatura	30	30
17	2171	15,67	16	146	F	Ajudante familiar	Básico (2º ciclo)	Básico (2º ciclo)	1	Básico (2º ciclo)	1	2
18	2174	18,6	19	5	M	Padeiros, pasteleiros e co	Básico (3º ciclo)	Secundário	10	Básico (3º ciclo)	1	11
19	2185	15	15	20	F	Padeiros, pasteleiros e co	Secundário	Secundário	10	Secundário	10	20
20	2200	13	13	7	F	Técnicos administrativos	Secundário	Formação Desconh	0	Secundário	10	10
21	2261	17,29	17	74	F	Director e gerente, de hot	Licenciatura	Licenciatura	30	Licenciatura	30	60

Figura 9 - Tabela com classificação académica do agregado familiar

FASE DA ANÁLISE DE DADOS

O próximo passo no *workflow* será a Análise dos Dados (*Data analysis*). Conforme recomendado pela literatura, é crucial iniciar **avaliando a qualidade dos dados**. Para começar esse processo, vai-se realizar-se uma análise dos dados (passos *Ask/answer questions* + *Analyse* + *Visualize*) que poderá revelar discrepâncias, incluindo dados em falta, *outliers*, e informações pouco fiáveis. Este também é o objetivo de investigação 1 que visa determinar se os dados são fiáveis ou não.

Após a triagem dos dados, será necessário proceder à **limpeza dos dados**, remoção dos registos dos alunos desistentes, transferidos para outras escolas ou reprovados por faltas, uma vez que estes não possuíam registos de classificação.

Após este processo, serão criadas tabelas de dados, derivadas da tabela geral, onde foram selecionados os dados relevantes para cada análise.

	A	C	D	E	F	G
1	Proc	Media	Média Arredondada	Ano	Tem computador	Tem internet
2	1831	11,75	12	12º ano	Sim	Sim
3	1981	15,4	15	12º ano	Não	Não
4	2029	10,67	11	11º ano	Não	Não
5	2046	19,6	20	12º ano	Sim	Sim
6	2048	17,4	17	12º ano	Não	Não
7	2049	13	13	11º ano	Sim	Não
8	2058	15,8	16	12º ano	Não	Não
9	2059	18,2	18	12º ano	Não	Não
10	2122	18,8	19	12º ano	Sim	Sim
11	2125	15,2	15	12º ano	Sim	Sim
12	2132	16,6	17	12º ano	Não	Não
13	2141	15,4	15	12º ano	Não	Não
14	2154	17,6	18	12º ano	Sim	Sim
15	2161	12,71	13	11º ano	Não	Não
16	2166	10,43	10	11º ano	Não	Sim
17	2171	15,67	16	12º ano	Não	Não
18	2174	18,6	19	12º ano	Não	Não

Figura 10 – Tabela com informação sobre acesso a Internet e Computador

Após a consolidação de todos os dados, optou-se por analisar exclusivamente os dados relativos aos alunos do ensino secundário regular. Esta decisão foi motivada pela amostra reduzida de alunos do 2º e 3º ciclo do ensino básico (ver capítulo da apresentação e discussão dos resultados), bem como pela natureza específica dos currículos e sistemas de avaliação do ensino profissional os quais diferem substancialmente do ensino regular.

No caso do ensino profissional, as classificações são distribuídas por diversos módulos, podendo cada disciplina ter mais de um módulo associado. No entanto, as classificações negativas não são registadas no *software*; quando um aluno não obtém uma classificação positiva, não lhe é atribuída qualquer nota no módulo em questão. Para a conclusão do curso e do ciclo de ensino, é exigido que o aluno do ensino profissional obtenha uma classificação positiva em todos os módulos.

A análise de dados centrou-se principalmente na influência do Encarregado de Educação e do agregado familiar no sucesso escolar do aluno. A decisão de seguir essa linha de

investigação deveu-se, em grande parte, à consulta da literatura, que apontava vários exemplos nesse sentido, e também à maior fiabilidade dos dados disponíveis.

Após a conclusão da fase de limpeza dos dados, procedeu-se à análise dos dados em conformidade com o objetivo de investigação 2.

OBJETIVO 3 - DESENVOLVER E AVALIAR UM MODELO DE PREVISÃO DO SUCESSO ESCOLAR COM BASE NOS DADOS DISPONÍVEIS.

Após a conclusão da limpeza e transformação dos dados, tornou-se necessário gerar novos conjuntos de dados com base nos registos existentes para alimentar o sistema de aprendizagem automática. Para treinar o algoritmo adequadamente, aproximadamente um terço dos registos originais precisaria ser reservado para fins de teste.

Assim, dos 656 registos originais, 198 foram reservados para testar o algoritmo, enquanto os restantes 458 foram utilizados para gerar os novos conjuntos de dados.

Para os novos conjuntos de dados destinados a alimentar o modelo preditivo, foram realizadas duas análises distintas. Uma análise mais robusta foi conduzida com base na informação de se o aluno foi reprovado ou não (média inferior a 10 ou média superior a 10) – **um modelo de classificação** - enquanto a outra análise teve como objetivo prever a classificação média do aluno – **um modelo de regressão**.

FASE DA COMUNICAÇÃO

Esta fase, composta pelos gráficos e relatórios resultantes da análise dos dados, representa as tabelas e gráficos que serão apresentados no capítulo de apresentação dos resultados.

INSTRUMENTOS DE RECOLHA DE DADOS

Para dar resposta ao objetivo 1, é necessário compreender o processo de registo. Durante o processo de matrícula online, os dados fornecidos pelos encarregados de educação no portal de matrículas são transferidos para o *software* INOVAR. Posteriormente, quando a matrícula é formalizada na escola, os serviços administrativos complementam esses dados no *software* de gestão escolar com informações adicionais provenientes do processo do aluno. Assim, identifica-se dois agentes humanos (pais e funcionários) que

estão responsáveis pela fiabilidade dos dados. Logo, é relevante compreender como os funcionários validam e complementam as informações do sistema. Para avaliar a qualidade desses dados produzidos pelos funcionários, será realizada uma entrevista com eles. O guião da entrevista é apresentado a seguir:

O guião da entrevista é apresentado a seguir:

- *Como são os dados introduzidos no processo dos alunos no software INOVAR?*
- *Os dados importados do portal das matrículas vêm totalmente preenchidos? Ou existem dados que são preenchidos pela secretaria?*
- *No caso dos dados que não são preenchidos, quais são?*
- *Esses dados são preenchidos à posteriori pela secretaria?*
- *Relativamente à informação da formação do* Encarregado de Educação *ou dos pais como é preenchida?*
- *No caso dos campos do formulário que não são importados quem preenche? Há um critério definido por todos os funcionários para o preenchimento dessa informação?*

Depois de recolher os dados com este instrumento de recolha de dados far-se-á uma transcrição da entrevista (apresentada no Anexo I).

INSTRUMENTOS DE ANÁLISE DE DADOS

Como tem vindo a ser referido, a técnica privilegiada de análise dos dados é a análise estatística. Para a entrevista, foi feita análise de conteúdo. A análise de conteúdo permite examinar e interpretar o conteúdo de texto da entrevista permitindo analisar padrões ou tendências por forma a obter *insights*.

CONSIDERAÇÕES ÉTICAS

Tal como referido no capítulo considerações éticas do uso de dados na educação, um dos princípios éticos envolve a salvaguarda da privacidade e proteção dos dados, garantindo a preservação da identidade do utilizador face às informações de identificação pessoal.

Assim, neste estudo, os dados pessoais dos alunos não foram utilizados; em vez disso, foram utilizados os números de processo do aluno para registo, o que impede a identificação individual dos alunos.

APRESENTAÇÃO E DISCUSSÃO DE RESULTADOS

Da limpeza de dados referida no capítulo anterior, o *dataset* (conjunto de dados) em análise é composto por 656 registos referentes aos alunos.

RESPOSTA AO OBJETIVO DE INVESTIGAÇÃO 1

Após uma análise das respostas à entrevista fornecidas pela funcionária, constatou-se que quando a informação não é completada no portal das matrículas, os campos em falta são preenchidos pelos assistentes administrativos. Contudo, não há uniformidade entre os funcionários quanto ao preenchimento desses campos em falta: alguns optam por inserir "Formação desconhecida", enquanto outros utilizam "Sem Formação". Embora essas variáveis não sejam objeto de análise neste estudo, verificou-se ainda que existem outros dados que não refletem a informação verdadeira, por exemplo: relativamente à profissão do Encarregado de Educação é comum, ao preencherem o formulário, selecionarem o primeiro campo disponível (valor pré-definido) na caixa de seleção relativa à profissão, que é "Agente de seguros".

Este procedimento compromete a precisão do processo de estratificação socioeconómica do agregado familiar com base na profissão.

Além disso, é importante considerar que não há garantia de veracidade dos dados importados do portal das matrículas. Por exemplo, o Encarregado de Educação pode preencher o campo da profissão como "Agente de seguros" apenas para acelerar o processo de preenchimento, sem que esta informação seja validada pela escola. Similarmente, no caso da escolaridade, essa informação pode ser inserida sem qualquer validação por parte da escola ou do Encarregado de Educação. Portanto, constata-se que existem variáveis cuja qualidade não está assegurada, o que comprometerá a sua utilização em modelos estatísticos.

Este problema pode estar relacionado com a **quantidade de dados** analisados; no entanto, é um facto que a escola regista poucas reprovações no ensino secundário regular. Além

disso, os alunos do ensino secundário regular têm a possibilidade de anular matrículas, o que resulta na ausência de qualquer classificação nas pautas. Isso, por sua vez, leva a que a sua classificação média seja apresentada como positiva.

RESPOSTA AO OBJETIVO 2

Como referido anteriormente, o *dataset* em análise é composto por 656 registos referentes aos alunos. Contudo, para efeitos de análise da sua qualidade e da correlação com o insucesso escolar, planeou-se analisar apenas 2/3 destes registos. Isto porque, e de forma a dar resposta ao terceiro objetivo, um terço destes registos seria intencionalmente deixado de fora da análise inicial com o propósito de validar o modelo preditivo que viria a ser desenvolvido. Ou seja, não se deseja que a informação desse 1/3 da informação "contamine" os dados que o modelo vai usar para o seu treino.

Porém, logo numa análise inicial para determinar se os dados iniciais seriam adequados para treino de um modelo preditivo, constatou-se que a o número de casos de reprovações (que representa o insucesso) era muito reduzido em comparação com os casos de sucesso:

Sucesso / Insucesso

	Alunos	**%**
Sucesso	615	94%
Insucesso	41	6%
TOTAL	**656**	

Tabela 1 - Sucesso e Insucesso geral

A escassez de dados de alunos reprovados pode ser um desafio significativo pois poderá ser mais difícil para o modelo aprender e generalizar padrões que ajudem na predição do insucesso escolar. Isso pode afetar a eficácia e a confiabilidade do modelo preditivo, uma vez que ele pode não ter dados suficientes para entender completamente os padrões associados ao sucesso e insucesso escolar. Desta forma, e de forma a dar resposta ao objetivo 2, analisaram-se os 656 registos.

Foram então produzidas diversas tabelas com o intuito de examinar a eventual associação entre o sucesso escolar e a acessibilidade à tecnologia, o nível de formação académica dos pais, a assiduidade dos alunos e o tipo de Encarregado de Educação.

IMPACTO DAS TIC NO SUCESSO ESCOLAR

No que concerne à relação entre o acesso à Internet, equipamento informático e o desempenho académico, os resultados são detalhados nas Tabelas 2 e 3. Ao observar essa associação, procurava-se compreender se existia impacto do acesso à tecnologia nos resultados escolares dos alunos.

Acesso a Internet / Sucesso					
Internet	**Insucesso**	**Sucesso**	**Total**	**Insucesso**	**Sucesso**
Com Internet	3	108	**111**	3%	97%
Sem Internet	38	507	**545**	7%	93%
TOTAL	**41**	**615**	**656**		

Tabela 2 - Relação entre acesso à Internet e sucesso escolar

Acesso a Computador / Sucesso					
Computador	**Insucesso**	**Sucesso**	**Total**	**Insucesso**	**Sucesso**
Com computador	2	108	**110**	2%	98%
Sem Computador	39	507	**546**	7%	93%
TOTAL	**41**	**615**	**656**		

Tabela 3 - Relação entre acesso a computador e sucesso

Dos alunos com acesso à Internet, apenas 3% apresentaram insucesso, enquanto 97% dos alunos apresentaram sucesso.

Por outro lado, 7% dos alunos sem acesso à Internet tiveram insucesso, enquanto 93% dos alunos sem acesso à Internet tiveram sucesso.

Com base nos dados apresentados na tabela, os resultados indicam que os alunos com acesso à Internet têm uma taxa menor de insucesso escolar em comparação com aqueles que não têm acesso à Internet. Tal é possível de analisar com maior clareza no gráfico 1.

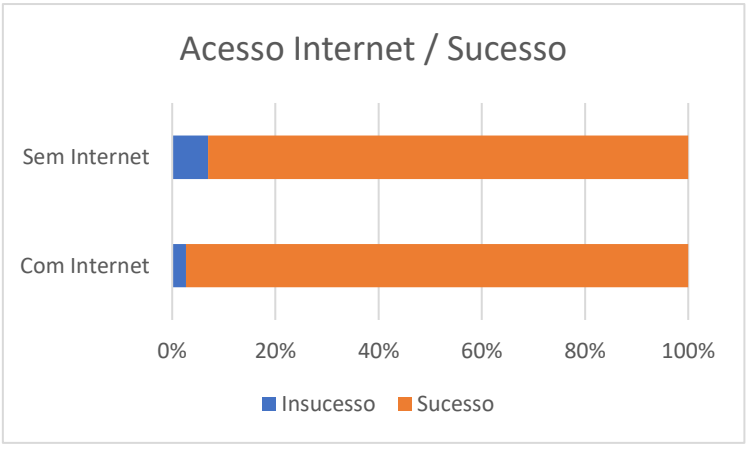


Gráfico 1 - Relação entre acesso à Internet e sucesso escolar

Relativamente ao acesso a computador, dos alunos que têm acesso a um computador, apenas 2% apresentaram insucesso escolar, enquanto 98% tiveram sucesso.

Por outro lado, entre os alunos sem acesso a um computador, 7 % apresentaram insucesso escolar e 93% tiveram sucesso.

À semelhança aos resultados referentes ao acesso à Internet, também estes resultados sugerem que o acesso a um computador pode estar associado a um melhor desempenho académico. Os alunos que têm acesso a um computador parecem ter uma taxa menor de insucesso escolar em comparação com aqueles que não têm acesso. Ver gráfico 2.

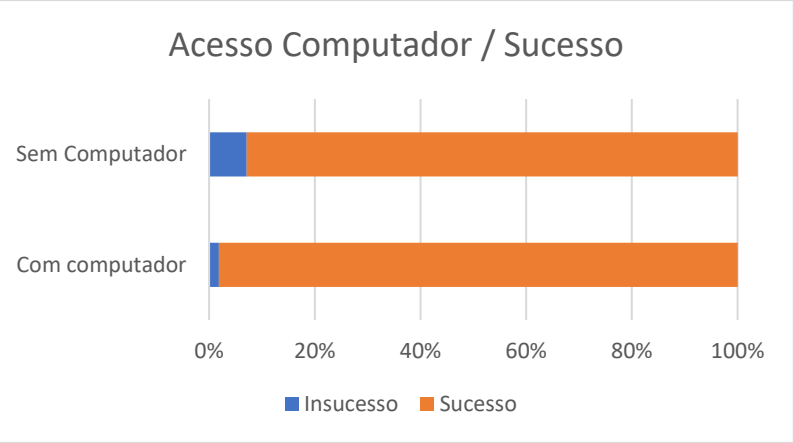


Gráfico 2 - Relação entre acesso a computador e sucesso

Nível de formação do Encarregado de Educação

O próximo aspeto a ser investigado é o impacto do nível de educação dos encarregados de educação, bem como do agregado familiar, sobre o desempenho académico dos alunos.

Dos 656 registos iniciais foram excluídos 119 registos devido ao nível de formação do Encarregado de Educação estar identificado como "Sem Formação" ou "Formação Desconhecida" e por isso foram considerados “poluídos”, assim restaram 537 registos.

A literatura sugere uma correlação consistente entre o nível escolar dos pais ou encarregados de educação e o desempenho académico dos alunos.

Formação EE / Sucesso					
Formação EE	**Insucesso**	**Sucesso**	**Total**	**Insucesso**	**Sucesso**
Básico	12	99	**111**	11%	89%
Secundário	6	181	**187**	3%	97%
Superior	5	234	**239**	2%	98%
	23	**514**	**537**		

Tabela 4 - Formação do EE / Sucesso

Na tabela 4 que relaciona o sucesso escolar com a formação do Encarregado de Educação, podemos observar o seguinte:

Para os Encarregados de Educação com formação de nível básico, a taxa de insucesso é de 11%, enquanto a taxa de sucesso é de 89%.

Quando o Encarregado de Educação possui formação de nível secundário, a taxa de sucesso aumenta para 97%, enquanto a taxa de insucesso é de 3%.

Nos casos em que o Encarregado de Educação possui formação de nível superior, a taxa de sucesso atinge 98%, com uma taxa de insucesso de 2%.

Estes resultados indicam que, quanto maior for o nível de formação do Encarregado de Educação, maior a probabilidade de sucesso do aluno. Consultar o Gráfico 3 para uma visualização mais clara e detalhada dos resultados.

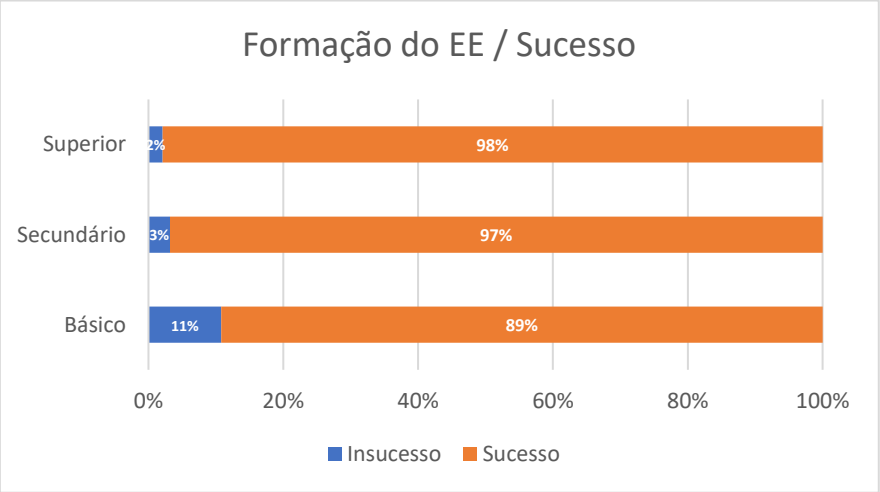


Gráfico 3 - Formação do EE / Sucesso

Pode-se verificar que existe a correlação mencionada na literatura teórica.

No que concerne à tabela 5 que relaciona a formação do Encarregado de Educação com a classificação do aluno:

Formação EE / Classificação

	Níveis																			
Formação EE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Básico	0%	0%	0%	0%	0%	0%	1%	5%	5%	8%	6%	17%	16%	8%	14%	11%	6%	1%	1%	1%
Secundário	0%	0%	0%	0%	0%	0%	1%	2%	1%	8%	7%	10%	14%	10%	21%	12%	6%	7%	1%	0%
Superior	0%	0%	0%	0%	0%	0%	0%	1%	1%	4%	8%	12%	9%	13%	15%	13%	11%	10%	2%	0%
	0%	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**2%**	**2%**	**7%**	**7%**	**13%**	**13%**	**10%**	**17%**	**12%**	**8%**	**6%**	**1%**	**0%**

Tabela 5 - Formação do EE / Classificação

Para os Encarregados de Educação com formação de nível básico, a distribuição das classificações varia consideravelmente, com notas entre o nível 7 e o nível 20.

No caso dos Encarregados de Educação com formação de nível secundário, observa-se uma tendência crescente nas classificações à medida que se avança nos níveis, com uma maior concentração, mais de 50% de classificações, a situar-se entre o nível 13 e 19.

Para os Encarregados de Educação com formação superior, verifica-se uma distribuição semelhante, mas com uma maior concentração de classificações nos níveis mais elevados de 16 a 18.

Estes resultados sugerem uma possível correlação entre a formação do Encarregado de Educação e o desempenho académico do aluno. Nota-se que, nos casos de insucesso, os alunos cujos Encarregados de Educação têm formação básica excedem em mais de três vezes os alunos cujos Encarregados possuem formação secundária, e em mais de cinco vezes no caso da formação superior. Este padrão evidencia uma tendência geral de melhores classificações associadas a um nível mais elevado de formação do Encarregado de Educação. No gráfico 4 pode observar-se com maior clareza a distribuição das classificações.

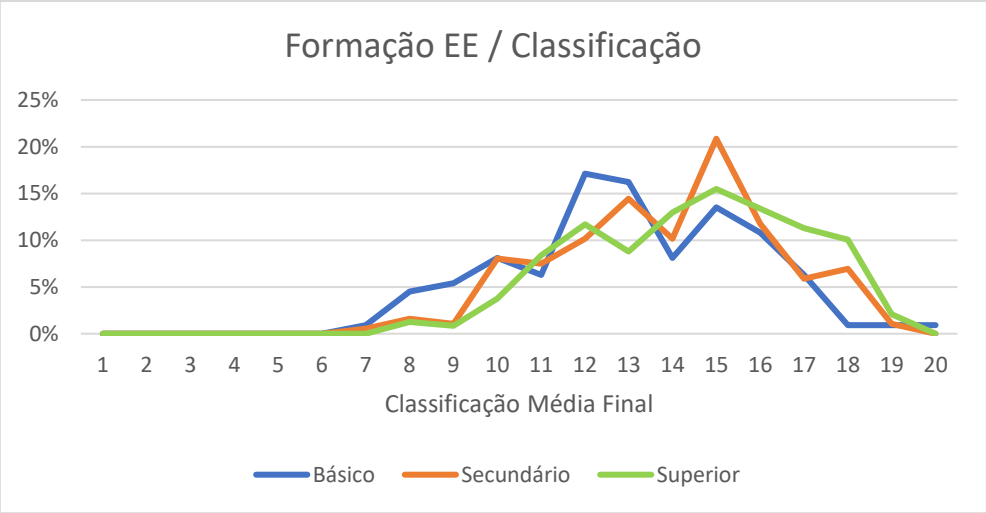


Gráfico 4 - Formação do EE / Classificação

NÍVEL DE FORMAÇÃO DO AGREGADO FAMILIAR

Relativamente à relação entre a formação do agregado familiar e o sucesso escolar os dados constam na tabela 6. Para estes dados foram retirados os registos em que na formação de ambos os pais estava o registo de "Formação desconhecida" ou "Sem formação", assim bastava que existisse apenas o registo de formação de um dos pais. Para a classificação da formação do agregado familiar foi considerado o nível mais elevado de formação perfazendo um total de 593 registos.

Formação Agregado Familiar / Sucesso

Formação Agregado Familiar	Insucesso	Sucesso	Total	Insucesso	Sucesso
Básico	12	105	**117**	10%	90%
Secundário	7	200	**207**	3%	97%
Superior	6	263	**269**	2%	98%
	25	**568**	**593**		

Tabela 6 - Formação do Agregado Familiar / Sucesso

Para os agregados familiares com formação básica, a taxa de insucesso é de 10%, enquanto a taxa de sucesso é de 90%. No caso dos agregados familiares com formação secundária, a taxa de insucesso diminui para 3%, enquanto a taxa de sucesso aumenta para 97%. Isso indica uma melhoria no sucesso escolar associada a uma formação secundária no agregado familiar.

Para os agregados familiares com formação superior, a taxa de insucesso continua a diminuir para 2%, enquanto a taxa de sucesso atinge 98%.

Em resumo, os resultados indicam que uma formação mais elevada no agregado familiar está associada a uma maior probabilidade de sucesso escolar, enquanto uma formação mais baixa está associada a um maior risco de insucesso, poderá verificar-se com maior clareza no gráfico 5.

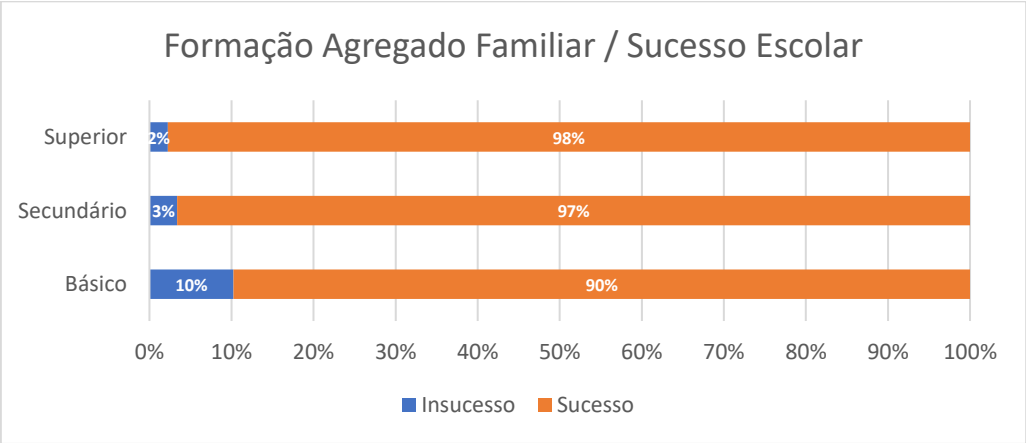


Gráfico 5 - Nível mais alto do agregado familiar / Sucesso

A literatura indica que os agregados familiares com maior nível de habilitação tendem a influenciar o sucesso escolar dos alunos. No entanto, embora se note que os agregados com formação básica apresentam resultados inferiores aos agregados com habilitações superiores, a diferença entre os níveis secundário e superior não é claramente distinta.
A tabela 7 mostra a distribuição das classificações dos alunos em relação aos diferentes níveis de formação do agregado familiar:

Formação Agregado Familiar / Classificação

	Níveis																			
Formação Agregado Familiar	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Básico	0%	0%	0%	0%	0%	0%	1%	3%	6%	9%	6%	21%	18%	8%	12%	9%	5%	1%	0%	1%
Secundário	0%	0%	0%	0%	0%	0%	0%	2%	1%	8%	7%	10%	14%	12%	21%	11%	5%	6%	1%	0%
Superior	0%	0%	0%	0%	0%	0%	0%	1%	1%	4%	7%	11%	10%	13%	14%	13%	12%	10%	2%	0%
	0%	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**2%**	**3%**	**7%**	**7%**	**14%**	**14%**	**11%**	**16%**	**11%**	**8%**	**6%**	**1%**	**0%**

Tabela 7 - Formação do agregado familiar / Classificações

Para os agregados familiares com formação básica, a distribuição das classificações dos alunos é mais concentrada em valores mais baixos (64% abaixo do nível 13), com uma proporção significativa de notas baixas (21% no nível 12, por exemplo). Isso sugere que

uma formação básica no agregado familiar está associada a um desempenho escolar mais fraco.

Nos agregados familiares com formação secundária, embora ainda haja uma proporção de notas nos níveis negativos, a maior concentração de classificações concentra-se entre o intervalo 10 e 16.

Já nos agregados familiares com formação superior, a maioria das classificações dos alunos está concentrada em valores mais altos, com uma proporção significativa de notas, mais de 50%, entre os níveis 15 e 19.

Isso sugere que uma formação superior no agregado familiar está associada a um melhor desempenho escolar.

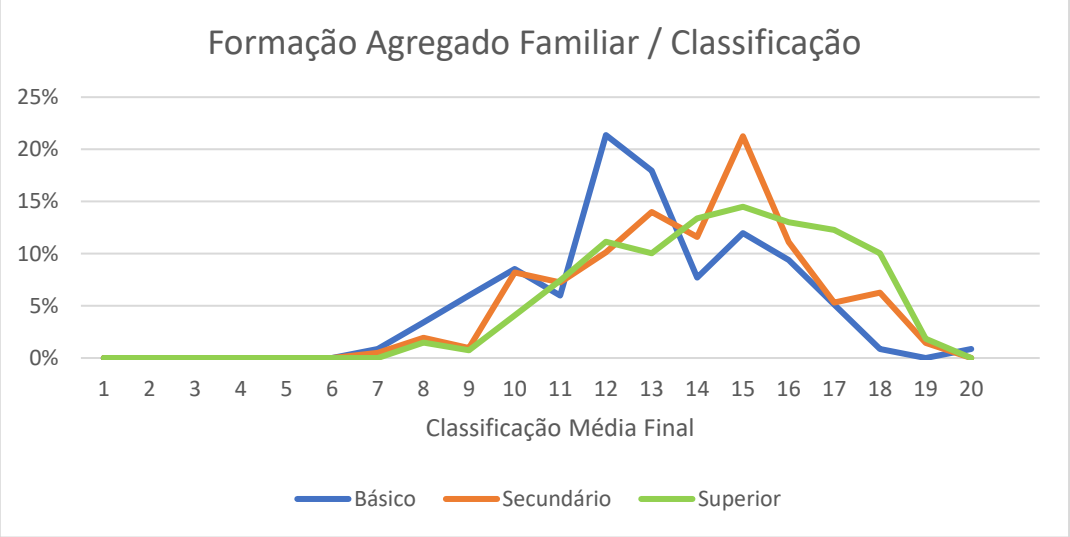


Gráfico 6 - Formação do Agregado Familiar / Classificação

Tipo de Encarregado de Educação

Em relação ao tipo de Encarregado de Educação e a sua influência no insucesso escolar, como previsto, a maioria dos encarregados de educação são mães, e estas são destacadas na literatura como tendo um impacto determinante no sucesso escolar; contudo, não foi identificada uma correlação evidente entre o tipo de Encarregado de Educação e o sucesso académico dos alunos. É relevante destacar que, no *software* de gestão escolar, quando o aluno atinge os 18 anos, ele próprio é automaticamente designado como seu Encarregado de Educação. A tabela 8 regista o sucesso por tipo de Encarregado de Educação.

Tipo de EE	Insucesso	Sucesso	Total	Insucesso	Sucesso
Mãe	21	459	**480**	4,4%	95,6%
Pai	6	95	**101**	5,9%	94,1%
Próprio	1	58	**59**	1,7%	98,3%
Outro	2	14	**16**	12,5%	87,5%
	30	**626**	**656**		

Tabela 8 - Tipo de EE / Sucesso

A maioria dos alunos tem a mãe como encarregada de educação, seguindo-se o pai como segundo maior representante, a terceira maior representatividade é o próprio (quando o aluno é o próprio Encarregado de Educação) e menor representado é "outros".

Observa-se que a percentagem de insucesso varia dependendo do tipo de Encarregado de Educação. Os alunos com "outro" como tipo de Encarregado de Educação apresentam a maior taxa de insucesso.

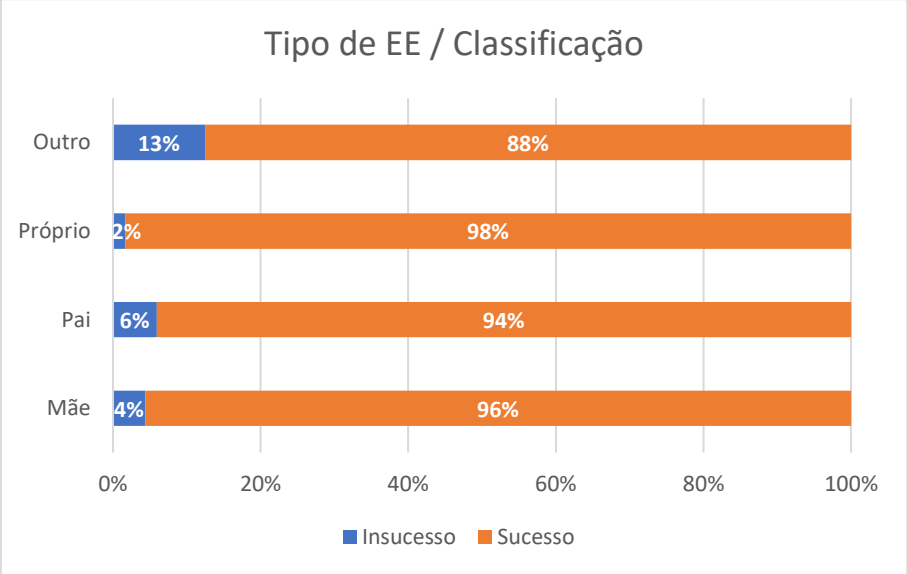


Gráfico 7 - Tipo de EE / Classificação

A tabela 9 apresenta a distribuição dos alunos de acordo com o tipo de Encarregado de Educação (EE) e a classificação dos alunos:

	Níveis																			
Tipo de EE	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Mãe	0%	0%	0%	0%	0%	0%	0%	2%	2%	6%	7%	13%	11%	11%	18%	12%	9%	8%	2%	0%
Pai	0%	0%	0%	0%	0%	0%	1%	4%	1%	6%	9%	10%	17%	16%	10%	16%	6%	4%	0%	1%
Próprio	0%	0%	0%	0%	0%	0%	0%	2%	0%	5%	2%	19%	19%	22%	15%	7%	7%	3%	0%	0%
Outro	0%	0%	0%	0%	0%	0%	0%	0%	13%	13%	0%	31%	6%	6%	25%	6%	0%	0%	0%	0%
	0%	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**2%**	**4%**	**7%**	**4%**	**18%**	**13%**	**14%**	**17%**	**10%**	**5%**	**4%**	**0%**	**0%**

Tabela 9 - Distribuição de níveis de classificação por tipo de EE

Os alunos com a mãe como encarregada de educação tendem a ter uma distribuição, com uma proporção considerável de alunos em níveis médios de classificação.

Os alunos com o pai como Encarregado de Educação apresentam uma distribuição de classificações que tende para uma proporção significativa de alunos nas classificações médias mais baixas que os alunos com Encarregado de Educação como mãe.

Os alunos que são o próprio Encarregado de Educação (Próprio) apresentam uma distribuição de classificações que tende a ser mais baixa.

Os alunos com outro tipo de Encarregado de Educação têm uma distribuição de classificações que tende a ser mais baixa, com classificações que não ultrapassam o nível 16 e com a maior proporção de insucesso dos quatro tipos de Encarregados de Educação analisados. No gráfico 8 pode observar-se a distribuição de níveis pelos diversos tipos de Encarregados de Educação.

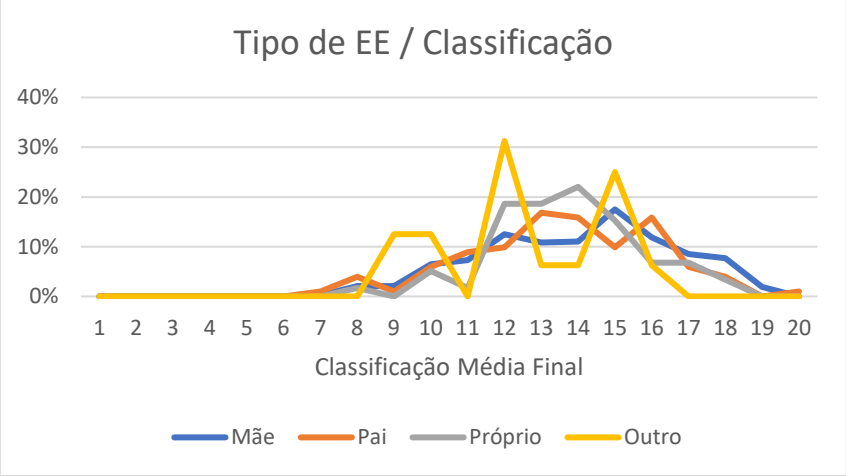


Gráfico 8 - Tipo de EE / Classificação

ASSIDUIDADE

Finalmente, uma das análises planeadas consistia em avaliar a relação entre a assiduidade dos alunos e a habilitação dos seus encarregados de educação, e consequentemente o seu sucesso académico. Para simplificar a interpretação dos dados, as faltas foram agrupadas em intervalos de 5.

A tabela 10 e gráfico 9 apresentam a relação entre a formação do Encarregado de Educação e a assiduidade dos alunos, agrupada em intervalos de faltas, observa-se uma distribuição uniforme da assiduidade tanto por níveis como por intervalos, não se observando correlações entre formação do Encarregado de Educação e assiduidade.

Assiduidade / Formação do EE

	Intervalos de Faltas					
Formação EE	**<5**	**Entre 5 e 10**	**Entre 10 e 15**	**Entre 15 e 20**	**Entre 20 e 25**	**>=25**
Básico (1º ciclo)	25%	21%	12%	14%	4%	24%
Secundário	24%	20%	18%	9%	10%	20%
Superior	25%	19%	17%	10%	6%	23%

Tabela 10 - Assiduidade / Formação do EE

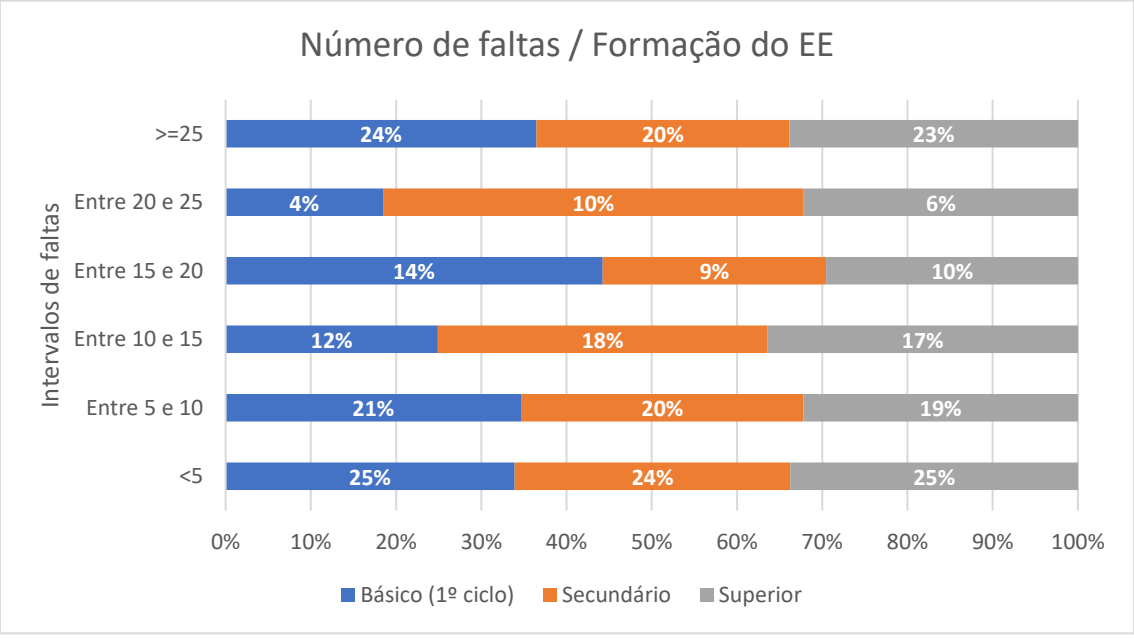


Gráfico 9 - Assiduidade / Formação do EE

A tabela 11 apresenta a distribuição dos alunos de acordo com o sucesso académico em diferentes intervalos de assiduidade.

Assiduidade / Sucesso

	Níveis																			
Intervalos de Faltas	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
<5	0%	0%	0%	0%	0%	0%	0%	3%	1%	5%	6%	4%	8%	9%	20%	17%	14%	9%	3%	1%
<10	0%	0%	0%	0%	0%	0%	0%	1%	2%	6%	8%	14%	13%	11%	18%	13%	8%	6%	1%	0%
<15	0%	0%	0%	0%	0%	0%	1%	2%	2%	5%	5%	18%	10%	20%	14%	8%	5%	9%	1%	0%
<20	0%	0%	0%	0%	0%	0%	2%	2%	0%	2%	5%	13%	20%	15%	20%	8%	8%	5%	0%	0%
<25	0%	0%	0%	0%	0%	0%	0%	0%	4%	6%	6%	16%	22%	14%	16%	10%	2%	4%	0%	0%
>=25	0%	0%	0%	0%	0%	0%	0%	5%	3%	11%	10%	17%	13%	11%	11%	11%	4%	3%	1%	0%
	0%	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**2%**	**2%**	**6%**	**7%**	**13%**	**12%**	**13%**	**16%**	**12%**	**8%**	**7%**	**1%**	**0%**

Tabela 11 - Assiduidade / Sucesso

Para os alunos com um número de faltas inferior a 5, observa-se uma tendência para um maior sucesso académico, com a maioria dos alunos alcançando classificações mais elevadas.

À medida que o número de faltas aumenta, a proporção de alunos com sucesso académico diminui gradualmente. No entanto, mesmo nos intervalos de faltas mais elevados (>= 25), ainda existem alguns alunos com sucesso académico.

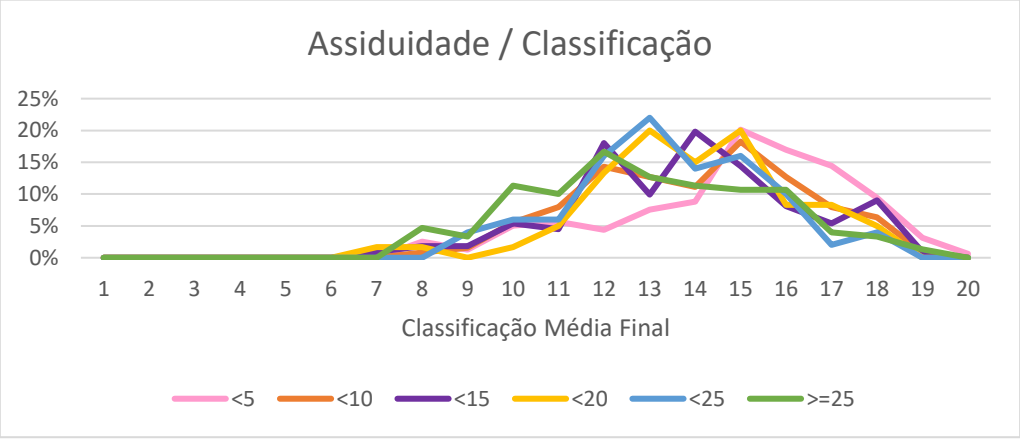


Gráfico 10 - Assiduidade / Classificação

Assim, dando resposta ao objetivo 2 (“Analisar a correlação entre certas variáveis e o sucesso ou insucesso escolar dos alunos”) constata-se que a correlação existe, mas é fraca. Os resultados indicam uma ligação entre as variáveis em análise e o desempenho académico dos alunos, no entanto, essa relação mostra-se limitada em termos de magnitude.

RESPOSTA AO OBJETIVO 3

No âmbito do objetivo 3 deste trabalho, tornou-se necessário realizar uma análise dos dados obtidos para determinar se estes poderiam ser utilizados para treinar um algoritmo de inteligência artificial capaz de prever o insucesso escolar com base em algumas das variáveis estudadas.

Para garantir a robustez desta análise, foram criadas tabelas, já demonstradas nos capítulos anteriores, contendo informações sobre a habilitação do Encarregado de Educação, uma das variáveis mencionadas na literatura como influente no sucesso escolar.

Neste contexto, existia a hipótese de desenvolver um modelo capaz de prever a nota quantitativa do aluno, numa escala de 0 a 20 valores, ou um modelo para antecipar a classificação qualitativa do aluno, isto é, “aprovado” ou “não aprovado”. Dada a insuficiência de dados para permitir que um modelo de inteligência artificial aprendesse os padrões necessários para prever com precisão as notas quantitativas (por exemplo, a amostra não incluía alunos com notas inferiores a 7 valores), foi decidido focar na variável qualitativa (categórica) "aprovado e reprovado".

Conforme observado na secção anterior, a categoria "aprovado" conta com 615 registos, enquanto a categoria "reprovado" possui 41 registos.

Treinaram-se e testaram-se 2 modelos comuns, o da **Regressão Logística** e o ***Random Forest***. Para isto usou-se dois terços dos dados para a aprendizagem de cada um dos modelos, e reservou-se um terço para o teste dos mesmos tal como referido no capítulo da análise preditiva. Ou seja, os modelos fizeram previsões com dados que os modelos não conheceram na fase do treino.

Para mineração dos dados usou-se o *software Orange Data Mining* com o *workflow* da figura 11.

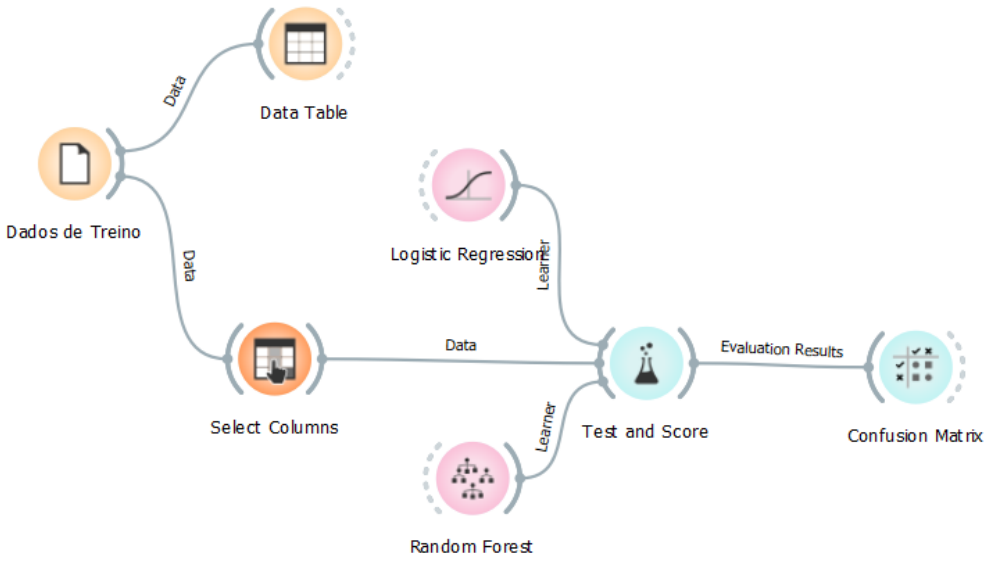


Figura 11 - Workflow do Orange Data Mining

O *workflow* termina com a elaboração de uma matriz de confusão, permitindo visualizar as previsões considerando os "erros" (falsos positivos e falsos negativos) e os "acertos" (verdadeiros positivos e verdadeiros negativos). A confiabilidade dos modelos é diretamente proporcional à diminuição da taxa de "erros" (falsos positivos e falsos negativos).

Como se pode verificar, o modelo *Random Forest* (Figura 12) é robusto na previsão de sucesso (acerta 99,7% das vezes) mas menos fiável na previsão de insucesso (só acerta 7,3% das vezes) e com uma taxa de erro de 92,7%.

		Predicted		
		Aprovado	**Não Aprovado**	**Σ**
Actual	**Aprovado**	99.7 %	0.3 %	**615**
	Não Aprovado	92.7 %	7.3 %	**41**
	Σ	**651**	**5**	**656**

Figura 12 – Modelo Random Forest

No caso do modelo Regressão Logística (Figura 13) a robustez na previsão do sucesso é de 99,5% mas relativamente ao insucesso é igualmente menos fiável na previsão com uma taxa de acerto de apenas 2,4% e com uma taxa de erro de 97,6%.

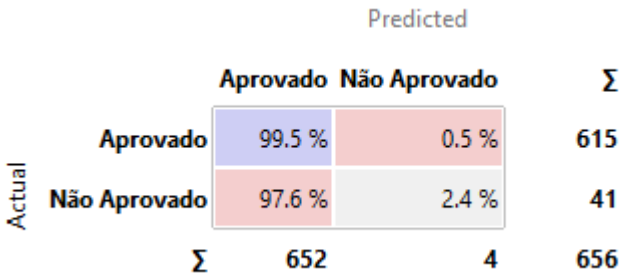


Figura 13 - Modelo de Regressão Logística

Conclui-se que a amostra de dados é inadequada para alimentar um sistema de aprendizagem automática. Isso ocorreu devido à **escassez de dados** relacionados com reprovações dos alunos (classe pouco representativa) em comparação com os alunos aprovados, tal como apresentado na tabela 1, o que levou a uma tendência do sistema em responder de forma enviesada para a aprovação do aluno (92,7% no caso do modelo *Random Forest* e 97,6% no caso do Modelo de Regressão Logística), independentemente de outros fatores.

Isso levaria o sistema de inteligência artificial a responder que o aluno foi aprovado quando não o era, porque a probabilidade do sistema errar respondendo que o aluno iria reprovar seria muito maior comparativamente com a resposta que o aluno iria aprovar.

CONCLUSÕES

No contexto da administração escolar, a precisão e confiabilidade dos dados desempenham um papel crucial na tomada de decisões informadas e na melhoria contínua dos processos escolares. Nesse sentido, a avaliação da qualidade da informação nas bases de dados do *software* de gestão escolar de uma instituição de ensino torna-se uma tarefa fundamental.

Um dos objetivos principais deste trabalho era examinar criticamente a qualidade dos dados armazenados no sistema de gestão escolar, identificando possíveis lacunas, inconsistências e falta de confiabilidade nas informações disponíveis. É importante ressaltar que a confiabilidade dos dados é essencial para garantir que as análises e relatórios gerados a partir desses dados sejam precisos e úteis para a tomada de decisões ou para utilizar em sistemas computacionais capazes de traçar modelos preditivos.

Ao avaliar a qualidade da informação uma das principais preocupações é a confiabilidade dos dados. Devido à natureza dinâmica das operações escolares e à entrada manual de dados por parte de diferentes intervenientes, verificou-se que nem todos os dados eram totalmente fiáveis. A informação no processo digital do aluno no *software* de gestão é completada de forma diferente pelos intervenientes e muitas vezes da forma mais célere, mas não precisa e tendo em conta a informação disponibilizada.

Assim, e respondendo à pergunta de partida, verifica-se que, na escola onde este estudo foi aplicado, os dados não possuem ainda a qualidade nem quantidade necessária para serem usados em modelos preditivos do insucesso escola.

Portanto, para que se possam utilizar os dados para alimentar mais tarde ferramentas de *machine learning* e para que estas possam auxiliar na tomada de decisão, será necessário uniformizar os processos de entrada de dados e dos métodos de validação utilizados pelo *software* de gestão escolar.

Além disso, também é importante considerar a integridade dos dados, garantindo que todas as informações relevantes estejam presentes e corretas. A completude dos dados é essencial para fornecer uma visão abrangente e precisa do desempenho escolar, das matrículas dos alunos, da frequência, entre outros aspetos essenciais da gestão escolar.

Em resumo, a avaliação da qualidade da informação nas bases de dados do *software* de gestão escolar de uma escola é um processo essencial para garantir a precisão e confiabilidade das informações utilizadas na administração escolar. Ao identificar e corrigir quaisquer problemas relacionados à qualidade dos dados, a escola estará melhor posicionada para tomar decisões informadas e promover a melhoria contínua de seus processos educacionais.

Observou-se ainda que existem variáveis cuja qualidade não está assegurada, o que comprometerá a sua utilização em modelos estatísticos. No entanto, esta questão pode ser corrigida através da padronização do processo junto dos serviços administrativos ou de alterações nas plataformas de *software* –. Por exemplo, na seleção da formação do Encarregado de Educação, existem duas opções que, apesar de semelhantes, são distintas: "Formação desconhecida" ou "Sem formação". Quando o utilizador que preenche os dados seleciona uma das opções, nem sempre considera a informação que consta no processo do aluno. Se for selecionada a opção "Formação desconhecida", o utilizador está, desde logo, a registar que o Encarregado de Educação tem uma formação, quando na realidade o que deveria fazer era registar essa formação no *software*

Estes procedimentos condicionam à posteriori, não apenas a análise de informação, mas a possível construção e ferramentas de IA que permitam auxiliar a tomada de decisão.

Concluindo, a avaliação da qualidade da informação nas bases de dados do *software* de gestão escolar é essencial para assegurar a precisão e fiabilidade dos dados utilizados na gestão escolar. A uniformização e padronização dos processos de entrada de dados, procedimentos administrativos e métodos de validação no *software* de gestão escolar é de suma importância para viabilizar a utilização dos dados em futuras aplicações de *machine learning*.

TRABALHO FUTURO

Em estudos futuros, pode ser pertinente considerar, além da classificação média, a variável que contabiliza o número de anulações de matrículas ou as classificações finais obtidas em exames nacionais. Dado que foi observada alguma correlação entre as variáveis, essa informação pode ser útil para um professor ponderar nos seus processos

de tomada de decisão. Logo, seria pertinente avaliar se os professores considerariam útil o uso de *dashboards* com este tipo de informação.

BIBLIOGRAFIA

As 6 Dimensões da Qualidade de Dados (Data Quality) - Data Science Academy. (2023, 26 de novembro). Data Science Academy Sua carreira elevada a outro nível. https://blog.dsacademy.com.br/as-6-dimensoes-da-qualidade-de-dados-data-quality/

Alexandron, G., Ruipérez-Valiente, J. A., Chen, Z., Muñoz-Merino, P. J., & Pritchard, D. E. (2017). Copying@Scale: Using Harvesting Accounts for Collecting Correct Answers in a MOOC. *Computers & Education*, *108*, 96–114. https://doi.org/10.1016/J.COMPEDU.2017.01.015

Bêa, M. (2018). Measuring school segregation: Evidence from Lisbon. A Work Project em Mestrado em Economia, Faculdade de Economia da Universidade Nova de Lisboa, Lisboa, Portugal.

Brooks, C., & Thompson, C. (2017). Predictive modelling in teaching and learning. Handbook of learning analytics, 61-68.

Brooks, C., Thompson, C., & Teasley, S. (2015). A time series interaction analysis method for building predictive models of learners using log data. In LAK '15: the 5th International Learning Analytics and Knowledge Conference. ACM. https://doi.org/10.1145/2723576.2723581

Awan, A. A. (2023, 13 de setembro). The Curse of Dimensionality in Machine Learning: Challenges, Impacts, and Solutions. Learn Data Science and AI Online | DataCamp. https://www.datacamp.com/blog/curse-of-dimensionality-machine-learning

Carneiro, P. (2008). Equality of opportunity and educational achievement in Portugal. Portuguese Economic Journal, 7(1), 17-41.

Carneiro, P., Meghir, C. and Parey, M. (2013), Maternal education, home environments, and the development of children and adolescents. Journal of the European Economic Association, 11: 123-160. https://doi.org/10.1111/j.1542-4774.2012.01096.x

Cunha, R. F. F. d. (2023). Dashboard como instrumento tecnológico para aprimorar o ensino na educação profissional e tecnológica. *Revista Educação em Debate*.

European Commission, Joint Research Centre, Hillaire, G., Ferguson, R., Brasher, A., Clow, D., Cooper, A., Hillaire, G., Mittelmeier, J., Rienties, B., Ullmann, T., & Vuorikari, R. (2016). *Research Evidence on the Use of Learning Analytics*. https://doi.org/https://data.europa.eu/doi/10.2791/955210

European Commission, Directorate-General for Education, Youth, Sport and Culture, Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators, Publications Office of the European Union, 2022, https://data.europa.eu/doi/10.2766/153756

Fernandes, G., & Martins, J. A. (2015). Influência da assiduidade no processo de ensino-aprendizagem no ensino Politécnico. Situação e estratégias no Instituto Politécnico da Guarda IPG-Portugal. In Congresos CLABES.

Firmino, J., Nunes, L., Reis, A., e Seabra, C. (2016). Class composition and student achievement in Portugal. Tese de Doutoramento em Economia, Faculdade de Economia da Universidade Nova de Lisboa, Lisboa, Portugal.

Gonçalves, Ana Rita Costa. (2012) O Papel das Tic na Escola, na Aprendizagem e na Educação, ISCTE - Instituto Universitário de Lisboa, Portugal.

GRILO JÚNIOR, Tarcísio Ferreira. Aplicação de técnicas de Data Mining para auxiliar no processo de fiscalização no âmbito do Tribunal de Contas do Estado da Paraíba. 2010. 103 f. Dissertação (Mestrado em Engenharia de Produção) - Universidade Federal da Paraí-ba, João Pessoa, 2010.

Hershkovitz, A., & Alexandron, G. (2020). Comprendiendo el potencial y los desafíos del *Big Data* en las escuelas y la educación. *Tendencias Pedagógicas, 35*, 7. https://doi.org/10.15366/tp2020.35.002

Holmes, N. (2018). Engaging with assessment: Increasing student engagement through continuous assessment. Active Learning in Higher Education, 19(1), 23–34. https://doi.org/10.1177/1469787417723230

Iberdrola. (2021) ANÁLISE PREDITIVA. https://www.iberdrola.com/inovacao/analises-preditivas

O que é engenharia de recursos: arte ou ciência? - Alteryx. (s.d.). Alteryx. https://www.alteryx.com/pt-br/glossary/feature-engineering

OVANOVIĆ, V. et al. Analytics of learning and teaching in educational ecosystems. In: SPECTOR, J. M.; LOCKEE, B. B.; CHILDRESS, M. D. (ed.). Learning, design, and technology: an international compendium of theory, research, practice, and policy. [S. l.]: Springer, 2019. p. 1-21

Mahanti (2019) — Qualidade de Dados: Dimensões, Medição, Estratégia, Gestão e Governança:

Mustard JF. Desenvolvimento cerebral inicial e desenvolvimento humano. Em: Tremblay RE, Boivin M, Peters RDeV, eds. Enciclopédia sobre o Desenvolvimento na Primeira Infância [on-line]. https://www.enciclopedia-crianca.com/importancia-do-desenvolvimento-infantil/segundo-especialistas/desenvolvimento-cerebral-inicial-e. Publicado: Fevereiro 2010 (Inglês). Consultado em 10 de dezembro de 2023.

Munshi, A., Rajendran, R., Penn, J. O., Biswas, G., Baker, R. S., & Paquette, L. (2018). Modeling learners' cognitive and affective states to scaffold srl in open-ended learning environments. *UMAP 2018 - Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 131–138. https://doi.org/10.1145/3209219.3209241

Mucharreira, P., Cabrito, B., Capucha, A., Capucha, L., Carvalho, H., Sebastião, J., Martins, S., Roldão, C., e Tavares, I. (2017). A dimensão das turmas no Sistema Educativo Português. Secretaria-Geral da Educação e Ciência. Lisboa: ISCTE- Instituto Universitário de Lisboa e CIES, Centro de Investigação e Estudos de Sociologia, 20-197

O que é Data Mining? – Tecnoblog. (s.d.). Tecnoblog. https://tecnoblog.net/responde/o-que-e-data-mining/. Consultado em 02-03-2024

O ano da Inteligência Artificial. Qual o impacto na Educação? (2023, 24 de março). NAU site. https://www.nau.edu.pt/pt/noticias/inteligencia-artificial-impacto-educacao/

Patel, N. (s.d.). Dashboard: guia definitivo. Neil Patel. https://neilpatel.com/br/blog/dashboard-o-que-e/ Consultado em 02-03-2024

Raykar, V., & Saha, A. (2015). Data Split Strategies for Evolving Predictive Models.

RIBEIRO, C. J. S. – Big Data: os novos Desafios para os Profissionais da Informação. Rio de Janeiro. UNIRIO. 2014

Santos, H. G. (2018). Comparação da performance de algoritmos de machine learning para a análise preditiva em saúde pública e medicina. Tese de Doutorado, Faculdade de Saúde Pública, Universidade de São Paulo, São Paulo. doi:10.11606/T.6.2018.tde-09102018-132826. Recuperado em 2024-03-02, de www.teses.usp.br

Shaw, Dhawal Jan 18th, 2018, By The Numbers: MOOCS in 2017, Consultado em 28-12-2022 Disponivel em: https://www.classcentral.com/report/mooc-stats-2017/

ONU.(2013). Objetivos de Desarrollo del Milenio Informe de 2013, Naciones Unidas,Nueva York.

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning Analytics Dashboard Applications. *American Behavioral Scientist*, *57*(10), 1500–1509. https://doi.org/10.1177/0002764213479363

What is a Data Science Workflow? (s.d.). Data Science Process Alliance. https://www.datascience-pm.com/data-science-workflow/

Wolff, E. (2023, 10 de abril). *Como a ciência de dados pode ajudar na gestão escolar?* SINEPE/RS. https://sinepe-rs.org.br/educacaoempauta/com-a-palavra/como-a-ciencia-de-dados-pode-ajudar-na-gestao-escolar/

ANEXO I – ENTREVISTA À FUNCIONÁRIA DA SECRETARIA

Eu – Bom dia Sra. H.

H. – Bom dia professor.

Eu – Necessito de compreender como os dados são preenchidos no processo dos alunos no *software* INOVAR, como são introduzidos?

H. – Numa primeira fase os dados são introduzidos através da importação do portal das matrículas e ficam registados no INOVAR após importação.

Eu – Os dados importados vêm do portal totalmente preenchidos? Ou existem dados que são preenchidos pela secretaria?

H – Habitualmente os dados obrigatórios no portal das matrículas vêm todos preenchidos, até porque a plataforma não permite efetuar a matrícula sem estarem preenchidos, no entanto os dados que são facultativos nem todos são registados no portal.

Eu – No caso desses dados que não são preenchidos, quais são?

H – Habitualmente são os dados referentes à formação e à profissão dos encarregados de educação e dos pais.

Eu - Esses dados são preenchidos à posteriori pela secretaria?

H – Apenas são preenchidos os dados de contactos dos encarregados de educação e no caso da informação estar toda presente no processo do aluno, por exemplo, no caso da profissão se o professor for verificar no INOVAR há uma grande percentagem de pais que têm como profissão "Agente de Seguros" porque é a primeira opção que aparece na caixa de seleção.

Eu – E no caso da formação? Reparei que por vezes aparece "Sem formação" e noutros casos aparece "Formação desconhecida", quem preenche e qual o critério para selecionar uma ou outra opção?

H – No caso da profissão quando não vem essa informação do portal das matrículas é preenchida por nós, no caso dessa informação acompanhar o processo do aluno colocamos no INOVAR, mas na maioria das vezes é colocada uma dessas opções. Como somos várias pessoas a preencher a nomenclatura a utilizar depende da pessoa que está a preencher, alguns colegas escolhem a opção “Sem formação” e outros “Formação desconhecida”.