Felipe Ferreira da Fonseca

MILAGE Aprender+:
aprendizagem personalizada suportada por aprendizagem máquina

Universidade do Algarve
Faculdade de Ciências e Tecnologia
2022

Felipe Ferreira da Fonseca

MILAGE Aprender+:
aprendizagem personalizada suportada por aprendizagem máquina

Mestrado em Engenharia Informática
Aluno: Felipe Ferreira da Fonseca

Trabalho efetuado sobre a orientação de:
Professora Doutora Paula Ventura Martins
Professor Doutor Mauro Figueiredo

Universidade do Algarve
Faculdade de Ciências e Tecnologia
2022

MILAGE Aprender+:
aprendizagem personalizada suportada por aprendizagem máquina

DECLARAÇÃO DE AUTORIA DE TRABALHO
Declaro ser o autor deste trabalho, que é original e inédito.
Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências bibliográfica incluída.

Felipe Ferreira da Fonseca

_____

Dedicatória

Dedico este trabalho a minha amada esposa Fabíola, que abdicou de tudo para estar ao meu lado, que sempre me incentivou, apoiou e lutou comigo durante toda esta jornada.

Dedico ao meu querido pai, Valdemir e a minha mãe, Eudes. Sem vocês nada disso teria acontecido. Vocês são minha inspiração no que se refere a luta, mérito, honra e coragem.

Ao meu amado irmão, Marcos, meu melhor amigo e grande professor.

Dedico a todos os professores e a todos os trabalhadores que impactam positivamente, diretamente ou indiretamente a Educação.

Agradecimentos

Agradeço a toda minha querida família que, desde o primeiro dia, me apoiou cada segundo ao longo deste projeto. Foi forte para suportar a saudade e forte para suportar todos os medos, forte para não perder o rumo mesmo em tempos tão duros.

À Professora Doutora Paula Ventura Martins e ao Professor Doutor Mauro Figueiredo, pelo apoio, paciência e confiança.

À Universidade do Algarve pelo acolhimento de um estrangeiro, ao apoio irrestrito de toda a equipa de trabalho, bibliotecárias, secretarias, equipa de suporte informática e em especial aos professores do curso de Mestrado de Engenharia em Informática, pela transmissão de conhecimentos e experiências.

Sou eternamente grato ao senhor Deus por ter me acolhido em vários momentos críticos desta jornada. Obrigado, Senhor.

E por fim, mas não menos importante, a todas as pessoas que direta ou indiretamente me apoiaram durante estes últimos anos e que contribuíram de certo modo, para concretizar este trabalho.

Resumo

Palavras-chave: Aprendizagem Máquina; MILAGE Aprender+; AM Automatizada; Ensino a distância;

O MILAGE Aprender+ é uma aplicação para dispositivos móveis com propósito educacional dentro e/ou fora do ambiente escolar. Os seus utilizadores têm acesso a uma ampla biblioteca de conteúdo para a realização de atividades diversas disciplinas escolares. Destaca-se por ser uma aplicação que prioriza a autonomia do aluno. Apresenta-se como uma ferramenta essencial para responder à atual procura relacionada com o ensino remoto e também presencial A aplicação MILAGE Aprender+ adota também uma característica jogo, contabilizando pontuações ao longo das realizações dos alunos. Desde o seu lançamento em 2018, possui um amplo conjunto de dados, um amplo histórico de atividades realizadas pelos seus utilizadores. No modelo atual de funcionamento o MILAGE Aprender+ é percebido um problema relacionado a forma que apresenta o conteúdo aos utilizadores. O problema é percebido pela forma que o conteúdo da aplicação é apresentado ao aluno, de forma totalmente uniforme e padronizada a todos os seus utilizadores, não tendo em consideração o contexto histórico do desempenho dos seus utilizadores ao longo do uso da aplicação. Não é aplicado nenhum tipo de recomendação de conteúdo ou apresentação de forma personalizada. Pelo que, o presente estudo, tem como objetivo avaliar a possibilidade de evolução da aplicação MILAGE Aprender+ com a adoção da tecnologia de Aprendizagem Máquina (AM) a fim de proporcionar uma melhor experiência ao seu utilizador, o aluno. Pretende-se recomendar conteúdos a partir de modelos que promovam melhorar o aproveitamento de pontos do aluno. A partir deste objetivo, o presente estudo aprofundou-se na AM, que é um método de análise e ciência de dados que automatiza a construção de modelos analíticos. Esta tese adotou a metodologia de desenvolvimento de projetos de prospeção e análise em ciência de dados CRISP-DM. A investigação desenvolveu uma ampla análise sobre o problema do MILAGE Aprender+, e a possibilidade da adoção da tecnologia de AM. O principal objetivo de permitir que os alunos, utilizadores da ferramenta, recebam a recomendação de conteúdos das disciplinas de forma adaptável, aproximada às suas necessidades e baseada no histórico de desempenho, para que assim possam assim ter um percurso de aprendizagem mais objetivo e também agradável. Como conclusão deste estudo evidenciou-se a possibilidade de adoção de modelos de AM no MILAGE Aprender+ para a identificação e recomendação de conteúdos aos alunos.

Abstract

Keywords: Machine Learning; MILAGE Learning; Automated Machine Learning; Distance learning.

MILAGE Learn+ is an application for mobile devices with educational purposes inside and/or outside the school environment. Its users have access to a wide library of content for carrying out activities in different school subjects. It stands out for being an application that prioritizes student autonomy. It is presented as an essential tool to respond to the current demand related to remote and face-to-face teaching. The MILAGE Learning+ application also adopts a game feature, counting scores along with the students' achievements. Since its launch in 2018, it has had a wide set of data, a wide history of activities carried out by its users. In the current operating model of MILAGE Learning+, a problem is perceived related to the way it presents the content to users. The problem is perceived by the way that the content of the application is presented to the student, in a totally uniform and standardized way to all its users, not considering the historical context of the performance of its users throughout the use of the application. No content recommendation or personalized presentation is applied. Therefore, the present study aims to evaluate the possibility of evolving the MILAGE Learning+ application with the adoption of Machine Learning (ML) technology to provide a better experience to its user, the student. It is intended to recommend content based on models that promote the improvement of the student's use of points. From this objective, the present study delved into AM, which is a data science and analysis method that automates the construction of analytical models. This thesis adopted the CRISP-DM data science prospecting and analysis project development methodology. The investigation developed a broad analysis of the problem of MILAGE Learn+, and the possibility of adopting ML technology. The main objective is to allow students, users of the tool, to receive the recommendation of contents of the disciplines in an adaptable way, close to their needs and based on the performance history so that they can thus have a more objective and pleasant learning path. As a conclusion of this study, the possibility of adopting ML models in MILAGE Learn+ was evidenced to identify and recommend content to students.

Índice

Índice de Figuras

Índice de tabelas

Capítulo 1: Introdução

1.1 Contexto

Este trabalho de mestrado propõe o estudo da aplicação MILAGE Aprender+ com o objetivo de adotar algoritmos de Inteligência Artificial (IA), em específico de Aprendizagem Máquina (AM) para adaptar e personalizar o conteúdo da aplicação.

A aplicação MILAGE Aprender+ é uma ferramenta em difusão e com notório reconhecimento dos alunos que aderiram à plataforma. Numa publicação do Jornal Expresso (Soares, 2020), foi noticiado que a tecnologia "já é utilizada por 50 mil alunos, não somente em Portugal como em Espanha, Chipre, Alemanha, Noruega ou Turquia". Aplicada num contexto dentro ou fora da sala de aula, a aplicação MILAGE Aprender+ tem como objetivo apoiar o aluno e o professor na aprendizagem escolar e providenciando uma ampla biblioteca de conteúdo, atividades e exercícios.

É a partir dos dados históricos dos utilizadores da aplicação e de todo o conjunto de interações realizadas no MILAGE Aprender+, que este projeto apresentará o resultado do estudo da adoção de modelos de AM, ou seja, a partir de algoritmos e modelos autônomos, que propõe resultados sucessivamente mais adaptados e respostas a partir de análise estatística (Hosch, 2009).

Com o objetivo de propor uma experiência mais adaptada e personalizada, o estudo direcionado para a adoção de AM no MILAGE Aprender+ consiste numa mudança gradual da apresentação do conteúdo da aplicação. Evoluindo a apresentação ao utilizador de uma experiência homogênea de um para muitos para uma experiência de aprendizagem mais personalizada e adaptada. A apresentação do conteúdo de uma forma personalizada é defendida em estudos de que modelos de educação baseados em dados permitirão a educação personalizada e melhorarão os resultados para alunos, educadores e administradores (King, et al., 2016).

Este projeto apresentará o resultado da adoção dos modelos de AM de modo que o utilizador, aluno, tenha uma melhor experiência relacionada com o conteúdo que lhe é recomendado. Ao professor também será proporcionada uma experiência personalizada no acompanhamento e gestão do aproveitamento e progresso do aluno, ao longo da utilização da aplicação MILAGE Aprender+, a partir da previsão dos resultados de aproveitamento com maior fiabilidade estatística.

1.2 Motivação

Com o advento de uma maior disponibilidade, capacidade de processamento e oferta de recursos computacionais e maior adoção e adaptação do uso destes recursos tecnológicos, tem-se proporcionado uma maior abertura para promover esta transformação e ampliação do modelo de ensino e respetivas ferramentas de ensino, para além da forma presencial. Este processo de transformação e ampliação do ensino, também para o ambiente à distância e remoto, foi amplamente acelerado no ano 2020 devido as exigências de distanciamento relacionado com a pandemia COVID-19 (Casatti, 2020). Segundo relatório da UNESCO (2020), "O fecho das escolas afectou cerca de 80% da população estudantil do mundo", "1,37 bilhão de estudantes estão em casa após o fechamento da escola, devido ao alastrar do COVID-19", "os ministros ampliam as abordagens multimédia para garantir a continuidade da aprendizagem".

Um outro indicador de referência que traduz este recente aumento da procura e necessidade, acentuada pela pandemia COVID-19, é apresentado pelo Google Trends, que avalia o comportamento do utilizador na internet a partir de pesquisas e buscas. Como podemos ver na Figura 2, as buscas e o interesse apresentaram um crescimento de 75% para os utilizadores portugueses entre o período de 2015 a 2020 para o assunto e palavra-chave "Educação a distância".

Figura 1 - "Educação a distância" desde 2015 até dezembro 2021 em Portugal (Google Trends, 2020) .

Acresce neste cenário a evolução da tecnologia na resolução de problemas com a adopção da AM, um ramo da Inteligência Artificial, que conciliado aos recentes avanços na tecnologia, tem ampliado e dado maior abertura para a resolução de problemas complexos (Sichman, 2021).

Neste contexto, onde se observa cada vez mais uma maior tendência e necessidade no recurso a ferramentas de ensino à distância, é importante identificar deficiências e oportunidades no âmbito da ferramenta MILAGE Aprender+. Pretende-se favorecer e oferecer aos utilizadores, alunos e professores, uma tecnologia que tire partido dos dados históricos e apresente uma trajetória de aprendizagem apoiada por uma experiência personalizada na disponibilização de conteúdos.

Cada vez mais, a IA tem transformado o quotidiano das pessoas. Com base numa análise de questionários entregues a investigadores de IA que possuem elevada experiência na área, os resultados referem que a IA superará os recursos humanos nos próximos dez anos em várias atividades, tais como, traduzir idiomas (até 2024), escrever ensaios de ensino médio (até 2026), conduzir um camião (até 2027), escrever um livro best-seller (até 2049) e trabalhar como um cirurgião (em 2053). Os investigadores acreditam ainda que existe uma hipótese de 50% da IA superar os humanos em todas as tarefas nos próximos 45 anos e de automatizar todos os trabalhos humanos em 120 anos (Grace et al., 2018). Atualmente, estamos rodeados por esta tecnologia em sistemas de estacionamento de veículos de forma autónoma, sensores inteligentes para tirar fotos espetaculares e assistência pessoal. Atualmente, estamos rodeados por esta tecnologia em sistemas de estacionamento de veículos de forma autónoma, sensores inteligentes para tirar fotos espetaculares e assistência pessoal.

Da mesma forma, há avanços na aplicação de IA na educação e os métodos tradicionais estão evoluindo à medida que o contexto e necessidades dos alunos e professores também se transformam. O ambiente de ensino torna-se cada vez mais personalizado e adaptável aos alunos. Esta transformação tem sido possível também a partir dos avanços recentes em IA. A tecnologia tem inúmeras aplicações que estão mudando a forma como aprendemos, tornando a educação mais acessível para alunos com computadores ou dispositivos inteligentes, seja dentro ou fora do ambiente escolar.

Os alunos não são os únicos que beneficiam, pois a IA também pode potencializar as atividades dos professores, facilitando ou simplificando a visão sobre desempenho, sugerindo e direcionando a tomada de decisão a partir da análise computacional.

De acordo com o relatório do mercado de IA para o setor educacional nos Estados Unidos (US Education Sector, 2018) avaliou-se que o investimento em IA para o sector da educação crescerá 47,5% até 2021 à medida que avançamos para um mundo mais conectado. O impacto da tecnologia existirá em qualquer lugar, do jardim de infância até o ensino superior, oferecendo a oportunidade de criar recursos de aprendizagem adaptativos com ferramentas personalizadas para melhorar a experiência do aluno. A tecnologia de AM, na área de IA, é

projetada de forma a possibilitar a interação direta com os alunos minimizando ou mesmo sem qualquer intervenção humana. A tecnologia de AM, na área de IA, é projetada de forma a possibilitar a interação direta com os alunos minimizando ou mesmo sem qualquer intervenção humana.

Na conhecida plataforma de transmissão de vídeos, Netflix, onde se apresenta ao utilizador um conjunto de recomendações de filmes e séries, estas são geradas a partir de uma avaliação do perfil individual. Segundo a própria definição e apresentação do produto Netflix, a adaptação e personalização do conteúdo ao seu utilizador é parte da estratégia do produto:

> "Usamos diversos algoritmos de AM e recomendamos em grande escala de forma a conduzir as nossas experiências de personalização e pesquisa. Fazemos iterações contínuas para melhorá-los através de experiências offline e testes A/B online. Trabalhamos para impulsionar o estado da arte, aprimorando essas áreas e procurando novas oportunidades para tornar a experiência mais personalizada. A personalização é um dos pilares da Netflix porque permite que cada membro tenha uma visão diferente do nosso conteúdo que se adapta aos seus interesses e pode ajudar a expandi-los ao longo do tempo. Esta perspetiva permite-nos ter não apenas um produto Netflix, mas centenas de milhões de produtos: um para cada perfil de membro." (Chandrashekar et al., 2020, p. 1)

Assim um breve exemplo da técnica aplicada ao Netflix, que além de personalizar o conteúdo também tem melhorado a apresentação desse conteúdo sob a adoção da mesma tecnologia de IA com a abordagem de modelos de AM, surge como motivação, mesmo em contextos e ambientes diferentes, para que estas abordagens sejam investigadas para futura aplicação no MILAGE Aprender+. O principal objetivo é disponibilizar um recurso de apoio ao aluno para este melhorar o seu aproveitamento.

Pretende-se que os alunos tenham e acedam a conteúdos das disciplinas de forma adaptável, aproximada a suas necessidades, para que possam assim ter um percurso de aprendizagem mais objetivo e também agradável (Wong & Oxman, 2014).

1.3 Enquadramento - MILAGE Aprender+

A partir de um projeto iniciado em 2015 e liderado pela Universidade do Algarve, a aplicação MILAGE Aprender+ (MathematIcs bLended Augmented GamE), publicada em 2018, foi alicerçada no conceito de "pedagogia móvel que se baseia na crença de que os dispositivos móveis podem apoiar a aprendizagem autodirigida e a autonomia do aluno"

(Figueiredo et al., 2018). A abordagem visa melhorar o desempenho escolar de todos os alunos já que se trata de uma ferramenta interativa, para o aluno e para o professor, servindo como uma ferramenta de apoio para o ensino remoto e presencial.

Na ferramenta é adotada uma abordagem de gamificação, onde para cada atividade ou tarefa o aluno será avaliado e poderá adquirir pontos, participando assim numa competição com os restantes alunos utilizadores da ferramenta.

Este estudo permitiu identificar uma oportunidade de evolução da ferramenta MILAGE APRENDER+, a fim de proporcionar uma melhor experiência aos seus utilizadores, alunos e professores, com a possibilidade de elevar o nível de adoção e aproveitamento na utilização da aplicação.

1.4 Domínio de Problema

O MILAGE Aprender+ constitui uma oportunidade relacionada com o modelo de disponibilização de conteúdos, como por exemplo fichas de questões e exercícios. O conteúdo disponibilizado ao aluno não é adaptado ou personalizado, ou seja, o conteúdo da aplicação não tem em consideração as necessidades, aproveitamento, progresso e/ou até mesmo os objetivos de cada aluno. Atualmente não é explorada a possibilidade de usar os dados históricos de desempenho, seja de um aluno ou grupo de alunos, para sugerir ou recomendar um conteúdo direcionado e focado nas reais necessidades dos alunos.

No MILAGE Aprender+ temos o professor que toma praticamente todas as decisões sobre a estrutura e a ordem do conteúdo que será atribuído ao aluno. É assumida uma relação de publicação de 1 (um) conteúdo para muitos alunos. Assim, os conteúdos de atividades são apresentados numa condição única e padronizada. Os alunos, como agentes passivos no processo de ensino aprendizagem, acedem ao conteúdo do percurso de aprendizagem, dentro de um roteiro padrão já estabelecido e realizam as atividades na ordem pré-definida e indicada pelo seu professor.

Na ferramenta, o aluno tem a possibilidade de escolher um capítulo, subcapítulo ou ficha de problemas a ser estudada, não existindo nenhum tipo de recomendação de conteúdo para que o aluno tenha alguma referência para uma navegação amigável. Na aplicação não há nenhum tipo de recurso que recomende e guie o aluno no seu estudo com base nos seus pontos ou aproveitamento. Todos os alunos seguem do mesmo ponto de partida planeado pelo professor, e mesmo no processo de utilização da ferramenta, ao longo da trajetória de aprendizagem, o

conteúdo não contempla variações, recomendações e ou adaptações específicas, não se adapta e não é minimamente personalizado de forma automática para cada aluno.

Como consequência deste modelo padronizado e linear, o aluno muitas vezes é exposto a conteúdos que não vão ao encontro das suas necessidades reais, apresentando conteúdo muitas vezes repetitivos ou desnecessários, podendo originar até mesmo o desinteresse pela utilização da ferramenta.

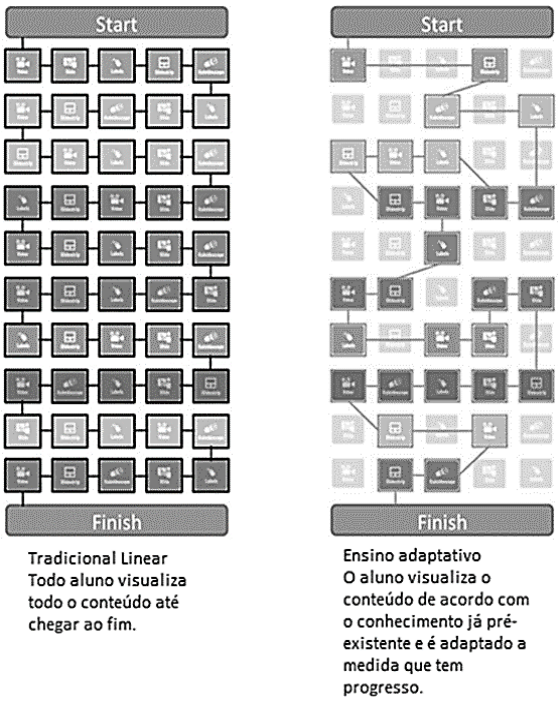


Figura 2 - Ensino tradicional linear e ensino adaptativo (Posner, 2017), adaptado por Felipe Fonseca, 2021)

Em outras palavras, o MILAGE Aprender+ e sua atual tecnologia não utiliza os seus dados e informação para propor conteúdo ao aluno e o professor não tem uma visão de análise ou projeção de resultados em relação ao aproveitamento dos seus alunos.

1.5 Objetivo Geral

Neste trabalho de mestrado pretende-se investigar a possibilidade de, a partir da adoção de IA, aplicar um modelo de AM para que no MILAGE Aprender+ apresente e recomende ao aluno o conteúdo de forma adaptativa e personalizada a partir do seu aproveitamento.

1.6 Objetivos específicos

i. Investigar o MILAGE Aprender+ a fim de identificar a atual estrutura de dados e modelo de funcionamento direcionando para a aplicação de modelos de AM para

oferecer ao aluno e/ou professor recursos que possibilitem um percurso de aprendizagem personalizada a partir da recomendação de conteúdo.

ii. Investigar se o MILAGE Aprender+ atende e/ou preenche os requisitos para a implementação de recursos de AM.

iii. Estudar a viabilidade de utilizar um modelo de AM. Para o ensino adaptativo, pretende-se investigar a viabilidade de um modelo, algoritmo, com a adoção de AM, que possa, a partir dos dados já existentes no MILAGE Aprender+, estimar o aproveitamento dos alunos e assim, a partir de grupos de desempenho e aproveitamento gerar informação para tomada de decisões, como por exemplo a recomendação de conteúdos específicos.

iv. Identificar no MILAGE Aprender+, informação atual e conjunto de dados, que possuam correlação forte, para que sejam aplicados nos modelos de AM.

v. Apresentar indicações e propostas que viabilizem a implementação de modelos de AM no MILAGE Aprender+.

1.7 Contribuições da tese

O trabalho visa alcançar o objetivo de avaliar a ferramenta MILAGE Aprender+ e propor a implementação de modelos de AM no MILAGE Aprender+ para que os beneficiários, ou seja, os utilizadores tanto alunos como professores, possam usufruir de uma aplicação que permita otimizar o percurso de aprendizagem dos alunos e torná-lo mais agradável no âmbito da experiência do utilizador.

Como ponto central, o trabalho encarregar-se-á de aprofundar e avaliar as condições necessárias, obstáculos, meios e recomendações para promover no MILAGE Aprender+ a adoção da AM.

Tratando-se de um primeiro estudo neste campo direcionado em específico ao MILAGE Aprender+, também será estudada a viabilidade da implementação e o respetivo benefício, obstáculos e dependências para a adoção de um modelo de aprendizagem máquina.

1.8 A organização da Tese

Este projeto de mestrado é apresentado em 6 (seis) capítulos: o Capítulo 1 envolve a introdução ao trabalho, com a apresentação do contexto, motivação, enquadramento, objetivos e contribuições do trabalho de mestrado.

No capítulo 2 é apresentado o funcionamento atual do MILAGE APRENDER+, a sua arquitetura tecnológica, e uma perspetiva de utilização a partir dos principais utilizadores (aluno e professor), explorando sobretudo a oportunidade para a adoção de modelos de AM. Também se apresenta a comparação de outras aplicações e ferramentas na área de educação que já aplicam recursos de IA.

O Capítulo 3 traz a revisão da literatura com a fundamentação teórica para o trabalho, com o objetivo de alicerçar e identificar o universo de conceitos de AM, na sua aplicação em problemas do mundo real. Apresentam-se as principais classes de conceitos e fases do processo de adoção.

No capítulo 4 será apresentada a metodologia CRISP-DM com seus conceitos e fases assim como os artefactos realizados nas fases de pesquisa do modelo AM deste trabalho de mestrado. Neste capítulo também serão apresentados os recursos utilizados, a análise de dados e as ferramentas tecnológicas utilizadas para cada uma das tarefas deste trabalho.

Os resultados e a avaliação da aplicação de modelos de AM ao MILAGE Aprender+ serão apresentados no Capítulo 5. Os resultados são apresentados a partir de uma abordagem comparativa entre modelos de AM permitindo uma generalização dos resultados e com mais significativo. Além disso, as limitações e os obstáculos serão identificados neste capítulo juntamente com um conjunto de recomendações.

Por fim, no capítulo 6, serão apresentadas as conclusões que integram as respostas às questões de investigação deste trabalho. O trabalho será concluído com considerações finais para trabalho futuro.

Capítulo 2: O MILAGE Aprender+

2.1 MILAGE Aprender+

A aplicação MILAGE Aprender+ para dispositivos móveis, permite aos alunos acederem a conteúdos pedagógicos, dentro e fora da sala de aula. Os alunos, de forma autónoma, além de realizarem os próprios exercícios, fazem a sua auto-avaliação. Cada exercício possui diferentes níveis de dificuldade. A ferramenta adota ainda a gamificação para estimular e apoiar os alunos na realização das suas atividades. Ao finalizar o exercício o aluno realiza a auto-avaliação comparando com a resposta disponibilizada pelo seu professor. Nesta auto-avaliação o aluno indica o total de pontos obtidos. A ferramenta também possibilita que alunos realizem a avaliação pelos pares, ou seja, um aluno realiza a avaliação de exercícios de outros alunos.

Embora o aluno tenha acesso a uma ampla biblioteca de conteúdo já carregada na aplicação, todo o conteúdo apresentado é planeado e carregado pelo professor, que por sua vez consegue visualizar o desempenho individual dos alunos e assim dedicar o seu tempo aos alunos que apresentem maiores dificuldades.

2.2 Modelo de Funcionamento do MILAGE Aprender+

O atual modelo de funcionamento do MILAGE Aprender+ é apresentado de seguida permitindo perceber a sua arquitetura tecnológica, sobretudo numa macro visão sobre o processo de trabalho para os principais utilizadores, aluno e professor.

A partir de um dispositivo tecnológico, como telemóveis inteligentes, tablet ou até mesmo um computador móvel, o aluno ou professor poderá aceder ao MILAGE Aprender+ através de uma aplicação que pode ser descarregada numa loja de aplicações. Para autenticação na ferramenta, os utilizadores efetuam o acesso com credenciais exclusivas.

Uma vez instalada a ferramenta, há a possibilidade de descarregar conteúdos, facilitando o acesso àqueles que tenham restrições de conexão com a internet, são os chamados e-books que contêm todo o conteúdo e exercícios da disciplina, inclusive vídeos explicativos.

Periodicamente, apresentações para professores e alunos são realizadas e são publicadas na internet para que um público seja instruído no uso da ferramenta.

2.3 O aluno

No ambiente MILAGE Aprender+, o aluno tem acesso à ampla biblioteca de conteúdo de diversas disciplinas e inclusive ao plano de estudos criado pelo seu professor.

O aluno é convidado a realizar os exercícios e atividades, através de fichas de questões, a partir do conteúdo planeado pelo seu professor. O aluno percorrerá o caminho proposto, realizando os respetivos exercícios que podem ser de resposta alternativa ou até mesmo dissertativa.

Todo o percurso de aprendizagem e seu conteúdo não é automaticamente personalizado, recomendado ou adaptado, esta tarefa fica à responsabilidade total do professor.

O processo de validação da resposta pode ser feito pelo professor e pelos seus pares. À medida que o aluno resolve o exercício, carrega a foto da sua resposta, realiza uma autoavaliação comparando com a resposta certa, submete ao professor. A aplicação também permite que um colega avalie a sua ficha de exercício, que é chamada de avaliação pelos pares. O aluno também pode ter acesso à ficha de outros alunos para que possa aprender e verificar como os seus colegas respondem aos exercícios. Neste passo, o aluno também atribui uma pontuação aos exercícios dos seus colegas.

O diagrama de casos de uso (figura 3) apresenta as principais funcionalidades da aplicação para o ator aluno.

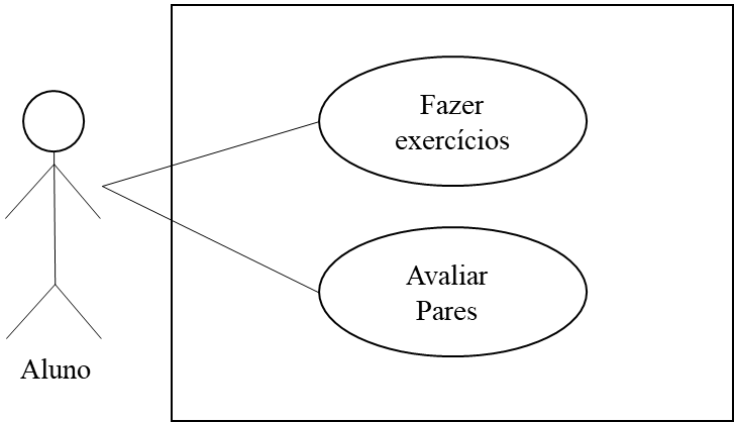


Figura 3 - Diagrama de Caso de Uso (Aluno)

2.4 O professor

O professor tem à disposição uma ferramenta dedicada, denominada MILAGE Aprender+ Professores que permite construir o plano de estudos que o aluno ou grupo de alunos irá realizar. Uma vasta biblioteca de conteúdos está disponível no MILAGE Aprender+, o professor também tem a possibilidade de carregar novo conteúdo como exercícios, textos e vídeos.

O desempenho e progresso do aluno, ou grupo de alunos, pode ser acompanhado para que o processo de avaliação e feedback possa ser realizado. A partir do resultado de desempenho dos alunos, o professor pode, além de interagir com mensagens diretas ao aluno através da própria ferramenta.

O professor tem acesso a fichas de exercícios realizadas e inclusive as fichas que foram avaliadas pelos pares, assim poderá verificar o progresso dos alunos. O professor pode ter uma visão organizada das turmas de alunos, por ano escolar, por escola e ou até mesmo grupo de escolas.

O diagrama de caso de uso (figura 4) apresenta as principais funcionalidades da aplicação para o ator professor.

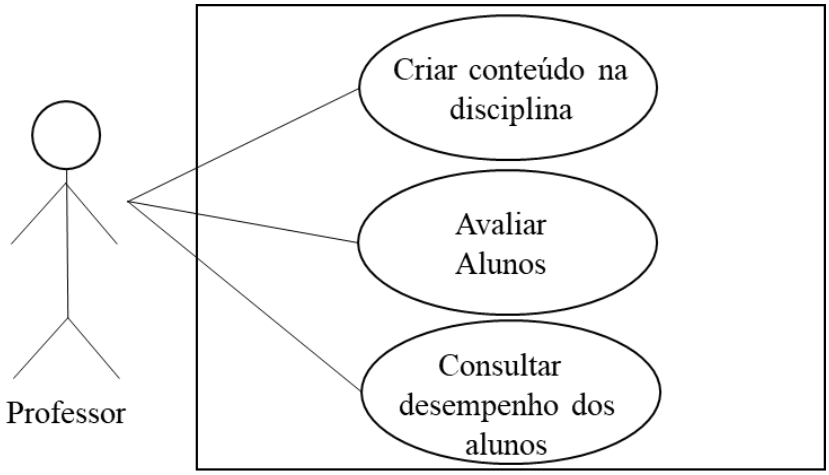


Figura 4 - Diagrama de Caso de Uso (Professor)

A ferramenta abre a possibilidade de criar planos individuais, entretanto todo este plano será construído fora da ferramenta e repassado ao aluno para que, também de forma manual, possa localizar os respetivos exercícios proposto pelo professor. Segundo (Mauro, Beata, & José, 2016) "Acreditamos que isso pode ser motivador para os alunos e também ajuda na realização de palestras, atividades práticas e customização de módulos de estudo". Todavia este plano individual é construído de forma manual e a depender totalmente do professor sem que o MILAGE apoie na apresentação ou recomendação de conteúdos, a partir da identificação dos pontos fracos ou de oportunidades do aluno.

Figura 5 - Tela do MILAGE Aprender+ – (MILAGE Aprender+, 2020)

2.5 Estrutura e arquitetura tecnológica

A estrutura e arquitetura tecnológica do MILAGE Aprender+ foi desenvolvida com uma base de dados relacional para armazenar dados e informações que são geradas pelos alunos e pelos professores. Todo o conjunto de dados e informações são organizados, a partir dos exercícios realizados e/ou carregados pelos professores, num determinado capítulo ou disciplina.

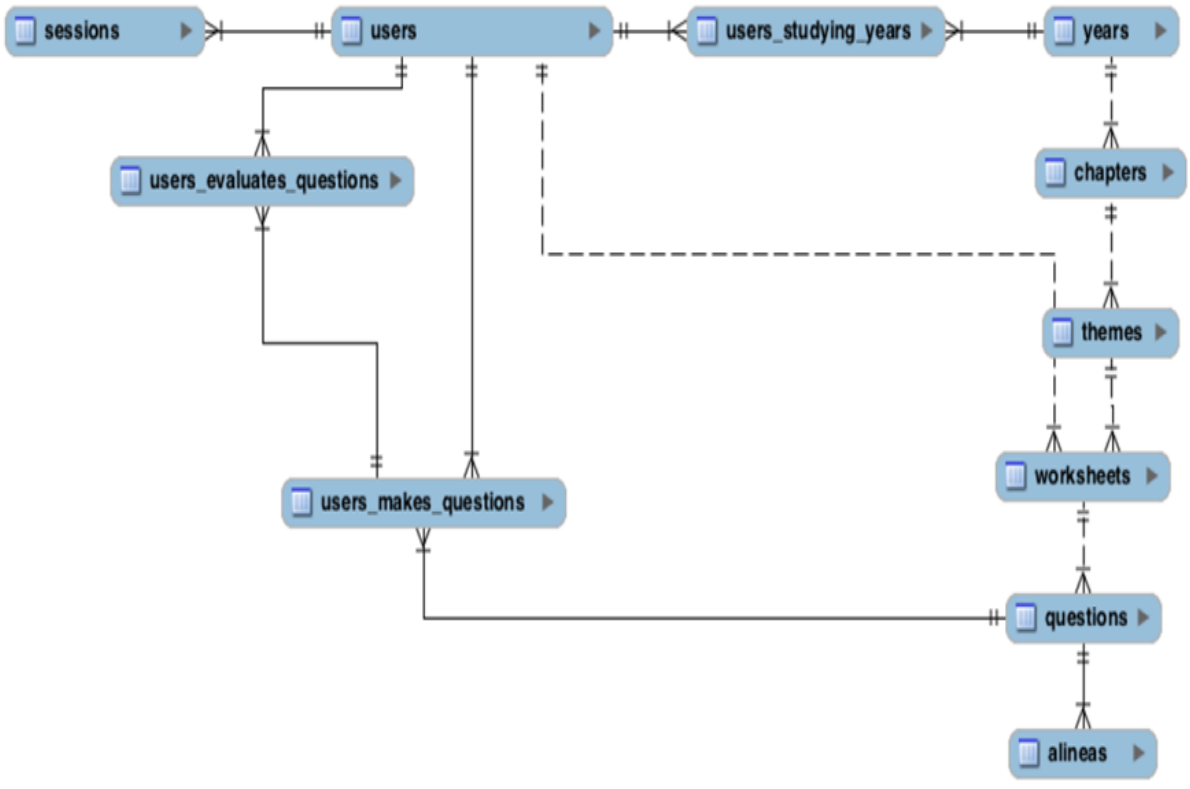


Figura 6 - Modelo relacional dos dados - (Mauro, Beata, & José, 2016)

A ferramenta tem como plataforma um servidor web que é acedido a partir de uma camada de interação, via computadores ou telemóveis inteligentes usando a aplicação MILAGE Aprender+ para aluno e professor. A camada de interação é apresentada através de ecrãs que foram desenhados para permitir o uso intuitivo por parte de todos os utilizadores.

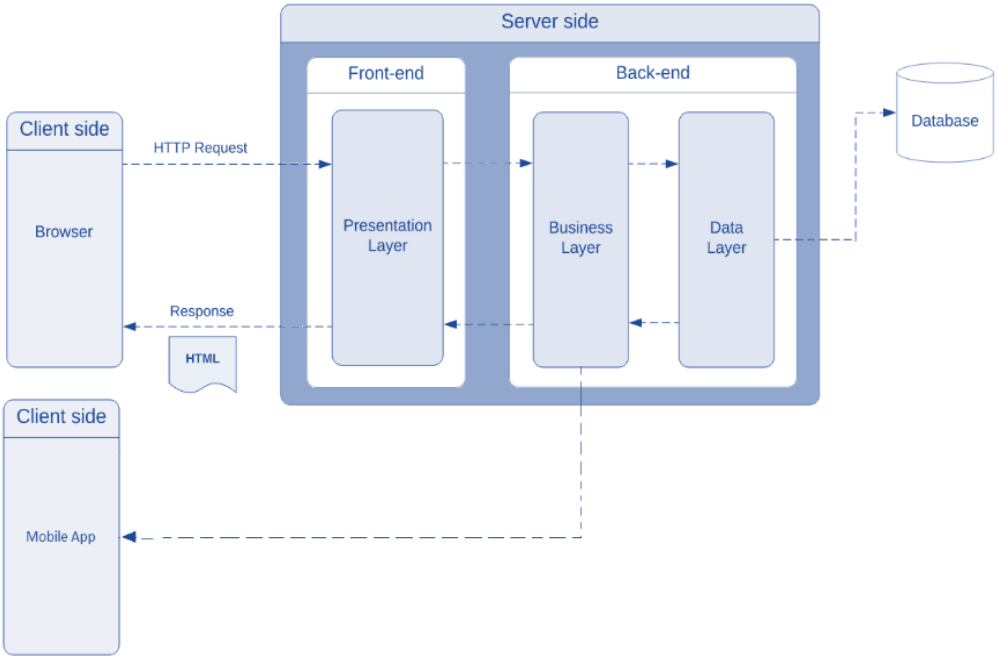


Figura 7 - Modelo de funcionamento – (Felipe Fonseca, 2020)

2.6 Análise comparativa entre os sistemas existentes

Para ampliar o espectro da investigação, a fim de perceber como outras ferramentas de educação à distância evoluem e são disponibilizadas aos alunos, foi realizado uma breve análise de um grupo de ferramentas de educação conforme se apresenta de seguida para servir como um referencial do estado da arte.

A análise concentrou-se a investigar principalmente as características relacionadas a adaptação, personalização e recomendação de conteúdos.

Segue-se uma breve descrição dessas ferramentas:

I. Thirdspace Learning

Projeto iniciado em 2013 orientado para o ensino misto, com uma solução que hoje está concentrada predominantemente na Inglaterra, com mais de 2000 escolas e mais de 60 mil alunos a adotar esta abordagem. A solução que tem como foco conteúdos de matemática, apresenta como característica fundamental o recurso à avaliação de diagnóstico pré e pós cada tópico para que se possam identificar lacunas que permitam personalizar o processo de aprendizagem.

II. Khan Academy

Proposta por uma organização sem fins lucrativos que propõe ensino e educação a distância. O ambiente disponibiliza principalmente conteúdos para Matemática, mas há também para Física, Química e Biologia. Antes de iniciar um novo tópico, o aluno tem a possibilidade de realizar um desafio e responder a um questionário que tem como objetivo posicionar o aluno no percurso de aprendizagem. Assim, para cada tópico é feito uma avaliação anterior e posterior ao tópico realizado, otimizando assim o percurso e possibilitando ao aluno saltar conteúdos já dominados.

III. Coursera

Um portal que reúne conteúdo amplo e cursos de diversas disciplinas. Muito dos conteúdos são oferecidos por grupos parceiros, com aproximadamente 200 Universidades de vários países e empresas de vários setores. Cada curso tem uma dinâmica própria ou modelo de ensino. Nos testes realizados, não foi detetado nenhum curso com verificação ou avaliação prévia de conhecimentos que permitisse apresentação de conteúdo adaptativo, nem mesmo ao longo do progresso do curso. As únicas avaliações apresentadas eram de verificação de progresso, mas o resultado apenas condicionava em habilitar ou não o avanço para o próximo capítulo.

IV. Duolingo

Tem como foco o ensino de idiomas. Contempla um modelo de ensino a partir da trajetória de um jogo, considerando pontuação e ranking de comparação com outros estudantes que são considerados como adversários. De início, aplica uma avaliação que permite detetar o nível de conhecimentos do idioma e à medida que avança é disponibilizado um desafio para que o aluno adquira mais pontos e possa assim avançar para outros níveis e capítulos. Aplica a adaptação de conteúdo desde o início e ao longo do percurso de aprendizagem.

V. Mindspark

Uma ferramenta com foco na matemática para crianças, mas que aplica uma lógica adaptativa ao longo do trajeto da prática de exercícios. Entre erros e acertos, o aluno é avaliado por perguntas distintas sobre o mesmo tópico com o objetivo de confirmar a sua aprendizagem. À medida que alcança um progresso significativo, o que a ferramenta chama de acerto com consistência, vai apresentando novos conteúdos. Percebe-se, de forma muito fácil e simples, que o Mindspark apresenta o conteúdo de perguntas levando em consideração o momento atual

do aluno em relação à aprendizagem do tema aplicando a adaptação de conteúdos seja no início ou no decorrer do percurso de aprendizagem.

VI. Lexplore

Projeto iniciado em 2015, na Suécia, sendo lançado em 2018, é uma ferramenta com foco no ensino de leitura. Captura imagens dos movimentos espontâneos dos olhos durante a leitura do aluno, a partir dos dados obtidos realiza uma análise estatística sobre o comportamento do aluno quanto à sua concentração, velocidade e atividade de leitura. Aplica IA e modelos de AM para sugerir ao professor, de forma adaptativa e personalizada, comparando com outros dados, um conjunto de orientações e direções que o professor de adotar para cada aluno, grupo de alunos, idade escolar, escola.

VII. Aleks

Sistema de avaliação e aprendizagem com IA, baseado na Web. Usa uma abordagem adaptativa para determinar com rapidez e precisão exatamente o que um aluno sabe e não sabe num curso. A ferramenta ALEKS instrui o aluno sobre os tópicos que ele está mais apto para avançar na aprendizagem, promovendo assim um cenário mais favorável e captando a atenção do aluno. Quando o aluno começa é realizada uma avaliação de diagnóstico, esta avaliação torna-se contínua e todo o conteúdo é adaptado de acordo com o seu desempenho, comparando os dados com outros estudantes seja do âmbito de grupo, escola, região, etc.

Como apresentado na tabela 1 que é apresentada de seguida, no sistema Coursera (III) não há recursos adaptativos e/ou personalizados para o conteúdo do curso realizado durante a fase de testes, *Pattern Discovery in Data Mining*. Estes dados foram obtidos a partir de testes que consistiam numa avaliação prática, assumindo o papel de aluno. Nessa avaliação foram criadas dez contas de diferentes utilizadores com perfil de aluno. Observou-se que o sistema Coursera, no curso *Pattern Discovery in Data Mining*, apresentava exatamente a mesma informação, mesmo em casos de progressos distintos de cada aluno. Ou seja, independentemente do avanço e ações realizadas pelos alunos, o conteúdo não era adaptado, modificado especificamente e personalizado para um utilizador aluno.

No entanto, temos de considerar que o Coursera é uma plataforma que conta com muitos cursos publicados a partir de diversas instituições de ensino e até empresas privadas, logo cada curso tem o seu modelo e método de ensino a aplicar. Por fim, mesmo na descrição e anúncio

institucional da plataforma, não há nenhuma menção ou publicação sobre a característica de cursos com conteúdo adaptativo e personalizado.

Nos restantes sistemas, Thirdspace Learning (I), Khan Academy (II), Duolingo (IV), Mindspark (V), Lexplore (VI) e Aleks (VII), usando a mesma prática de testes, criando dez utilizadores de teste com perfil de aluno, foi possível detetar e perceber, a partir das ações de progresso realizada pelos utilizadores alunos, que o conteúdo era apresentado de forma distinta e dinâmica, para cada aluno, resultando assim num processo mais adaptativo e personalizado na apresentação do conteúdo. Todos estes sistemas incluem ainda, nas suas páginas institucionais ou de apresentação, uma publicação das suas características dando ênfase ao ensino com conteúdo adaptativo e personalizado.

Um ponto relevante ao longo do desenvolvimento desta investigação foi a constatação da evolução de uma das ferramentas analisadas. Em maio de 2020, na ferramenta Khan Academy, não foram identificados recursos de avaliação prévia para oferecer a adaptação de conteúdo. Porém, em novembro de 2020 foi detetado a inserção desta funcionalidade.

Como podemos observar na Tabela I abaixo, a abordagem personalizada, adaptativa, com recomendação de conteúdo ao aluno, tem sido uma prática real para alguns sistemas renomeados e bem avaliados.

William e Steven (2014) afirmam que não adaptar o conteúdo a partir do contexto, ignorando conhecimento prévio do aluno ou não levando em consideração o seu desempenho e progresso ao longo do percurso de aprendizagem, poderá estar relacionado com um resultado de aprendizagem menos eficiente.

Esta análise do estado da arte que consistiu na comparação de vários sistemas, anteriormente mencionados com o MILAGE Aprender+, permitiu identificar lacunas e oportunidades, que apoiam e justificam o trabalho de investigação proposto. Como anteriormente mencionámos, no MILAGE Aprender+ não há uma recomendação, adaptação da apresentação do conteúdo de forma personalizada ao seu aluno.

Neste enquadramento, é válido acrescentar também que o MILAGE Aprender+ é uma ferramenta em difusão e com notório reconhecimento nos alunos que aderiram à plataforma. Numa publicação do Jornal Expresso (2020), foi noticiado que a tecnologia “já é utilizada por 50 mil alunos, não só em Portugal como em Espanha, Chipre, Alemanha, Noruega ou Turquia”. Ou seja, o MILAGE é uma ferramenta que dia após dia tem ampliado a sua presença, em diversas escolas portuguesas, mas também noutros países.

A tabela 1 que se segue, apresenta outras ferramentas, com tecnologias de apoio ao ensino e com propósito equivalente ao MILAGE Aprender+.

id.	Plataforma de Educação a distância	Foco da plataforma	Endereço de referência (URL)	Curso escolhido para pesquisa	Aplica recurso de IA para a apresentação de conteúdo
I	Thirdspace Learning	Foco em Matemática	https://thirdspacelearning.com/	Estatística	Sim
S	Khan Academy	Há disciplinas como Biologia, Química e Física, mas é percebido um maior volume de conteúdo na Matemática	https://pt-pt.khanacademy.org/	Probabilidades e estatística	Sim
III	Coursera	A um leque bem diverso de cursos de várias áreas e disciplinas.	https://pt.coursera.org/	Pattern Discovery in Data Mining	Não
IV	Duolingo	Foco em Idiomas	https://pt.duolingo.com/	Inglês	Sim
V	Mindspark	Foco em Matemática e aulas de Inglês.	https://www.mindspark.in/	Counting numbers	Sim
VI	Lexplore	Ensino a leitura	https://www.lexplore.com/	Inglês	Sim
VII	Aleks	Predominantemente Matemática	https://www.aleks.com/	Matemática	Sim

Tabela 1 - Comparativo de ferramentas de ensino a distância.

É neste contexto e representatividade relacionado com o MILAGE Aprender+, também considerando o recente crescimento da procura e necessidade pelo ensino remoto e à distância, que se identifica a oportunidade de investigação que tem como objetivo principal melhorar e evoluir a aplicação MILAGE Aprender+, recorrendo à IA e adotando modelos de AM para viabilizar a oferta de ensino adaptativo e personalizado.

Desta forma, usando como referência os casos analisados no estado da arte, a investigação concentrar-se-á em potencializar a capacidade de tomada de decisão do professor, direcionando o foco para estudantes que possuam desempenho e aproveitamento abaixo do esperado, permitindo acompanhar e disponibilizar conteúdos tendo em consideração as forças e fraquezas de cada aluno, ou seja, uma combinação a acrescentar no processo de aprendizagem. Para o aluno pretende-se potencializar a ferramenta com recursos de recomendação de conteúdo a partir do seu aproveitamento no MILAGE Aprender+.

Capítulo 3: Revisão da Literatura

3.1 Aprendizagem Máquina

O termo Aprendizagem Máquina (AM), traduzido do inglês *Machine Learning*, foi definido como um campo de estudo que fornece capacidade de aprendizagem ao computador sem ser explicitamente programado (Samuel, 1959).

Numa das definições mais conhecidas para AM, o autor (Mitchell, 1997) apresentou uma definição mais apropriada ao significado de AM, “Ao programa de computador é dito para aprender com a experiência *E* relacionada com algumas classes de tarefas *T* e medidas de desempenho *P*, se o seu desempenho nas tarefas T, medida por P, melhora com a experiência E”.

Conforme a tabela que se segue, desde 1950 que a área de IA tem sido pesquisada, desenvolvida e melhorada (Alzubi et al., 2018).

1950	Alan Turing criou o *"Turing Test"* para verificar a inteligência de uma máquina. Para passar no *“Turing Test”*, a máquina deveria ser capaz de convencer os humanos de que estão realmente falando com um humano e não com uma máquina
1952	Samuel criou um algoritmo de aprendizagem, altamente capacitado, que podia jogar o jogo das damas consigo mesmo e realizar auto-treino.
1956	Martin Minsky,John McCarty, Claude Shannon e Nathan Rochester organizaram uma conferência em Dartmouth em 1956, onde realmente nasceu a Inteligência Artificial.
1958	Frank Rosenblatt criou o *Perceptron*, que lançou a pedra fundamental para o desenvolvimento da Rede Neuronal Artificial (RNA).
1967	O algoritmo do vizinho mais próximo foi proposto, o qual poderia ser usado para "Reconhecimento de padrões".
1979	Os alunos da Universidade de Stanford desenvolveram o *"Stanford Cart"*, um robô sofisticado que podia navegar por uma sala e evitar obstáculos no seu caminho.
1981	O *Explanation Based Learning (EBL)* foi proposto por Gerald Dejong, segundo o qual um computador pode analisar os dados de treino e criar regras para remover dados inúteis.
1985	O *NetTalk* foi inventado por Terry Sejnowki, o qual aprendeu a pronunciar palavras em inglês da mesma forma que as crianças aprendem.
The 1990s	O foco da Aprendizagem Máquina mudou de Orientada ao Conhecimento para Orientada a Dados. A AM foi implementada para analisar grandes blocos de dados e tirar conclusões a partir deles.
1997	A IBM inventou o computador *Deep Blue* que venceu o campeão mundial de xadrez Gary Kasparov.

2006	O termo *"Deep Learning"* foi aplicado por Geoffrey Hinton, o qual consiste numa nova arquitetura de redes neurais que usava várias camadas de neurônios para a aprendizagem profunda.
2011	O *Watson* da IBM, construído para responder a perguntas feitas em linguagem natural, derrota um concorrente humano no jogo Jeopardy
2012	Jeff Dean, do Google, desenvolveu o GoogleBrain, que é uma *Deep Neural Network* para detetar padrões em Vídeos e Imagens.
2014	O Facebook inventou o algoritmo *"DeepFace"* baseado em *Deep Neural Network* capaz de reconhecer rostos humanos em fotos.
2015	A Amazon propôs a sua própria plataforma de AM. A Microsoft criou o *"Distributed Machine Learning Toolkit"* para resolução de problemas de AM através do poder computacional em nuvem (*cloud computing*). Elon Musk e Sam Altman, criaram uma organização sem fins lucrativos - OoeaAI, com o objetivo de utilizar a IA para servir o ser humano.
2016	A Google propôs o *DeepMind*, considerado o jogo de tabuleiro mais complexo. O programa *Google AlphaGo* torna-se o primeiro programa *Computer Go* a vencer um jogador humano profissional. Este baseia-se na combinação de técnicas de AM e pesquisa em árvores de decisão.
2017	A Google propôs telefones baseados em *Google Lens, Google Clicks, Google Home Mini e Google Nexus* que usam AM e Algoritmos de Aprendizagem Profunda. A Nvidia propôs *GPUs NVIDIA - The Engine of Deep Learning*. A Apple propôs o *Home Pod*, que é um dispositivo interativo de AM.

Tabela 2 - Evolução da AM (Alzubi et al., 2018).

A AM é uma subárea da IA dedicada a algoritmos e técnicas que permitem ao computador aprender e possibilita a resolução de problemas complexos, isto é, permite ao computador aperfeiçoar o seu desempenho na realização de alguma tarefa.

Recorrendo aos modelos estatísticos, o principal objetivo da AM é entender a estrutura dos dados, aplicando distribuições teóricas em dados bem compreendidos[1]. Os modelos de AM são modelos estatísticos que podem ser comprovados matematicamente, mas estes modelos de AM requerem que os dados também sigam determinados pressupostos de qualidade para que possam sustentar e apoiar as respostas dos modelos matemáticos. Dados com qualidade são fundamentais para qualquer compromisso de ciência de dados.

[1] Compreensão dos dados (traduzido de *Data Understanding*) abrange atividades de construção de conjunto de dados e sua qualidade. Cada problema de análise de negócios tem requisitos de dados específicos e diferentes padrões podem ser aplicados. A compreensão de dados, da ciência de dados, responde à pergunta: Os dados recolhidos representam o problema a ser resolvido? Existem duas fases principais de compreensão de dados: a avaliação de dados e a exploração de dados.

Para os autores (Alzubi et al., 2018) um dos objetivos da AM é permitir que os computadores executem a tarefa de forma sofisticada, sem qualquer intervenção de seres humanos com base na aprendizagem e no aumento constante da experiência para compreender a complexidade do problema e a necessidade de adaptação.

A AM foi desenvolvida recorrendo à capacidade dos computadores para examinar a estrutura dos dados, mesmo se não soubermos determinar manualmente essa estrutura. Como a AM geralmente usa uma abordagem iterativa para aprender com os dados, a aprendizagem pode ser facilmente automatizada. Para (Brownlee, 2016) os algoritmos de AM funcionam para estimar a função de mapeamento (f) das variáveis de saída (y), dadas as variáveis de entrada (x), ou seja, $Y = f(X)$.

É salutar reforçar que não existe um modelo pronto, ou modelo já pré-concebido e estabelecido para resolver todos os problemas por antecipação. Contudo, podem ser utilizados modelos já pré-estabelecidos como ponto de partida. Estes também podem ser combinados com outros modelos sucessivamente até obter a melhor generalização[2] dos dados. Há que trabalhar diferentes combinações de algoritmos, realizar experiência sobre a estrutura de dados, só assim saberemos qual será a melhor abordagem, qual funcionará melhor como modelo para estimar a função (f), com conjunto de variáveis (X) para melhor prever e estimar o resultado de saída (Y).

Ao analisar um grande volume de informação, os algoritmos de AM podem ser capazes de identificar padrões e de tomar decisões com o auxílio dos modelos. Ou seja, as máquinas serão capazes de fazer previsões através do processamento de dados.

Ainda a respeito dos algoritmos de AM o autor (Brownlee, 2016) divide-os em dois grupos, os paramétricos ou não paramétricos. Os paramétricos fazem grandes suposições sobre o mapeamento de variáveis de entrada (X) para a variável de saída (Y) e geralmente requerem menos dados e têm maior agilidade de resposta. Já os não-paramétricos têm características opostas, fazendo menor suposição e tendo uma exigência de mais dados, pelo que se tornam mais lentos para treino, mas resultam em modelos com melhor desempenho na previsão.

Para entender a AM é também preciso conhecer as três principais técnicas, aprendizagem supervisionada, não-supervisionada e aprendizagem por reforço.

[2] A capacidade de fazer previsões bem sucedidas com dados não observados a partir dos dados observados (dados de treino) é chamada de generalização, ou seja, reunir numa classe geral, termo ou proposição, um conjunto de seres ou fenômenos similares. O objetivo fundamental da AM é generalizar além dos exemplos no conjunto de treino.

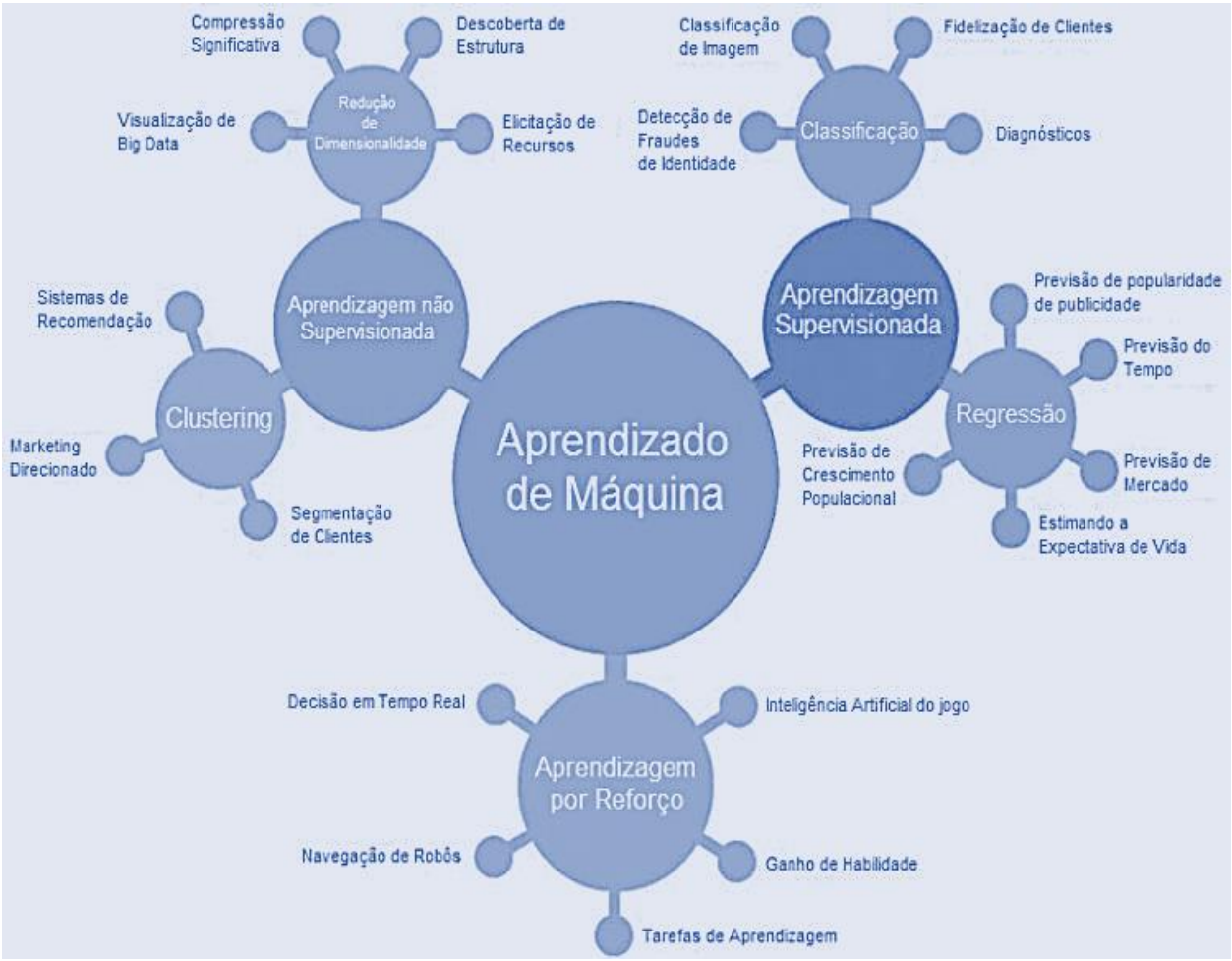


Figura 8 - AM (Jafar Alzubi, Anand Nayyar, & Akshi Kumar, 2018), adaptado por Felipe Fonseca.

3.2 Aprendizagem Supervisionada, Não-Supervisionada e por Reforço.

A AM tem uma classe de algoritmos orientado aos dados, ou seja, contrariamente aos algoritmos "normais", são os dados que "informam" qual é a "melhor resposta". Identificar o tipo e características do problema relacionado é fundamental para o processo de obtenção da melhor resposta (Y). Existem 3 tipos de AM que se baseiam na forma como os algoritmos são criados. Eles são categorizados em supervisionada, não-supervisionada e por reforço.

3.3 Aprendizagem Supervisionada

Esta categoria baseia-se num conjunto de dados e grupo de variáveis de entrada (X) e de exemplos de respostas já classificadas e conhecidas (Y). Com a resposta (Y) previamente identificada e servindo como guia de referência, o algoritmo na sua fase de treino passa a aprender a atribuir uma função (f) que se aproxime e generalize com menor custo de erro. Com novos dados (X), já na fase de teste, poderá prever as variáveis de saída (Y) usando a função aprendida (f) com alto nível de aproximação e menor erro. Ou seja, no modelo Supervisionado,

são apresentados um conjunto de dados já com a respostas e previamente rotulados (Y), para que estes sirvam como exemplo de alvo para o processo de aprendizagem ao mapear uma função (f).

Para (Brownlee, 2016) aprendizagem supervisionada pode ainda ser organizada em problemas caracterizados como:

- Classificação: um problema de classificação ocorre quando a variável de saída (Y) é uma variável categórica como por exemplo um resultado positivo ou negativo para uma doença, uma classificação entre aprovado ou reprovado, spam[3] ou não é spam, fraude ou não-fraude, vai chover ou não vai chover, etc. O objetivo é atribuir uma função de saída que prevê duas ou mais classes.

- Regressão: Um problema de regressão ocorre quando a variável de saída (Y) é um valor numérico real, como por exemplo na previsão pontos num campeonato, o aproveitamento numa prova escolar, valor de uma moeda, valor de venda de uma casa, temperatura, probabilidade de chuva, etc.

3.4 Algoritmos de Aprendizagem Supervisionada

De seguida, são apresentados alguns dos algoritmos mais utilizados dentro da categoria de Aprendizagem Supervisionada da AM (IBM Cloud Education, 2020);

- Redes neuronais: Aplicadas principalmente para aprendizagem profunda. As redes neurais processam dados de treino imitando a interconectividade do cérebro humano através de camadas de nós. Cada nó é composto de entradas, pesos, uma tendência (ou limite) e uma saída. Se esse valor de saída exceder um determinado limite, ele "dispara" ou ativa o nó, passando os dados para a próxima camada da rede. As redes neurais aprendem essa função de mapeamento ajustando com base na função de perda e através do método Gradiente Descendente. Quando a função de custo está igual ou próxima de zero, podemos confiar na precisão do modelo para gerar a resposta correta.

[3] O SPAM da sigla, Enviar e Postar Publicidade em Massa (traduzido do inglês Sending and Posting Advertisement in Mass) é uma mensagem eletrônica que chega ao usuário sem a sua permissão ou sem seu desejo em recebê-lo.

- Naive Bayes: É uma abordagem de classificação que adota o princípio da independência condicional de classe do Teorema de Bayes (1701-1761). Significa que a presença de um recurso não tem impacto na presença de outro recurso em relação à probabilidade de um determinado resultado, e cada previsor tem um efeito igual naquele resultado. Existem três tipos de classificadores Naïve Bayes: Multinomial Naïve Bayes, Bernoulli Naïve Bayes e Gaussian Naïve Bayes. Esta técnica é usada principalmente em sistemas de classificação de texto, identificação de spam e recomendação.

- Regressão linear: A regressão linear é usada para identificar a relação entre uma variável dependente e uma ou mais variáveis independentes e normalmente é aplicada na previsão de resultados futuros. Quando há apenas uma variável independente e uma variável dependente, é conhecida como regressão linear simples. À medida que o número de variáveis independentes aumenta, é conhecida como regressão linear múltipla. Para cada tipo de regressão linear, procura-se traçar uma linha de melhor ajuste, que é calculada pelo método dos mínimos quadrados. No entanto, ao contrário dos outros modelos de regressão, esta linha é reta quando se traça um gráfico.

- Regressão logística: Enquanto a regressão linear é aplicada quando as variáveis dependentes são contínuas, a regressão logística é selecionada quando a variável dependente é categórica, significa que têm saídas binárias, como "verdadeiro" e "falso" ou "sim" e "não". Embora ambos os modelos de regressão procurem entender as relações entre as entradas de dados, a regressão logística é usada principalmente para resolver problemas de classificação binária, como a identificação de spam.

- Máquina de vetores de suporte (SVM *Support Vector Machine*): um modelo popular de aprendizagem supervisionada desenvolvida por Vladimir Vapnik, usado para classificação e regressão de dados. Normalmente é aplicado em problemas de classificação, construindo um hiperplano onde a distância entre duas classes de pontos de dados está no seu máximo. Esse hiperplano é

conhecido como limite de decisão, separando as classes de pontos de dados (por exemplo, laranjas vs. maçãs) em ambos os lados do plano.

- K-Vizinho mais próximo (KNN *K-nearest neighbors*): É um algoritmo não paramétrico que classifica os dados com base na sua proximidade e associação com outros dados disponíveis. Este algoritmo assume que pontos de dados semelhantes podem ser encontrados próximos uns dos outros. Como resultado, procura calcular a distância entre os dados geralmente através da distância euclidiana, e então atribui uma categoria com base na média mais frequente. A facilidade de uso e baixo tempo de cálculo tornam-no no algoritmo preferido dos cientistas de dados. Porém, quando o conjunto dos dados de teste cresce, o tempo de processamento aumenta, tornando-o menos atraente para tarefas de classificação. O KNN é normalmente usado para mecanismos de recomendação e reconhecimento de imagem.

- Floresta aleatória (*Random Forest , RF*): outro algoritmo flexível de aprendizagem máquina supervisionada usada para fins de classificação e regressão. A "floresta" faz referência a uma coleção de árvores de decisão não correlacionadas, que são combinadas para reduzir a variabilidade e criar previsões de dados com maior precisão.

3.5 Aprendizagem Não-Supervisionada

Ao contrário da Aprendizagem Supervisionada que possui dados já classificados ou com a resposta corretas (Y), na Aprendizagem Não-Supervisionada temos apenas os dados (X) sem respostas. Logo a Aprendizagem Não-Supervisionada tem como objetivo realizar uma descoberta para perceber e entender o conjunto de dados e apresentar a estrutura e padrão dos dados (Brownlee, 2016). No modelo de Aprendizagem Não-Supervisionada não há o que chamamos de professor, o modelo de aprendizagem encarregar-se-á de encontrar uma resposta e padrões a partir do conjunto de dados apresentado. Os modelos de aprendizagem Não-Supervisionada também se dividem em dois tipos, sendo eles:

- Agrupamento (*Clustering*): Caracteriza-se como um problema de *Clustering* quando se tem como objetivo a descoberta dos agrupamentos inerentes e relacionado aos dados, como por exemplo agrupar por desempenho os

jogadores de ténis a partir do seu desempenho nos jogos durante um campeonato (Brownlee, 2016).

- Associação: Também a partir do conjunto de dados não classificados, este método pretende descobrir regras ou padrões que evidenciam as características de um conjunto de dados. Por exemplo, uma rede de supermercados pretende verificar a relação entre a compra de um produto X e a tendência de compra do produto Y (Brownlee, 2016).

3.6 Aprendizagem por Reforço

A aprendizagem por reforço é um modelo de aprendizagem comportamental (Johnson, 2021). Com a resposta do algoritmo, usa-se essa resposta da análise de dados, orientando o utilizador para o melhor resultado. A aprendizagem por reforço difere dos outros tipos de aprendizagem supervisionada porque o sistema não é treinado com um conjunto de dados de amostra, traduz-se na prática por uma permanente interação com o ambiente através de políticas de tentativa e erro que, maximizando um sinal de reforço, permite atingir um conjunto de objetivos pretendidos. O modelo desenvolve-se e aprende através de tentativa e erros. Portanto, a partir de um uma sequência de acertos, com decisões bem-sucedidas, atribuirá uma função que continuamente será reforçada e melhorada.

Esta abordagem é geralmente aplicada nos jogos de computador. O modelo de Aprendizagem por Reforço usa o treino de modelos de AM para processar uma sequência de ações e decisões. O modelo permitirá atingir um objetivo num ambiente incerto e potencialmente complexo. Com o objetivo de maximizar o ganho, o algoritmo recebe recompensas ou penalidades pelas ações que executa.

Existem três abordagens para implementar um algoritmo de Aprendizagem por Reforço;

- Baseado em valor: envolve maximizar uma função de valor V (s). Nesse método, o agente espera um retorno de longo prazo dos estados atuais sob a política π.

- Baseado em políticas: envolve criar uma política de forma que a ação realizada em cada estado ajude a obter o máximo de recompensa no futuro. Uma política pode ser determinística ou estocástica. Uma política determinística, a = π (s), é essencialmente um mapeamento direto de um estado para uma ação,

determinando a ação (a) que deve ser tomada por encontrar-se num estado qualquer (s). Já uma política estocástica, a = π (a|s) = P(A = a|s = s), determina a probabilidade de um agente escolher uma ação (a) dado que o agente se encontra no estado (s).

- Baseado em modelos: envolve criar um modelo virtual para cada ambiente. O agente aprende a atuar naquele ambiente específico.

Como exemplo de algoritmos para Aprendizagem por Reforço temos:

- Processo de decisão de Markov *(Markov Decision Process):* A abordagem matemática para mapear uma solução em Aprendizagem por Reforço é reconhecida como um Processo de Decisão de Markov ou (MDP). Geralmente o MDP é utilizado com o propósito de modelar situações onde é necessário executar ações de tomada de decisões sequenciais em ambientes com incerteza. O principal objetivo é encontrar uma função que especifique uma boa ação para tomar a decisão em cada tipo de estado. Sempre com o objetivo de maximizar as recompensas atribuídas para cada tomada de decisão (Puterman, 2014).

- Aprendizagem-Q *(Q-learning):* É o algoritmo mais comum para a Aprendizagem por Reforço pois é um modelo que não tentará compreender todo o ambiente, mas, em vez disso, seguirá uma abordagem de política e recompensa para cada decisão tomada e baseada na qualidade da ação. É um algoritmo sem modelo, ou seja, estima a sua política de otimização sem estimar toda a dinâmica e condições do ambiente (Johnson, 2021).

3.7 Análise dos três tipos de AM

A partir dos três tipos de Aprendizagem, de seguida será apresentado uma tabela resumo com os tipos, *I. Supervisionada, II. Não Supervisionada e III. Por reforço*. A tabela ainda relaciona algumas características chaves para cada tipo de dados, domínio (problema), e os respetivos principais algoritmos.

	Característica	Alvo de resposta	Objetivo	Entrada dos dados	Tipo de Processo	Principais Algoritmos	Domínio
I. Supervisionada	Os dados de entrada são previamente rotulados, com respostas previamente já conhecidas e classificadas. Serve como referência na fase de treino. Orientado por tarefa.	Mapeamento (Prever ou estimar o próximo valor).	O objetivo da aprendizagem supervisionada é estudar muitos exemplos rotulados como estes, e depois ser capaz de fazer previsões sobre futuros pontos de dados.	Dados categóricos / qualitativos	Classificação	KNN	Deteção de Fraudes, Detetar *Spam*, Classificação de imagens, etc.
						Naive Bayes	
						Árvore de Decisão	
						Redes Neurais	
						SVM	
						Regressão Logística	
				Valores contínuos / quantitativos	Regressão	Polinomial	Avaliação de risco, prever resultados numéricos, prever preços, pontos, etc.
						Árvore de Decisão	
						Floresta Aleatória	
						Redes Neurais	
						Regressão Linear	
II. Não-Supervisionada	Os dados não têm rótulos associados. O objetivo é organizar os dados de alguma forma ou descrever a sua estrutura. Orientado aos dados.	Prever Classes e Grupos de informação	Agrupa dados em clusters, ou encontra diferentes formas de analisar dados complexos para que pareçam mais simples.	Não há dados de resposta / alvo	Redução de dimensionalidade	PCA	Identificar *fake news*
						SVD	
					Associação ou Agrupamento	K-means	Segmentação de Clientes, Perfis de utilizadores, etc.
						Redes Neurais	
						Hierarquia	
						Modelos Fuzzy	
III. por Reforço	Na aprendizagem de reforço, o algoritmo pode escolher uma ação em resposta a cada ponto de dados	Estado/Ação (Aprender com os erros)	O modelo desenvolve-se e aprende através de tentativa e erros. A partir de um uma sequência de decisões bem-sucedidas, atribuirá uma função que continuamente será reforçada e melhorada.	Dados categóricos / qualitativos	Classificação	Processos de decisão de Markov	Tomada de decisão em tempo real, Jogos *(Gaming)*, Robótica, Carros autônomos, etc.
				Não há dados de resposta / alvo	Controlo	Aprendizagem-Q	Carros autônomos

Tabela 3 – Resumo da análise dos tipos e modelos de Aprendizagem Máquina.

A partir da análise dos modelos apresentados na tabela 3, é possível chegar a conclusão de que para cada tipo de problema e objetivo, sobretudo para cada tipo de dados há um tipo de estratégia de modelo a ser adotado. Embora possamos identificar que alguns algoritmos, como Redes Neurais, possa ser aplicado para mais de um tipo de cenário e problema, é percebido a relevância de ter uma visão holística, e não somente do modelo ou algoritmo somente. Ter uma visão clara sobre o ambiente e domínio, qual a disponibilidade de dados e suas características

(por exemplo, dados contínuos ou categóricos), é fundamental para avançar e progredir na utilização ou extensão de modelos já existentes.

3.8 MILAGE Apender+ e a Aprendizagem Máquina

Com o domínio do problema do MILAGE Aprender+ apresentado no capítulo 2 e a partir do potencial das tecnologias de AM apresentados neste capítulo 3, constata-se a oportunidade de aplicação da AM para resolver os problemas detetados ao MILAGE Aprender+. A tabela que se apresenta de seguida, relaciona alguns problemas do MILAGE Aprender+ com as potenciais alternativas de resolução a partir da adoção da AM.

Problema (Oportunidade)	**Alternativa identificada para resolução com AM**
i. O MILAGE Aprender+ apresenta seu conteúdo de forma linear e padronizada ao seu utilizador e não adota nenhum tipo de recomendação de conteúdo.	Pretende-se investigar um modelo de AM que possa recomendar ao aluno, um conteúdo baseado na projeção de ganho (pontuação) e classificação de desempenho a partir dos dados históricos do MILAGE Aprender.
ii. O MILAGE Aprender+ não apresenta ao utilizador Professor uma projeção de aproveitamento ou desempenho a partir do resultado de pontos dos alunos.	A partir do resultado histórico do aluno investigar um modelo de AM que possa estimar e prever o resultado de pontos e classificação do aluno para conteúdo ainda não realizado.

Tabela 4 – A oportunidade identificada no MILAGE Aprender+.

Considerando o tipo de problema exposto na tabela 4 e aplicando os conceitos apresentados no Capítulo de revisão da literatura a respeito de AM e possível assumir alguns pressupostos para o trabalho:

- A técnica de AM que será aplicada no âmbito deste projeto será de Aprendizagem Supervisionada uma vez que os dados já estão rotulados, ou seja, já contêm respostas identificadas. No MILAGE Aprender+, todas as atividades realizadas pelos alunos já possuem um resultado de pontos atribuídos.
- A partir dos pontos, dado histórico do MILAGE Aprender+, considerando que os dados de pontos são do tipo contínuo (numérico real), aprofundaremos a utilização dos modelos de regressão.

- A partir dos dados de desempenho avaliaremos a possibilidade de utilização de modelos de classificação a partir de variáveis categóricas que poderão ser construídas para definir o nível de classificação dos utilizadores.

Considerando os pressupostos anteriormente apresentados, podemos aprofundar sobre o âmbito do trabalho a realizar no MILAGE Aprender+ no que diz respeito à adoção da AM:

i. **Pretende-se apresentar ao utilizador do MILAGE Aprender+ o conteúdo de forma personalizada;**

   A apresentação de conteúdo ao aluno passa pela utilização de um modelo para prever Temas (conjunto de exercícios) que possam potencializar o seu desempenho por pontos a partir do histórico de exercícios realizados por outros utilizadores do MILAGE Aprender+. Os alunos podem assim receber a recomendação de conteúdos que potencializem seus pontos dentro da aplicação dentro do contexto de gamificação.

   Pretende-se ainda apresentar a recomendação de conteúdo a partir do nível de classificação de desempenho do aluno.

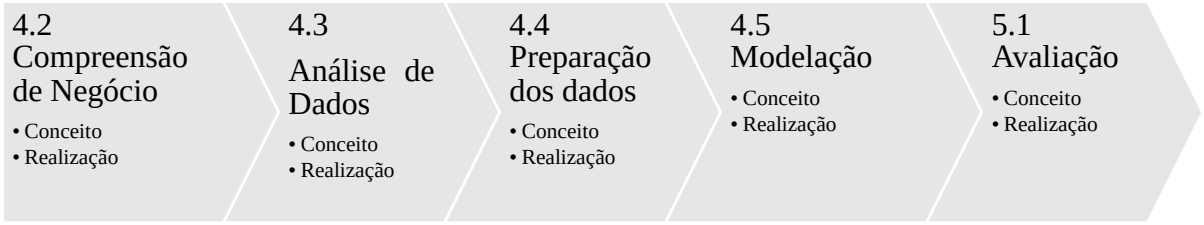
ii. **Pretende-se apresentar aos utilizadores, principalmente aos professores, uma perspetiva com previsão do aproveitamento do utilizador aluno a partir do dado histórico;**

   A partir da projeção e previsão de pontos dos alunos, o modelo deverá projetar o desempenho por pontos de cada aluno, facilitando assim ao utilizador professor o conhecimento sobre a trajetória de aprendizagem do aluno. Os professores poderão ter a visão antecipada sobre o desempenho de pontos a partir dos dados correntes do aluno permitindo assim que o MILAGE Aprender+ auxilie o professor no processo de decisão e suporte ao aluno.

Capítulo 4: Metodologia e as realizações da tese

Neste capítulo 4 será apresentada a metodologia identificada para o projeto. Cada fase da metodologia será apresentada com a definição do conceito teórico e sucessivamente com as respetivas realizações de cada fase da metodologia aplicada ao projeto. Assim, conceito e prática serão apresentados em simultâneo para que seja possível a melhor interpretação da realização deste projeto de mestrado.

Para implementar os requisitos identificados para este projeto de mestrado optou-se por seguir uma metodologia. Constatou-se que a metodologia CRISP-DM (Cross Industry Standard Process for Data Mining) alinha-se com o desenvolvimento deste trabalho pois o principal objetivo, além da apresentação fundamentada e teórica, é a produção de artefactos e experiências para que os resultados possam ser avaliados e melhorados iterativamente.



4.1 CRISP-DM

CRISP-DM é reconhecida como a metodologia mais comum para projetos de prospeção de dados (*Data-Mining*) e análise em ciência de dados, é utilizada desde a sua publicação em 1999, a metodologia apresenta seis fases que traduzem o ciclo de vida de um projeto de ciência de dados (Hotz & Saltz, 2021).

Quinze anos após a publicação do guia CRISP-DM sobre a popularidade de metodologias e abordagens, numa pesquisa de mercado realizada em 2014 pela KDnuggets constatou-se que a metodologia CRISP encontra-se à frente de todas as outras metodologias e abordagens (Piatetsky, 2014). CRISP-DM é uma metodologia que contou com a contribuição de mais de 300 organizações para o seu desenvolvimento. Envolveu a criação de um grupo de trabalho, dando origem a um consórcio com o acrónimo CRISP, do inglês (Cross-Industry Standard Process for Data Mining). (Chapman, et al., 2000)

Um dos principais benefícios desta metodologia é a adaptabilidade da metodologia às necessidades e expectativas deste projeto de mestrado. Nesse sentido, após a fundamentação teórica, em cada fase, apresenta-se também a respetiva componente prática. As várias iterações

do projeto, em cada uma das suas fases, permitiram avaliar os resultados das experiências realizadas.

Cada fase da metodologia CRISP-DM orienta e estrutura o todo do trabalho para que se possa direcionar, planear e organizar a implementação do projeto de AM. Nas próximas secções serão apresentadas cada uma destas fases da metodologia, com a definição do conceito e a respetiva componente prática deste projeto de mestrado.

A partir dos próximos tópicos serão apresentados cada uma destas fases do processo, o conceito e a realização a partir deste projeto.

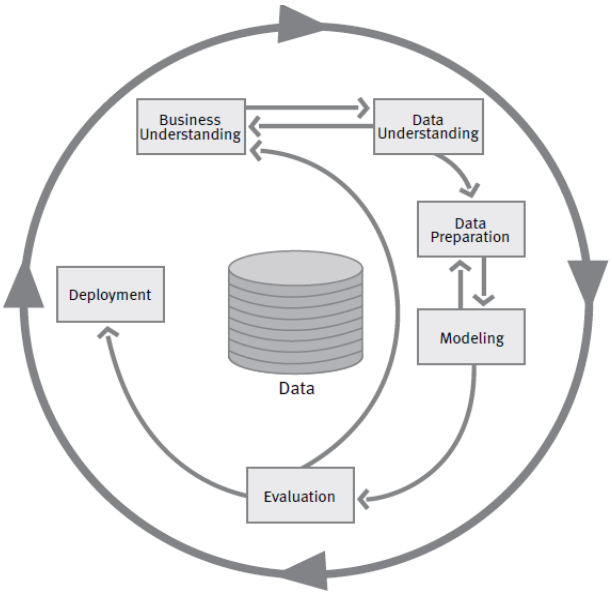
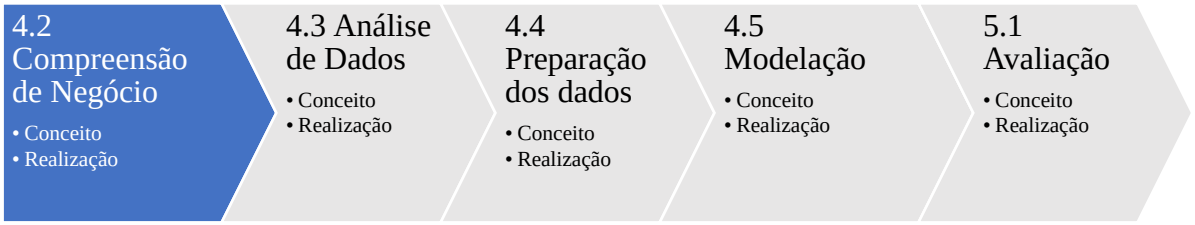


Figura 9 - Fases do modelo CRISP-DM (Pete, et al., 2000).

4.2 Compreensão de negócios



4.2.1 Conceito

Nesta fase de compreensão de negócios, o objetivo principal é avaliar e perceber as expectativas das partes interessadas[4] no projeto. Nesta fase, as partes interessadas apresentam

4 Para Montes & Patz (2017), Partes interessadas (traduzido do termo inglês, *stakeholders*) são os indivíduos e as organizações envolvidos no projeto. Ou seja, quem têm algum tipo de interesse no projeto. Podem ser

as suas expectativas e especificam os objetivos gerais do projeto. O alvo e objetivos do projeto são apresentados, entendidos e discutidos para que seja possível avaliar o contexto e cenário atual *(As-is)* e os objetivos e as necessidades futuras *(to-be)*. Também são discutidos recursos os envolvidos no projeto, como por exemplo que informações serão disponibilizadas, que sistemas estão envolvidos, equipa de projeto, etc.

4.2.2 Realização da Compreensão de Negócios

Com base na apresentação dos conceitos na fase de Compreensão de Negócio, nesta secção é apresentada a sua realização a partir de reuniões e sessões de trabalho com a equipa de projeto do MILAGE Aprender+. Foi identificada como a parte interessada do projeto o Doutor Professor Mauro Figueiredo.

Para além da apresentação da aplicação MILAGE Aprender+, assim como o seu funcionamento atual, foram discutidas expectativas para a realização deste trabalho de mestrado, com foco na aplicação de um modelo de AM que permita:

- estimar o desempenho dos alunos.
- usar informação histórica para desenvolver um modelo que suporte a recomendação de conteúdos aos alunos e oriente os professores no apoio aos alunos.

Nestas sessões de trabalho com as partes interessadas foram apresentados como objetivos principais:

- Utilizar a informação histórica do MILAGE Aprender+ para que o percurso de aprendizagem do utilizador aluno seja otimizado com a recomendação de conteúdo.
- Permitir, aos professores, previsão de desempenho dos alunos e/ou grupo escolar.

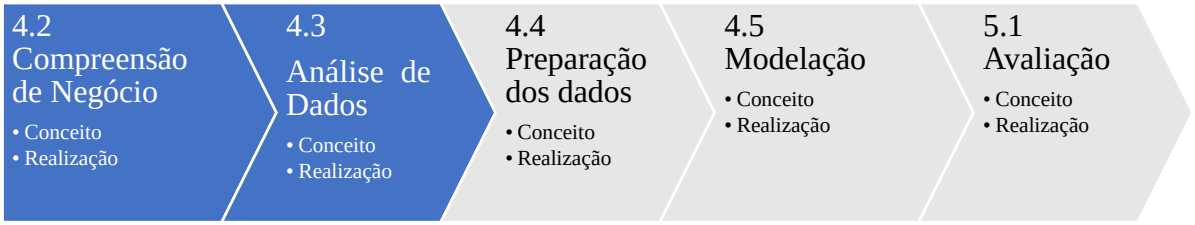
As reuniões de trabalho com a equipa MILAGE Aprender+, para a definição e compreensão do amplo contexto do MILAGE Aprender+, ocorreram várias vezes, mesmo na fase seguinte de análise dos dados. À medida que nova informação era aprendida e/ou

positivamente ou negativamente afetados com a sua execução e podem influenciar o projeto e/ou seu resultado. O projeto irá atender necessidades das partes interessadas e elas, por sua vez, são responsáveis por desempenhar o papel acordado para atender o objetivo do projeto.

identificada, durante estas interações tornou-se possível a correção e/ou otimização do plano do projeto.

De seguida, apresenta-se a fase de análise de dados. Contudo, na realização prática, por vezes retornámos à fase de Compreensão de Negócios devido à constante dependência com as tarefas realizadas na fase seguinte. Esta dependência contínua entre as fases devesse sobretudo à necessidade de alinhamento entre o estado atual de funcionamento do MILAGE Aprender+ e os objetivos da parte interessada para o projeto.

4.3 Análise dos dados



4.3.1 Conceito

Na fase de análise de dados, com o principal objetivo de compreensão dos dados, é realizada a identificação dos dados chave que poderão enriquecer o modelo. Na metodologia CRISP-DM, o principal objetivo da fase de Análise de Dados é identificar e analisar o conjunto de dados que agregarão valor ao modelo para que o objetivo definido na fase de Compreensão de negócios seja alcançado. Ou seja, avaliar quais dados ou informação serão úteis e contribuirão para que o modelo de AM seja definido.

Em geral esta fase é a mais dispendiosa já que requer o maior esforço durante o projeto. Nesta fase o trabalho concentra-se em perceber combinações, avaliar dados históricos, verificar e definir variáveis relevantes para o modelo, identificar as propriedades da informação identificada, avaliar tipos e relacionamentos entre dados.

Outra atividade que deve ser aplicada nesta fase é a limpeza dos dados. Significa que, dados limpos, sem ruídos são extremamente vitais para o avanço e início da modelação, já que os modelos de AM são muitos sensíveis aos dados. Além de compreensão plena dos dados que serão utilizados, garantir a qualidade dos dados é palavra de ordem. É esperado que todo o dicionário de dados, o seu diagrama de relacionamentos, bem como o seu significado seja devidamente documentado.

Nesta fase é realizada a exploração e visualização dos dados. A partir deste tipo de análise exploratória será possível produzir um resultado inicial de compreensão dos dados, bem como avaliar o seu impacto ao longo do projeto. Geralmente a representação gráfica permite uma melhor percepção e apoia o processo de decisão em ações futuras.

Uma regra comum é 80% do tempo do projeto na preparação de dados (Hotz & Saltz, 2021). Segundo relatório do MITSloan (Redman, 2017), o custo de dados com ruídos[5] é de 15% a 25% da receita para a maioria das empresas. A IBM estimou que o impacto anual apenas na economia dos Estados Unidos é de espantosos US $ 3,1 trilhões. Isso ocorre por conta de dados mal definidos, com erros, incompletos, com ruídos, etc. Os cientistas de dados gastam aproximadamente 80% do tempo de um projeto localizando, limpando e organizando dados, deixando apenas 20% de seu tempo para realizar análises. Segundo o relatório do MITSloan (Redman, 2017), o custo de dados com ruído[6] é de 15% a 25% da receita para a maioria das empresas. A IBM estimou que o impacto anual apenas na economia dos Estados Unidos é de espantosos US $ 3,1 trilhões. Esta situação ocorre devido a dados mal definidos, com erros, incompletos, com ruído, etc. Os cientistas de dados gastam aproximadamente 80% do tempo de um projeto localizando, limpando e organizando dados, deixando apenas 20% do seu tempo para realizar análise.

4.3.2 Realização da Análise de dados

Esta fase de Análise de Dados consistiu em realizar uma ampla análise dos dados do MILAGE Aprender+. Esta fase do projeto, como indicado na secção sobre conceitos, foi a que consumiu mais tempo para permitir realizar completamente todas as atividades do projeto.

Inicialmente foi feita uma análise detalhada do diagrama de entidade relacionamento para perceber toda a informação registada e armazenada na base de dados MILAGE Aprender+. Este modelo de dados é apresentado na figura 10.

6 Ruído pode ser definido como uma parte num conjunto de dados que aparentemente é inconsistente com os restantes dados existentes, pois não segue o mesmo padrão dos demais. Ruído num conjunto de dados pode reduzir o desempenho das técnicas de Aprendizado de Máquina (AM) aplicadas e aumentar o tempo de construção da hipótese induzida, assim como a sua complexidade.

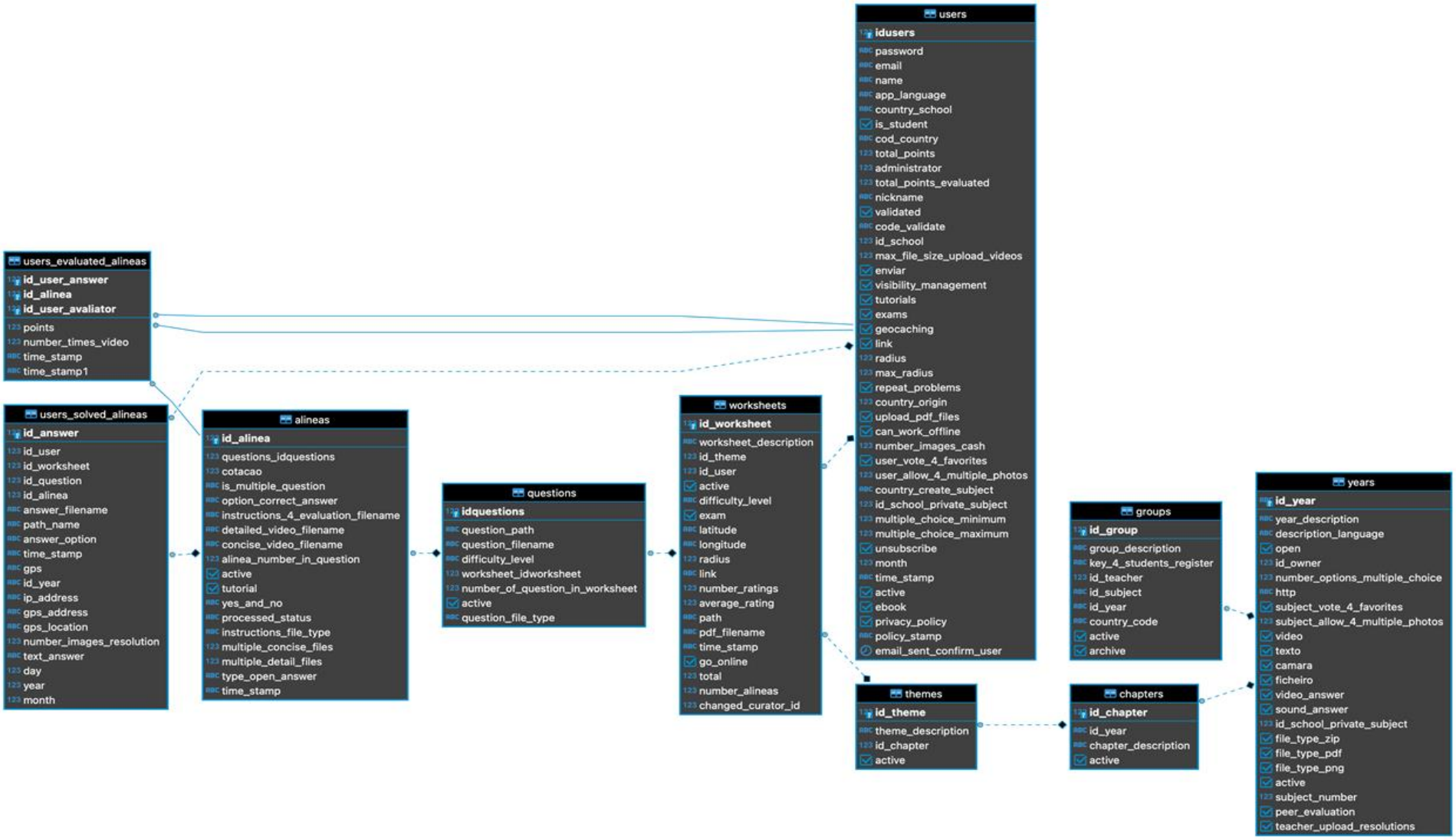


Figura 10 - Modelo entidade relacionamento do MILAGE Aprender+ (MILAGE Aprender+ 2021)

Para além de observar o modelo dos dados da aplicação MILAGE Aprender+, foi feita uma análise ao seu funcionamento para perceber e identificar cada um dos dados, relacionando-os com a análise funcional da aplicação.

Na figura 11 é apresentada cada secção de alguns ecrãs do MILAGE Aprender+ identificando a sua função no sistema, assim como a relação com os dados identificados e que serão utilizados no projeto.

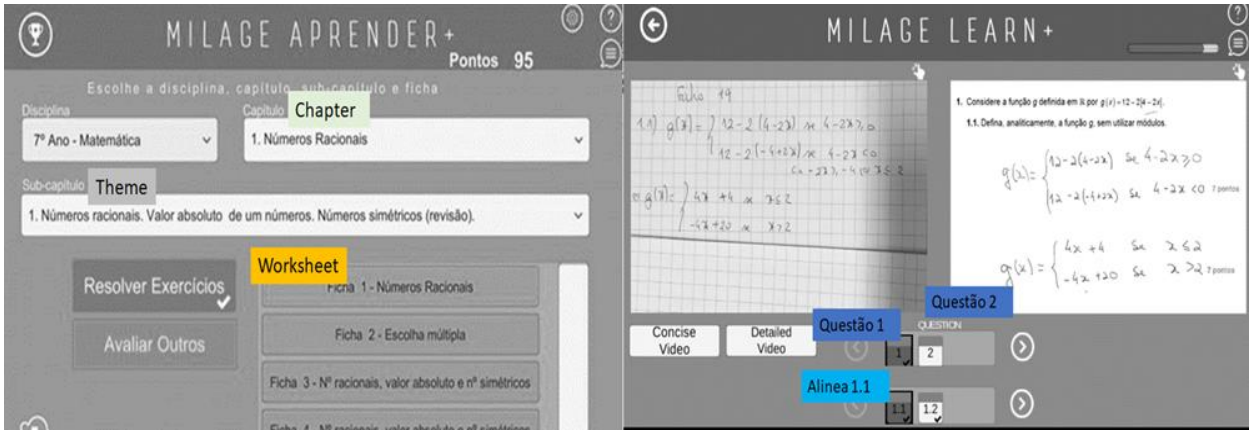


Figura 11 . Ecrãs do MILAGE Aprender+ com identificação dos dados (Felipe Fonseca, 2021)

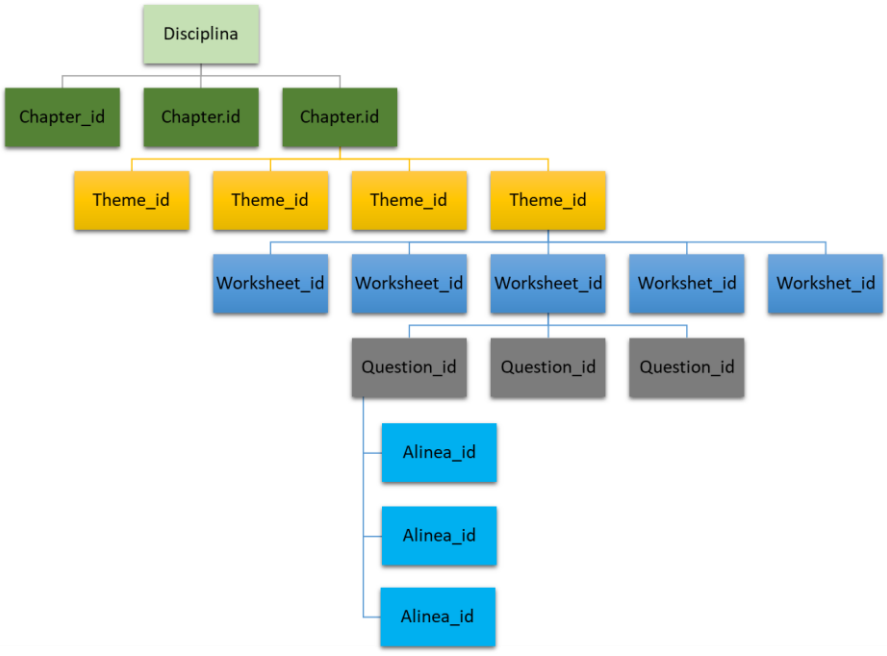


Figura 12 - Relação hierárquica do conteúdo e dados MILAGE Aprender+ (Felipe Fonseca, 2021)

A partir dos principais dados do MILAGE Aprender+, suas tabelas, relações e campos da aplicação, foi criado um dicionário de dados para apoiar na compreensão dos dados analisados, onde se incluiu uma coluna com a sua descrição. Na tabela 5 foram identificados os dados selecionados para ajudar na construção do modelo.

id	Nome da variável	Tabela	Descrição	Tipo
1	id.Alínea	Alíneas	Identificação dos exercícios Alíneas	Txt
2	cotacao	Alíneas	Valor máximo de ponto da Alínea	Int
3	active	Alíneas	Estado true/false da Alínea	Binário
4	is_multiple_question	Alíneas	Tipo de resposta da Alínea	Binário
5	id_question	Question	Identificação das questões que contém Alíneas	Txt
6	difficulty_level	Question	Nível de dificuldade da Questão	Int
7	active	Question	Estado true/false da Alínea	Binário
8	id_worksheet	Worksheet	Identificação das fichas que contém questões	Txt
9	difficulty_level	Worksheet	Nível de dificuldade da Ficha	Int
10	active	Worksheet	Estado true/false da Alínea	Binário
11	id_theme	Theme	Identificação dos temas que contém fichas	Txt
12	active	Theme	Estado true/false do tema	Binário
13	id_chapter	Chapter	Identificação dos Capítulos que contém temas	Txt
14	id_school	Schools	Identificação da Escola do utilizador	Txt
15	id_user	User_evaluated_Alíneas	Identificação do utilizador	Txt
16	id_user_avaliator	User_evaluated_Alíneas	Identificação do utilizador que avalia	Txt
17	points	User_evaluated_Alíneas	Total de pontos da Alínea	Int
18	number_times_video	User_evaluated_Alíneas	Total de vezes que foi assistido vídeo	Int
19	time_stamp	User_evaluated_Alíneas	Data e hora da realização da Alínea	Int
20	is_student	Users	É tipo aluno ou não?	Binário
21	total_points	Users	Total de pontos do aluno	Int
22	time_stamp	Users	Data de criação do utilizador	Txt
24	nickname	Users	Identificação do utilizador	Txt

Tabela 5 – Mapeamento de dados selecionados e seu significado.

Para além dos campos apresentados na tabela 5, foram também identificados outros diversos campos e que foram eliminados na construção do modelo. O critério para utilização dos dados teve em consideração alguns princípios recomendados, como por exemplo pela

Microsoft, para modelos confiáveis e responsáveis (Microsoft, 2021): responsabilidade, inclusão, confiabilidade, segurança, justiça, transparência, privacidade e segurança:

- Dados do género e/ou até mesmo como país, região do utilizador foram desconsiderados.
- Dados como códigos de verificação, endereços de correio eletrônico, nome de registo ("nickname") e qualquer outro dado particular do utilizador foram eliminados de qualquer análise, preservando a Lei de Proteção geral dos Dados (LGPD)[7].
- Outros dados relacionados com opções de personalização técnica do utilizador também foram desconsiderados na análise, por exemplo: Se permite upload de ficheiros, total de imagens autorizadas para cache, etc. Estes dados não têm qualquer tipo de correlação no âmbito do projeto.

Após a compreensão dos dados e respetivo significado foi realizada a exploração dos dados a fim de identificar eventuais anomalias e problemas na qualidade dos dados. Todo o trabalho de exploração foi realizado a partir de amostras de dados reais do MILAGE Aprender+ que serão apresentadas de seguida.

4.3.3 Amostras

O trabalho foi realizado em três amostras de dados:

I. Amostra I: Os dados inicialmente trabalhados foram especificamente da disciplina de Química Orgânica. Uma amostra com 2931 registos de Alíneas contendo o total de pontos da auto-avaliação e também da avaliação concedida pelos pares. Entretanto, detetamos que nesta amostra da disciplina de Química Orgânica havia uma condição muito especifica, que foge ao padrão geral do MILAGE Aprender+. Foi detetado que para esta amostra havia apenas uma Alinea por questão. Descartámos esta amostra pois mostrou ser uma disciplina com comportamento muito específico e que não acrescentaria valor ao modelo em estudo. O período dos dados utilizado foi todo o intervalo de dados disponível na aplicação, desde a primeira Alínea respondida em 2010 até 31 de julho de 2020.

[7] O Regulamento Geral sobre a Proteção de Dados (RGPD), Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016, apresenta um conjunto único de regras relativas à proteção das pessoas singulares no que diz respeito ao tratamento de dados pessoais e à livre circulação desses dados.

II. Amostra II: A partir deste momento trabalhei com os dados da disciplina de Matemática do 9º ano. A nova amostra da disciplina de Matemática com um total de 266.086 registos de Alíneas com respostas e pontuação dos alunos foi utilizada na exploração dos dados. O período dos dados utilizado foi todo o intervalo de dados disponível na aplicação, desde a primeira Alínea respondida em 2010 até 31 de julho de 2020.

III. Amostra III: Para amplificar a análise foi realizada a todo o conteúdo do 9º ano (várias disciplinas), trabalhando assim com uma amostra de 829.709 registos de Alíneas com respostas (Tabela: "user evaluated alineas"). O período dos dados utilizado foi todo o intervalo de dados disponível na aplicação, desde a primeira Alínea respondida em 2010 até 31 de julho de 2020.

Capítulos	Temas	Fichas	Questões	Alíneas	Utilizadores	Alíneas respondidas e avaliadas	Escolas
1.075	1.531	5.987	15.590	24.761	53.260	829.709	2.459

Tabela 6 – Resumo da Amostra de dados II do MILAGE Aprender+.

4.3.4 Exploração de dados

Amostra II – Somente disciplina Matemática 9º. Ano;

A análise que se segue foi feita com dados reais do MILAGE Aprender+ usando como amostra de dados as Alíneas realizadas pelos alunos do 9º ano de Matemática. Desta amostra, foram avaliadas somente aquelas Alíneas que estavam no estado "já avaliadas", ou seja, Alíneas que já receberam, além da pontuação de autoavaliação do aluno, também a avaliação do par e ou do professor. Esta análise possibilitou a seguinte constatação:

- Total de Alunos: 11.096 alunos responderam a alíneas com pelo menos a autoavaliação e uma segunda avaliação.
- Total de Alíneas: 1.237 Alíneas com resolução e avaliação

A figura 13 apresenta todos os registos de Alíneas resolvidas e avaliadas pelos 11.096 alunos. Cada ponto apresentado no gráfico representa um único aluno fazendo a intersecção entre o seu total de Alíneas resolvidas e avaliadas (x) com o total de pontos obtidos (y).

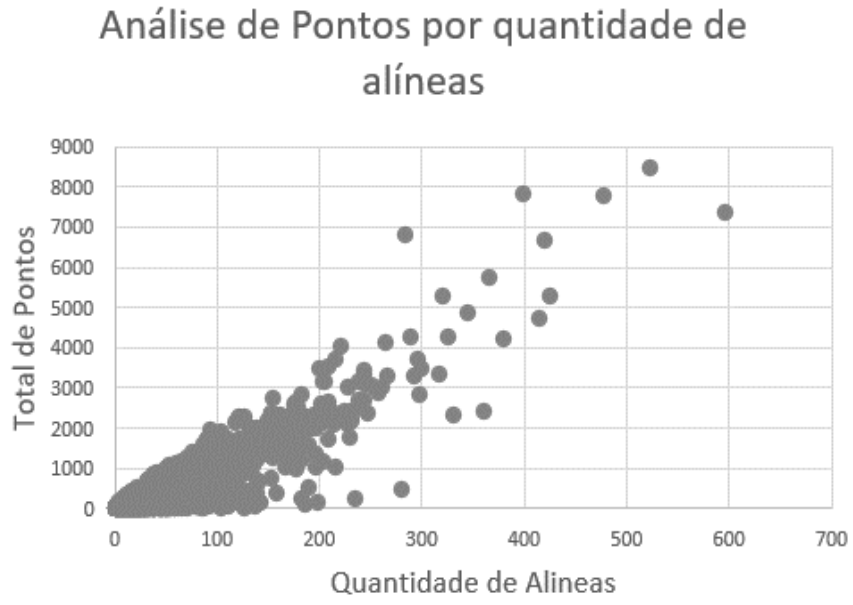


Figura 13 -Todos os alunos e todas as Alíneas avaliadas por pontos (Felipe Fonseca, 2021)

A partir da figura 13, é possível fazer algumas observações:

- Do total de 11096 alunos, apenas 6 (0,05%) alunos realizaram mais do que 400 Alíneas.
- Do total de 11096 alunos, apenas 6 (0,05%) alunos alcançaram mais do que 6.000 pontos.
- Um aluno, mesmo que com menos de 300 Alíneas realizadas, alcançou mais do que 6000 pontos.
- Dois alunos que realizaram mais do que 400 Alíneas não conseguiram alcançar 6000 pontos.
- Um aluno que fez aproximadamente 600 Alíneas não foi quem mais pontuou.
- O aluno que mais pontuou não foi quem fez mais fez Alíneas.
- Do total de 11096 alunos, apenas 33 (0,29%) alcançaram 4000 ou mais pontos.
- Do total de 11096 alunos, apenas 307 (2,76%) realizaram mais do que 100 Alíneas.

A figura 14 apresenta a análise sobre as Alíneas respondidas pelos alunos, onde se observa a pontuação média de todos os alunos por Alínea versus a quantidade de Alíneas. Observámos que nas 1237 Alíneas, contabilizaram-se 231288 respostas.

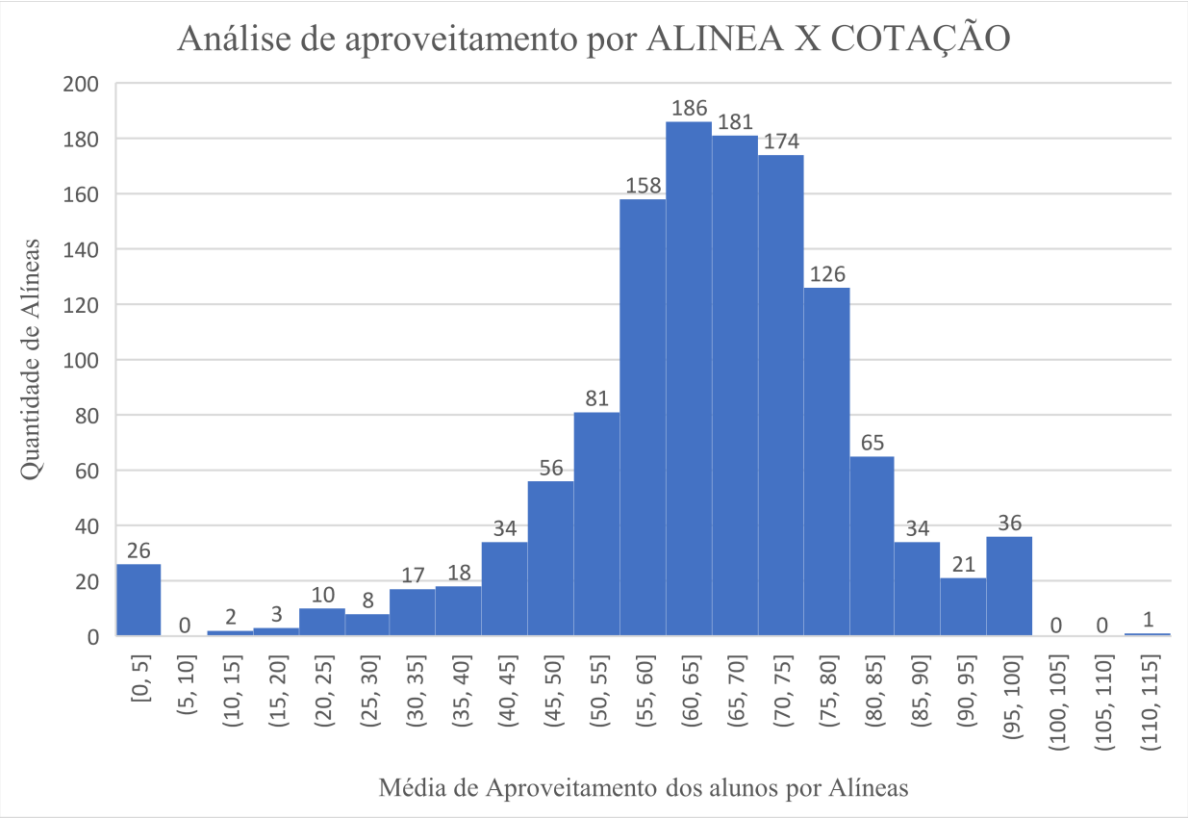


Figura 14 - Análise de aproveitamento a partir da pontuação média dos alunos nas 1237 Alíneas (Felipe Fonseca, 2021)

A partir da figura 14, é possível fazer algumas observações:

- No total das 1.237 Alíneas há um caso onde o total de pontos excede a cotação total da questão (entre 110 e 115%).
- No total das 1.237 Alíneas, apenas em 58 Alíneas (4,68%) os alunos tiveram aproveitamento igual ou superior a 90%.
- No total das 1.237 Alíneas, 413 Alíneas (33.3%) apresentam aproveitamento inferior a 60%.

Para aprofundar a compreensão sobre a utilização do MILAGE Aprender+ pelos alunos e identificar um comportamento padrão na realização e aproveitamento das Alíneas, a figura 16 apresenta uma visão das Alíneas que tiveram pelo menos 100 respostas, ou seja, Alíneas que foram razoavelmente acedidas e respondidas pelos utilizadores da ferramenta. Como resumo das definições da análise apresentada na figura 16, consideramos:

- Para o total de 1.237 Alíneas e para um total de 231.288 respostas, ao avaliar como média padrão, cada alínea tem 186,97 respostas.

- Entretanto, para se perceber o desvio desta média, apenas 562 Alíneas possuem cem ou mais respostas.
- A análise abaixo é feita a partir de uma amostra de 562 Alíneas que possuem 100 ou mais respostas.
- A média no aproveitamento para estas 562 Alíneas é de 67,03%.

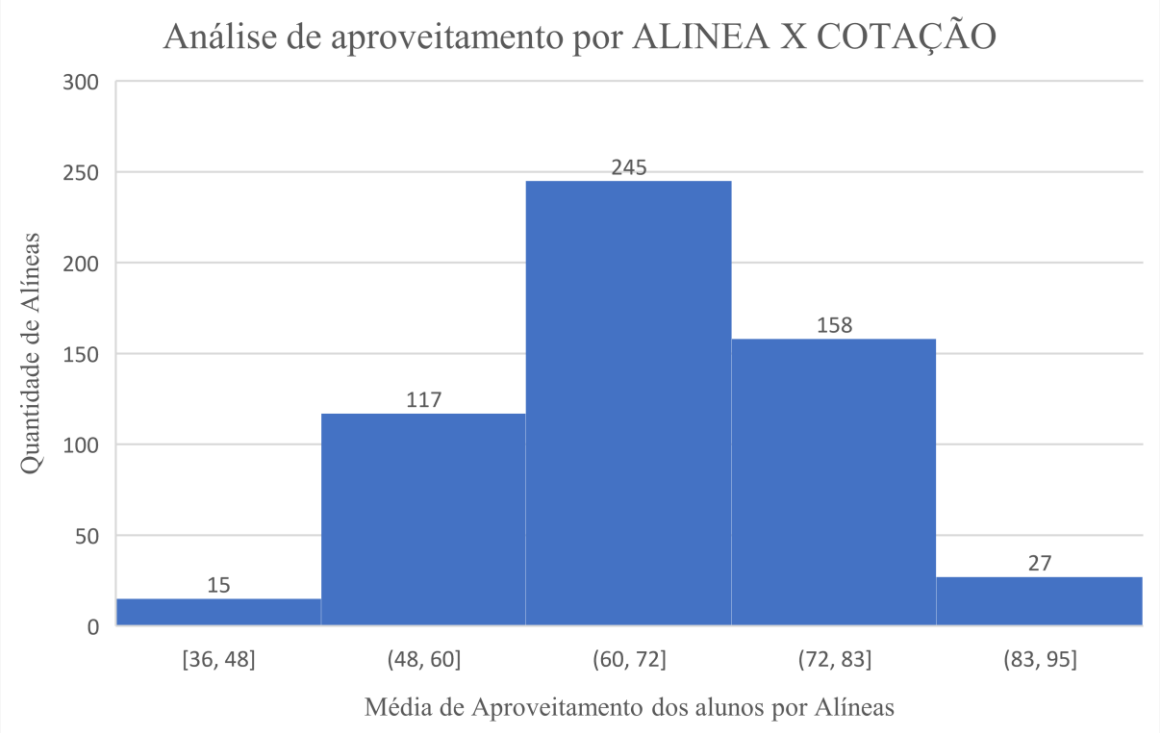


Figura 15 -Desempenho médio dos alunos nas Alíneas com mais de 100 respostas.

A partir da figura 15 é possível fazer algumas observações da distribuição:

- É possível observar que o aproveitamento se concentra na sua maior parte (43,5%) entre o intervalo de 60% a 72% de aproveitamento médio nas Alíneas. 67% da Alíneas apresentam médias de pontuação inferior a 72%.
- Podemos observar que apenas 32% das Alíneas possibilitaram aos alunos alcançar aproveitamento superior a 72%.
- Apenas 4.8% das alíneas tem alunos com média igual ou superior a 83%.

A partir desta análise, conforme figuras 15 e 16, permitiram verificar que os alunos têm dificuldade em obter aproveitamento igual ou superior a 70%.

Ainda durante a fase de Análise de dados, realizando validação e análise da qualidade dos dados, foram detetadas anomalias que se apresentam de seguida:

- I. Foi detetado que cada item de conteúdo como Alínea, Questão, Worksheet, Theme e Chapter possui uma variável de indicação de estado (True/False). Este estado indica se aquele conteúdo se encontra ou não disponível. Entretanto foram observados diversos conteúdos onde o estado com relação hierárquica não é propagado para os conteúdos subsequentes. Por exemplo: Alínea está no estado ativo, entretanto a Questão encontra-se como inativa.

- II. Caso o professor faça a avaliação de uma Alínea de um aluno, a pontuação é simplesmente alterada no registo do aluno. Ou seja, a nota da auto-avaliação realizada pelo aluno é simplesmente substituída pela nova nota do professor.

- III. Foram detetadas situações de registos com pontuação negativa, conforme evidencia na coluna "points" da tabela que se segue.

id_user_answer	id_Alínea	points	id_user_avaliator	time_stamp
33690	11296	-10	33690	19/10/2020 10:07
61628	836	-10	61628	18/10/2020 19:33
82947	11845	-30	82947	10/03/2021 12:26
82947	22949	-60	82947	10/03/2021 12:25
82947	22951	-9	82947	10/03/2021 11:38
32141	34923	-10	32141	23/10/2020 14:14
32165	34925	-10	32165	23/10/2020 13:24
57801	34930	-10	57801	23/10/2020 14:12
61628	834	-10	61628	18/10/2020 19:24

Tabela 7 – Evidência de entradas com pontuação negativa.

- IV. Foram detetados registos de pontos nas respostas às Alíneas do aluno que excedem a cotação máxima definida para a Alínea. A tabela seguinte apresenta uma evidência, ou seja, o registo do *id_user_answer* 336, com pontuação de 35 numa Alínea onde a cotação máxima é de 25 pontos.

id_user_answer	id_Alínea	points	id_user_avaliator	number_times_video	time_stamp
336	703	35	390	0	NULL

questions_idquestions	id_Alínea	cotacao	is_multiple_question
312	703	25	n

Tabela 8 – Mais pontos atribuídos do que o permitido na Alínea.

- V. Foram encontrados registos onde há mais de 2 avaliações para uma Alínea. Por regra, o número de avaliações por Alínea consiste na auto-avaliação do aluno e na do par. Porém, conforme evidência abaixo, há casos onde há uma terceira avaliação.

id_user_answer	id_Alínea	points	id_user_avaliator	time_stamp
25509	10978	5	23763	25/06/2020 11:02
25509	10978	4	30720	24/06/2020 12:17
25509	10978	6	25509	16/02/2020 22:29
id_user_answer	**id_Alínea**	**points**	**id_user_avaliator**	**time_stamp**
30720	24342	6	23765	28/02/2020 21:37
30720	24342	6	25509	27/02/2020 19:01
30720	24342	6	30720	26/02/2020 15:51
id_user_answer	**id_Alínea**	**points**	**id_user_avaliator**	**time_stamp**
478	19349	15	561	31/03/2019 22:28
478	19349	15	469	20/03/2019 23:31
478	19349	15	478	17/03/2019 19:41

Tabela 9 – Mais de 2 registos de avaliação de uma Alínea por aluno.

- VI. Observamos que caso o total de alunos numa turma ou grupo seja ímpar, um aluno ficará sem a avaliação por pares. Por exemplo, numa turma de 9 aluno, pelo menos 1 aluno ficará sem ter os pontos da avaliação do par. Penalizando o aluno que ficará com pontuação não avaliada ou pendente.

- VII. Foram observados diversos registos com caraterísticas de teste, mesmo na base de dados de produção. Foram identificados registos como por exemplo: criação de fichas, questões e alíneas de teste, criação de utilizadores para demonstração da ferramenta, utilizadores alunos para simulação com pontuação aleatória.

- VIII. Ao realizar uma auditoria à pontuação do Aluno Id25509 observamos um total de pontos de 8461. Do total de 8461 pontos acumulados, 5990 foram atribuídos pelo próprio aluno. Os outros demais pontos foram atribuídos por outros Id´s na avaliação do par do tipo "aluno". Ou seja, não há nenhum registo de avaliação ou aprovação de um utilizador do tipo professor.

Histórico do Aluno:	23763	23765	23772	24151	25509	30720	Total pontos
25509	***731***	***517***	***70***	***197***	***5990***	***956***	***8461***

Tabela 10 – Auditoria de pontos do aluno id 25509.

- IX. Foram detetados registos onde o aluno conseguiu responder mais de uma vez à mesma Alínea.

[id_user_answer]	[id_Alínea]	[points]	[id_user_avaliator]	[number_times_video]	[time_stamp]
733	697	10	733	0	24/11/2016 10:52
733	697	10	820	0	01/12/2016 14:14

[id_user_answer]	[id_Alínea]	[points]	[id_user_avaliator]	[number_times_video]	[time_stamp]
57964	697	20	57964	0	17/11/2020 14:53
57964	697	7	57901	0	22/11/2020 14:23

Tabela 11 – Evidência de mais de uma resposta para a mesma alínea pelo mesmo id user.

- X. Foram detetados registos com informação de que o aluno saltou a opção de auto-avaliação através da coluna com a data "time_stamp" da tabela das Alíneas. Nestes casos o campo ficou registado com a indicação de "Student skipped auto-evaluation". Desta forma, não se consegue obter a data de realização da Alínea.

Todos os 9 tópicos identificados nesta análise de dados foram apresentados ao grupo de trabalho do MILAGE Aprender+ juntamente com notas de recomendações que serão apresentadas no capítulo 5, secção 5.4.

Amostra III – 9º. Ano com todas as disciplinas;

Com o objetivo de aumentar o tamanho da amostra para perceber se existiam outros tipos de anomalias nos dados e ou mais temas relacionados com a qualidade dos dados, foi também realizado uma análise exploratória considerando os dados de todo o 9ª ano, ou seja, incluindo outras disciplinas.

Por tratar-se de uma massa de 829.709 registos de Alíneas respondidas por 32.436 alunos, foram construídos painéis de dados, a partir do uso da ferramenta Microsoft Power BI para a análise de qualidade dos dados.

Todo o trabalho foi concentrado na tabela "user evaluated alineas", que é justamente a tabela que contém o histórico das realizações dos alunos que foram avaliadas (829.709 registos).

Validação e qualidade dos dados da amostra de todo o 9ª ano:

I. Foram detetadas datas com formato dd/mm/aaaa hh:mm:ss e no mesmo campo dados (time_stamp) com a indicação de que o aluno não respondeu à auto-avaliação. Em nenhum outro local ficou armazenada a data da resposta à Alínea.

id_user_answer	id_Alínea	id_user_avaliator	points	time_stamp
1	14	1	0	Student skipped auto-evaluation
1	6736	1	0	Student skipped auto-evaluation
1	7099	1	0	Student skipped auto-evaluation
1	7939	1	0	Student skipped auto-evaluation
1	7941	1	0	Student skipped auto-evaluation
1	8440	1	0	Student skipped auto-evaluation
1	10692	1	0	Student skipped auto-evaluation

Tabela 12 – Evidência de Alíneas onde o aluno não realizou a auto-avaliação.

- II. No mesmo campo da data "time_stamp" foram detetados 4 registos de Alíneas datadas no ano 2033.

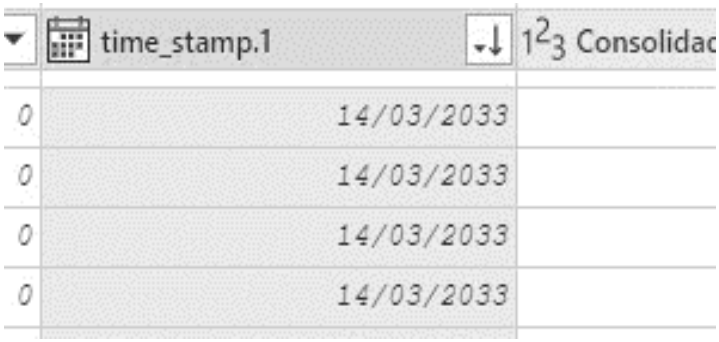


Figura 16 - Data de realização de Alínea com indicação em data futura.

- III. Foram identificados 23 registos contendo apenas dois dígitos no campo de data (22), faltando mês, ano e hora.
- IV. Foram identificados 2301 registos em que o campo de data de realização da Alínea ("time stamp") não foi preenchido.

id_user_answer	id_Alínea	id_user_avaliator	points	time_stamp
1	12	1	10	
1	84	1	10	
1	86	1	2	
1	87	1	3	
1	88	1	10	
1	90	1	10	
1	92	1	10	
1	95	1	5	

Tabela 13 – Evidência de várias Alíneas respondidas sem data registada.

- V. Foram identificados 10 registos com pontuação negativa.

id_user_answer	id_Alínea	id_user_avaliator	points	time_stamp
182	536	69	-5	
303	892	319	-2	
605	1502	138	-5	09/11/2016 12:56:51
2247	1451	2247	-25	09/10/2017 14:00:18
2610	4318	2610	-2	08/09/2017 22:30:18
2763	1857	2763	-1	27/09/2017 12:38:41
2768	1853	2763	-30	27/09/2017 12:49:36
2808	20	2629	-20	06/02/2018 11:22:57
34669	5041	35441	-9	09/02/2020 18:54:17
34669	5043	35441	-9	09/02/2020 18:54:25

Tabela 14 – Evidência de várias Alíneas respondidas com pontuação negativa.

- VI. Foram identificados 314 registos de respostas a Alíneas com pontuação superior à cotação máxima da Alínea.

- VII. Foram detetados 2573 registos de utilizadores sem a indicação de data de criação desse registo de utilizador.

- VIII. Foram identificados 180 registo de Alíneas no estado ativo em que a respetiva questão se encontrava no estado false, ou seja, inativo. O estado não foi propagado para o conteúdo hierárquico subsequente.

- IX. Foram identificados 28 registo de worksheets no estado ativo em que o respetivo tema se encontra no estado false, ou seja, inativo. O estado não foi propagado para o conteúdo hierárquico subsequente ao conteúdo inativado.

- X. Foram identificados 9313 registos de respostas de utilizadores tipificados como não alunos. (is Student: “false”).

idusers	app_language	country_school	is_student	cod_country
1	Portuguese	Portugal	FALSO	pt
5	Portuguese	UK	FALSO	pt
6	Portuguese	Portugal	FALSO	pt
7	Portuguese	Portugal	FALSO	pt

Tabela 15 – Evidência de várias Alíneas respondidas por utilizadores com valor falso no tipo de aluno.

- XI. Foram detetados 532 registos sem indicação se a questão é do tipo questão múltipla ou não. Nesses registos o campo “is_multiple_question” está vazio

id_Alínea	questions_idquestions	cotacao	is_multiple_question
499	244	0	
511	248	0	
514	249	0	
520	250	0	
528	251	0	
529	251	0	
540	256	0	

Tabela 16 – Evidência de indicação de coluna com valores "em branco".

- XII. Foram identificados 17 registos de alunos que, responderam a Alíneas, mas que, entretanto, não possuem eventos de avaliação. Baseado na regra do sistema, nenhuma Alínea estará na tabela "users_evaluated_Alíneas" sem que existisse pelo menos uma avaliação do id_user (aluno), mesmo que estivesse indicado "Student skipped auto-evaluation" o evento deveria estar registado.

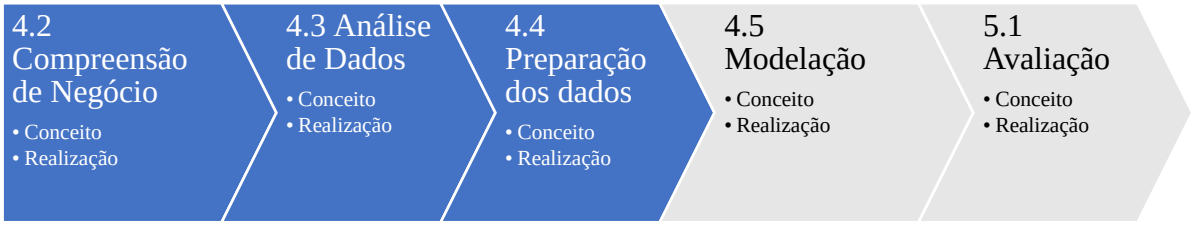
id_theme	id_chapter	id_user	users.id_school
416	140	**40765**	1923
417	140	**36912**	1206
484	174	**22829**	5
484	174	**37644**	1134
532	179	**39360**	781
540	181	**41718**	1563
581	197	**17785**	745
645	252	**19727**	1399
707	269	**15782**	937
707	269	**20891**	1349
707	269	**22648**	1232
707	269	**25782**	1048
707	269	**26516**	1015
1093	442	**14297**	1148
1237	274	**26900**	712
1499	610	**36521**	5
1605	657	**16375**	729

Tabela 17 – Evidência de Alunos identificados na tabela de realização de Alíneas.

- XII. Foram detetados 286 registos de Alíneas onde a indicação sobre o tutorial está vazia. Destaca-se que do total de 24761 Alíneas, apenas 532 contêm valor "verdadeiro" indicando que têm tutorial.

Todos os tópicos identificados nesta análise de dados foram apresentados ao grupo de trabalho do MILAGE Aprender+ juntamente com notas de recomendações que serão apresentadas no capítulo 5, secção 5.4.

4.4 Preparação dos dados



4.4.1 Conceito

O objetivo desta fase é definir o conjunto de dados que serão utilizados para o desenvolvimento de um modelo de AM. Na fase de preparação de dados são tomadas decisões sobre os dados usados na análise e desenvolvimento do modelo. Na decisão são considerados critérios de relevância, qualidade e avaliação de limitações, volume (quantitativo) e tipo de dados. Durante a preparação dos dados são consideradas a seleção e avaliação de atributos (colunas) e também os registos (linhas) do conjunto de dados. Também são aplicados critérios de exclusão e/ou limpeza dos dados. No final desta fase, todas as decisões e ações realizadas para a preparação dos dados são apresentadas e documentadas. É aplicado o critério de exclusão e ou limpeza dos dados, assim como descrição destas decisões.

Ainda na preparação de dados, são realizadas a integração e combinação de dados. Os dados unificados também cobrem agregações. A agregação e/ou seleção de recursos, traduzido do inglês *Feature Selection*[8], refere-se a operações em que novos valores são calculados resumindo informação de vários registos e/ou tabelas.

Uma tarefa importante neste processo é a formatação dos dados, organizando a informação sem alterar o seu significado, mas garantindo um padrão para que o processo de Modelação ocorra com fluidez.

Novos atributos são construídos a partir de um ou mais atributos, também conhecido como Engenharia de Recursos, traduzido do inglês *Feature Engineering*[9]. Por exemplo: Num

[8] Feature engineering: Seleção de recursos é o processo de seleção do subconjunto principal de recursos para reduzir a dimensionalidade do problema de treinamento (Microsoft, 2020).
[9] Feature engineering: Engenharia de recursos é o processo de criação de novos recursos a partir de dados brutos para aumentar o poder preditivo do algoritmo de aprendizagem. Os recursos de engenharia devem capturar informações adicionais que não são facilmente aparentes no conjunto de recursos original (Microsoft, 2020).

conjunto de dados que contém o atributo Peso e o atributo Altura, podemos criar um novo atributo conhecido como índice de massa corporal (IMC = Peso(kg) / Altura(m)^2. O objetivo desta criação de atributos é apoiar o modelo e ter informação com maior significado.

4.4.2 Realização da preparação dos dados

Para a preparação dos dados foram considerados todos os problemas identificados relacionados com a qualidade dos atributos (colunas) e também alguns registos (linhas) conforme apresentado na fase de análise dos dados. De acordo com os conceitos apresentado anteriormente, nesta secção será apresentada toda componente prática da preparação dos dados.

A amostra III, apresentada no capítulo 4.3.3, foi selecionada e aplicada nesta fase de preparação dos dados. A escolha baseou-se justamente na amplitude de informação, possibilitando um estudo mais amplo com todo o conteúdo dos exercícios resolvidos por alunos do 9º ano. Seguem-se as subsecções sobre: limpeza de dados, engenharia de recursos e seleção de recursos.

4.4.2.1 Limpeza dos dados

Nesta tarefa foram removidos os registos ou atributos que apresentavam as seguintes características:

- O atributo “time stamp” estava com a indicação que o aluno saltou a avaliação, ou seja, que não realizou a auto-avaliação.
- O atributo “time stamp” apresentava erro na composição da data.
- O atributo “time stamp” apresentava uma data no ano 2033.
- Os registos onde a data de registo e criação do utilizador estavam em branco.
- O atributo "is_multiple_question" estava vazio.
- O atributo "tutorial" estava vazio.
- Os registos que não contêm eventos de avaliação.
- Foi observado que muito dos problemas de qualidade apresentados estavam relacionados com o período inicial de uso da ferramenta, quando ainda estava no processo de amadurecimento. Por este facto, formam eliminados os registos

(linhas) de eventos anteriores a 2019. Considerando assim somente o período de 2019 a junho de 2020.

- Embora existissem muitos registos sobre Alíneas já no estado Indisponível e/ou que pertencessem a Questões ou fichas já desativadas, foi utilizado este conjunto de dados pois reuniam o histórico de realização e obtenção de pontos pelo aluno.
- Após a limpeza (exclusão) dos dados, o total inicial da amostra com 829.709 registos de Alíneas com respostas avaliadas foi reduzido a 648.234 registos.
- Ao realizar esta limpeza, usando como referência principal a tabela user evaluated alineas, a amostra de dados reduzida a 648.234 registos de respostas realizadas, passou a contar com os seguintes totais distintos: 293 capítulos, 746 Temas, 24.414 alunos e 820 escolas.
- Como exemplo, na Amostra III, o MILAGE Aprender contém 53.260 alunos registados, sendo 32.436 com respostas realizadas e avaliadas, mas após toda a limpeza de dados, ficamos com 24.414 utilizadores.

4.4.2.2 Engenharia de Recursos

Com o objetivo de derivar nova informação a partir de atributos já existentes no conjunto de dados, foram criados os seguintes atributos:

Nome do Atributo	Tipo	Descrição
QtdAlíneasnoThema	Numérico	Acumulador de Alíneas pertencentes ao Tema.
CotacaoTotalThem	Numérico	A cotação total do Tema.
GlobalMediaAprovPontosTema	Numérico	A média global de todos os alunos para dado tema.
AlunoQtdAlíneasRealizadasTema	Numérico	Quantidade de Alíneas realizadas pelo aluno num dado tema.
AlunoMediaAproveitamentoPontosRealizadoTema	Numérico	(Pontos*100)/Cotacao do referido tema.
AlunoPercentualRealizacaodoTema	Numérico	(QtdAlíneasnoThema*100)/AlunoQtdAlíneasRealizadasTema

AlunoMediaAprovPontosGERAL	Numérico	(Pontos*100)/Cotacao de todas as Alíneas realizadas.
AlunoMediaPontosGeral	Numérico	(pontos/Total de Alíneas realizadas)
AlunoDiferençaAprovMédiaGeral	Numérico	GlobalMediaAprovPontosTema - AlunoMediaAproveitamentoPontosRealizadoTema
AlunoFrequenciaUso	Numérico	Total de Alíneas realizadas/(Data de Criação do Aluno - 01/Junho/2020)
AlunoQtdAvaliacoesRealizadas	Numérico	Total de Alíneas avaliadas por dado aluno.
EscolaAlunoPontuacaoMedia	Numérico	Media de Pontos da Escola por dado aluno.
EscolaAlunoAproveitamentoMedia	Numérico	Media de Aproveitamento da Escola por dado aluno.
EscolaAlunoPontuacaoMinima	Numérico	Min Pontuação da Escola
EscolaAlunoPontuacaoMaxima	Numérico	Max Pontuação da Escola

Tabela 18 – Atributos, tipo de atributos e sua descrição.

Com o objetivo de reduzir a dimensão do conjunto de dados, foi realizada uma agregação dos dados no nível do Tema das Alíneas. Nesta agregação os dados de Alíneas, Questões e Fichas pertencentes ao agrupamento Tema foram resumidos a partir de novos atributos que foram construídos conforme tabela 18. Uma vez que os registos estavam no nível mais granular no nível atributo Alíneas, foi aplicado o agrupamento dos dados por Tema, considerando que este será o conteúdo recomendado ao utilizador.

Após a agregação, a dimensão do conjunto de dados passou de 648.234 registos para 77.975 registos. Neste processo de agregação mantivemos todos o conhecimento derivado dos dados e informações, realizando apenas uma agregação dos resultados.

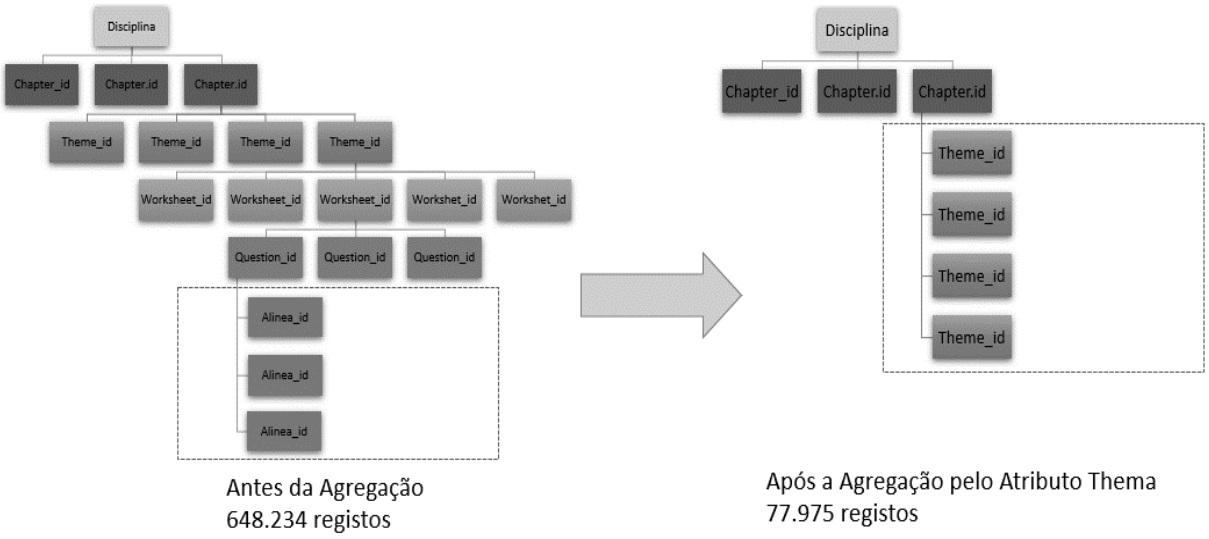


Figura 17 – Agregação dos dados para a partir do atributo Tema.

Com o objetivo de realizar experiências com modelos de classificação, também foi elaborada, a partir da amostra III, uma organização dos dados baseada nos alunos. Ou seja, usando como chave primária cada aluno e não apenas o Tema. De seguida este conjunto de dados é apresentados com os respetivos campos e detalhes.

id_user_answer	Identificação do Aluno
QtdDiasUtilizacao	Total de dias que o aluno tem no Milage Aprender
FrequenciaUso	Média de realizações pelo total de dias que o aluno tem no Milage Aprender+
QtdAvaliacoesRealizadas	Total de avaliações que o aluno realizou
QtdPontosObtidos	Total de pontos obtidos geral no Milage Aprender+
DificuldadeMediaAlineasRealizadas	Média de dificuldade das Alineas que o aluno realizou
MediaPontosObtidos	Média de pontos obtidos das alineas que o aluno realizou
CotacaoMediaAlineasTrabalhadas	Cotação média das alineas realizadas pelo aluno
PontuacaoMinima	Pontuação mínima do aluno em todas as alineas realizadas
PontuacaoMaxima	Pontuação máxima do aluno em todas as alineas realizadas
AproveitGeralnaPontuacao	Aproveitamento geral do aluno no Milage Aprender
QtdAlineasRealizadas	Total de alineas realizadas pelo aluno
QtdFichasRealizadas	Total de Fichas realizadas pelo aluno
QtdQuestõesRealizadas	Total de questões realizados pelo aluno
QtdTemasRealizados	Total de temas realizados pelo Aluno
Classificação	Classificação do Aluno a partir do Aproveitamento de pontos

Tabela 19 – Dados orientados a partir dos dados consolidados dos alunos.

4.4.2.3 Seleção de Recursos

Foi realizada uma análise abrangente sobre os atributos existentes na base de dados MILAGE Aprender+, sobretudo para perceber o seu significado e relação com o objetivo do projeto. Juntamente com os atributos que foram criados, conforme apresentados na tabela 18, os atributos selecionados para estabelecer o conjunto de dados e gerar os modelos aparecem descritos na tabela 20.

Nome do Atributo	**Tipo**	**Descrição**
id_theme	Categórica	Identificação do Tema
id_chapter	Categórica	Identificação do Capítulo
id_user_answer	Categórica	Identificação do Aluno
user_id_school	Categórica	Identificação da Escola
number_times_videoTema	Categórica	Total de vezes que o aluno assistiu o vídeo tutorial do referido tema.

Tabela 20 – Seleção de atributos.

Ainda no processo de preparação de dados foi realizada uma verificação da qualidade de cada atributo (coluna) e respetivos registos com o objetivo de verificar se ainda persistia qualquer outro tipo de anomalia para todos os conjuntos de dados. Estes resultados são apresentados na figura 18 e 19.

Figura 18 – Qualidade dos atributos a partir da ferramenta MS-Power BI – Conjunto de dados Temas

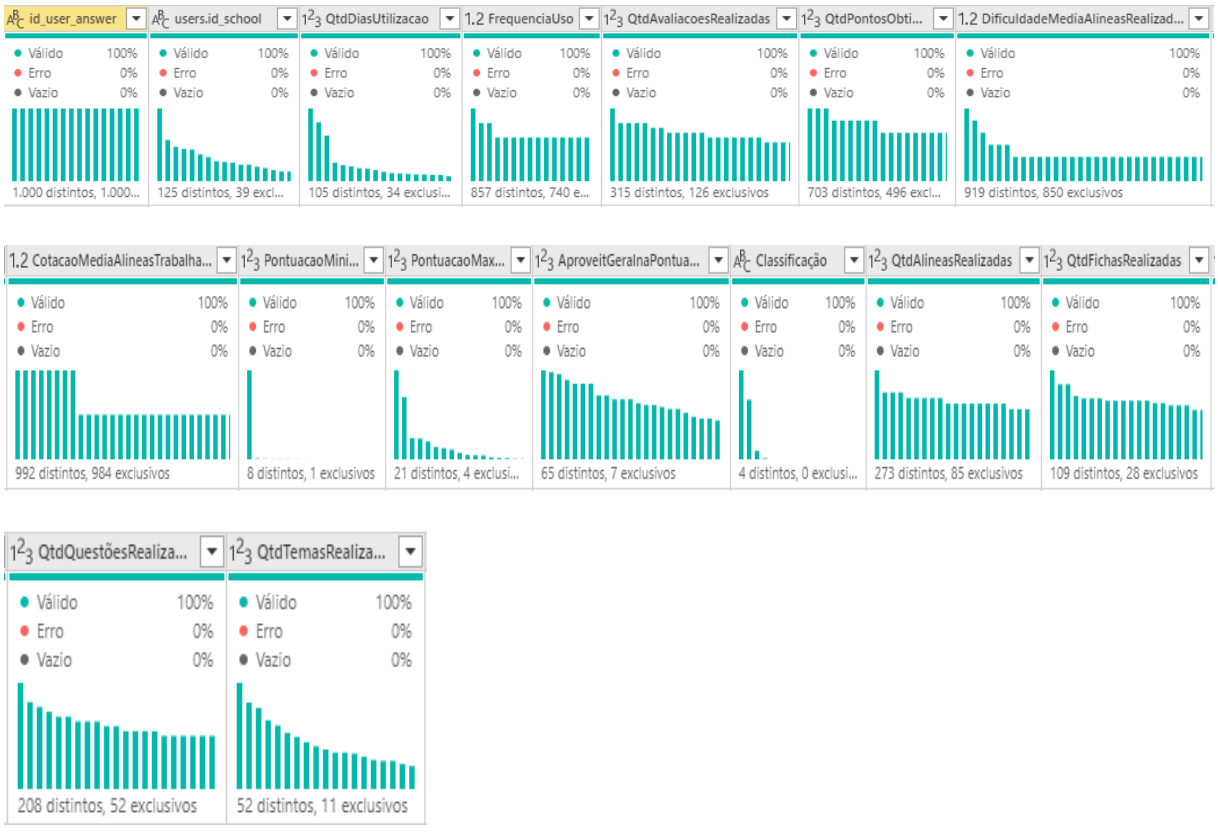


Figura 19 – Qualidade dos atributos a partir da ferramenta MS-Power BI – Conjunto de dados Alunos.

Na figura 18 e 19, apresentada a partir do Power BI, podemos avaliar o resultado de qualidade dos dados em três aspetos, válido (100%), erro (0%) e vazio (0%). Considerando os vários aspetos, o resultado final indica que os dados estão em conformidade para avançar.

Adicionalmente realizou-se uma segunda avaliação com a ferramenta RapidMiner Studio 9.9, o que permitiu comparar os resultados da análise. A figura 20 e 21 apresentam visualmente os resultados da qualidade dos dados com o RapidMiner.

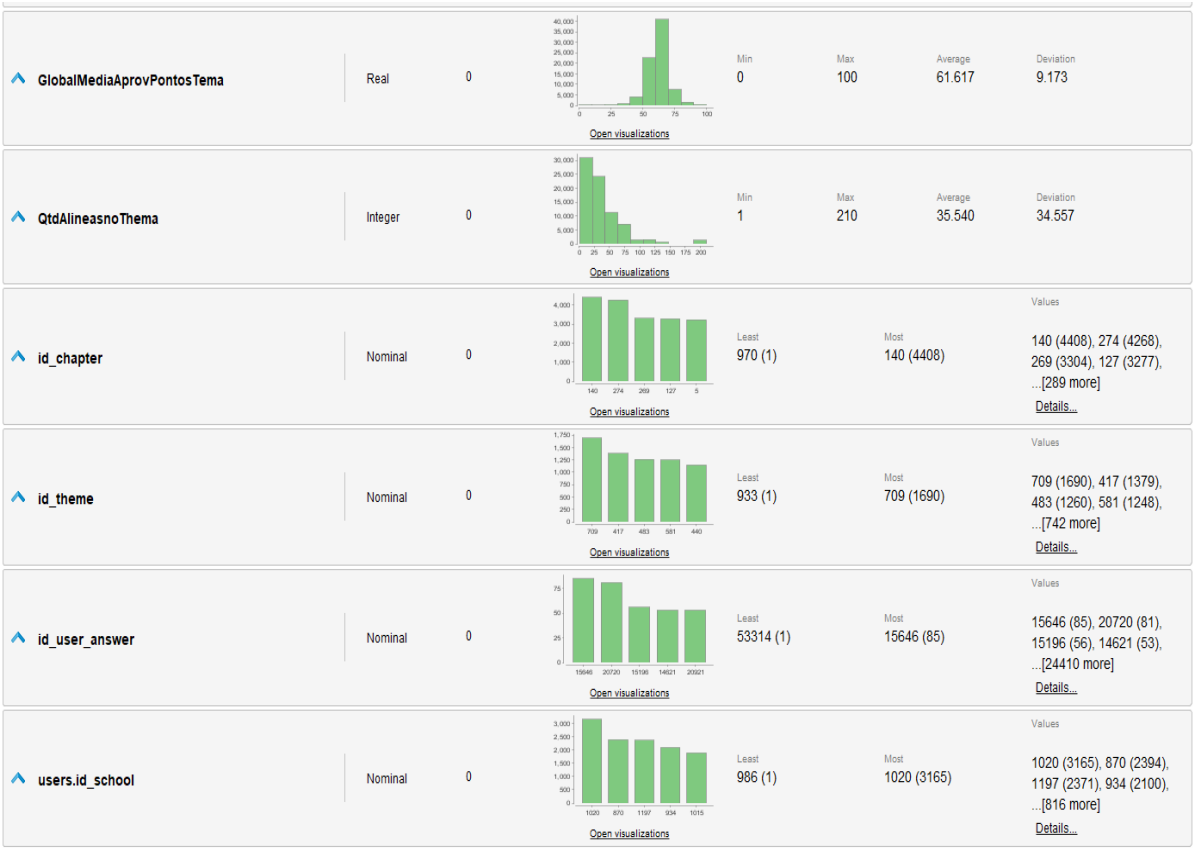


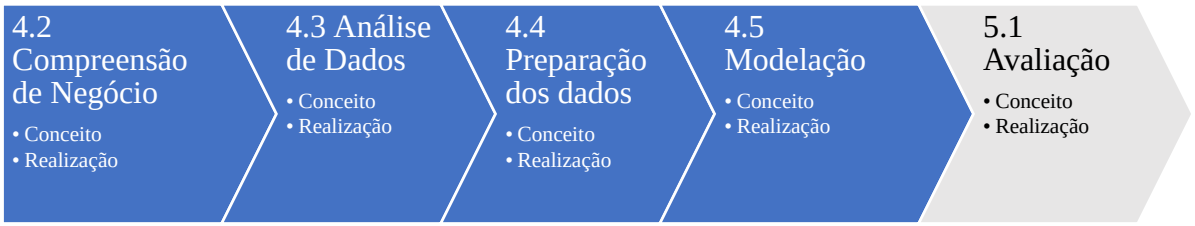
Figura 20 – Qualidade dos atributos a partir da ferramenta RapidMiner (parte II).

Figura 21 – Qualidade dos atributos a partir da ferramenta RapidMiner (parte II).

Em oposição ao Power BI, nas figuras 20 e 21 obtidas ano RapidMiner, os dados são apresentados para que seja feita a avaliação a partir da informação que o sistema resume para o utilizador. O RapidMiner não faz uma auto-avaliação de qualidade geral e não apresenta, como o Power BI, indicadores de cores para identificar eventuais anomalias ou resultados finais de qualidade. A informação apresentada para a avaliação do utilizador são, propriedade dos dados que foi identificado (Real, Inteiro, Nominal), Erro (0), Gráfico de Distribuição dos valores, Mínimo/Máximo, Média, etc. A partir dos valores apresentados podemos também observar que não temos dados ausentes, *outliers*, valores negativos, etc.

Como podemos observar, nas figuras 18, 19, 20 e 21, os dados são apresentados sem erros, sem valores ausentes, sem *outliers* ou valores negativos para atributos pontuação, cotação das alíneas, etc. Em ambas as ferramentas de análise de dados pudemos concluir que toda a preparação dos dados resultou num trabalho que possibilita avançarmos para a fase de Modelação que será apresentada de seguida.

4.5 Modelação



4.5.1 Conceito

Nesta fase é esperada a aplicação de técnicas de modelação para construir um modelo, como por exemplo, árvore de decisão, rede neuronal, etc. Com os dados preparados nas fases anteriores, o próximo passo de construção do modelo servirá para prever a classificação ou resposta ao problema proposto neste projeto. Na fase de modelação constrói-se um modelo, ou uma função que viabilizará a previsão de respostas com o menor custo possível. Antes de se iniciar a construção do modelo, é necessário gerar um procedimento ou mecanismo para testar o modelo no que se refere à qualidade e validação dos dados. Por exemplo, em tarefas supervisionadas de prospeção de dados, como classificação, é comum usar taxas de erro como medida de qualidade nos modelos criados. Portanto, é recomendada a separação do conjunto de dados em conjuntos de treino e de teste, ou seja, constrói-se o modelo com o conjunto de treino e estima-se a sua qualidade no conjunto de teste.

Separar dados em conjuntos de treino e teste é uma parte fundamental da avaliação de modelos de prospeção de dados. A separação de um conjunto de dados num conjunto de treino

e num conjunto de teste tem como objetivo separar a maioria dos dados para treino e uma parte menor dos dados é usada para teste. A partir desta divisão dos dados, quando realizarmos o teste com o conjunto de dados que o modelo não conhece, poderemos avaliar o que o modelo realmente aprendeu com a amostra de dados de treino.

Ao construir e testar modelos, devemos listar as qualidades dos modelos gerados (por exemplo, em termos de precisão) e classificar a sua qualidade em relação aos outros modelos. Devemos repetir o processo sucessivamente observando e fazendo ajustes.

4.5.2 Realização da Modelação

A fase de Modelação foi realizada com recurso a ferramentas que possibilitam a criação de modelos a partir de AM automatizada[10]. Foi utilizado o Microsoft Power BI com o recurso de AutoML[11] , a aplicação MATLAB[12] com o recurso App Toolbox Machine Learning e o software RapidMiner[13] com o recurso *Auto Model*.

Nesta etapa de modelação o trabalho observou as possibilidades de apresentação da recomendação de conteúdo a partir dos objetivos definidos para o projeto. O objetivo de desenvolver um ou mais modelos tem como prioridade avaliar a possibilidade de resolver o problema indicado no objetivo deste projeto, que foca na recomendação de conteúdo para o aluno para que possa motivá-lo na utilização da ferramenta, aumentando seu desempenho de pontos no contexto de gamificação e ou na recomendação de conteúdos que possam lhe ser mais assertivos para que possam obter uma melhor classificação de desempenho.

Com o objetivo de clarificar a realização do trabalho de Modelação, o diagrama apresentado na figura 22 inclui a sequência de passos até ao início da criação dos modelos.

---

[10] A AM automatizada é a automação de processos ou fases que permitem a aprendizagem máquina. a AM automática torna o processo de aprendizagem máquina simples e repetível. Basicamente simplifica o processo de construção de um modelo de AM, automatizando todo o processo. Permite economizar muito tempo e esforço e apresenta os recursos da aprendizagem de máquina (Rapidminer, 2021).

11 Em 2019 a Microsoft disponibilizou no produto Power BI o recurso Auto ML. Com o AutoML no Power BI, analistas de negócios sem grande experiência em AM podem criar modelos de AM para resolver problemas de negócios que antes exigiam cientistas de dados. A maior parte da ciência de dados por detrás da criação dos modelos de AM é automatizada pelo Power BI, ao mesmo tempo que dá visibilidade ao processo usado para criar esse modelo de AM. Outros softwares, assim como o PowerBI também disponibilizam este recurso AutoML como uma extensão ou recurso adicional.

12 O MATLAB foi criado no fim dos anos 1970, entre vários recursos e ferramentas disponíveis, também facilita a AM. Com ferramentas e funções para lidar com big data, bem como aplicações para tornar a AM acessível, o MATLAB é um ambiente ideal para aplicar a AM na sua análise de dados (Mathworks, 2021).

[13] O Rapidminer é um software de ciência de dados desenvolvido pela empresa de mesmo nome e fornece soluções para preparação de conjunto de dados, aprendizagem máquina, prospeção de texto e análise preditiva. O software suporta todas as fases do processo de AM, incluindo preparação de dados, visualização de resultados, validação e otimização de modelos.

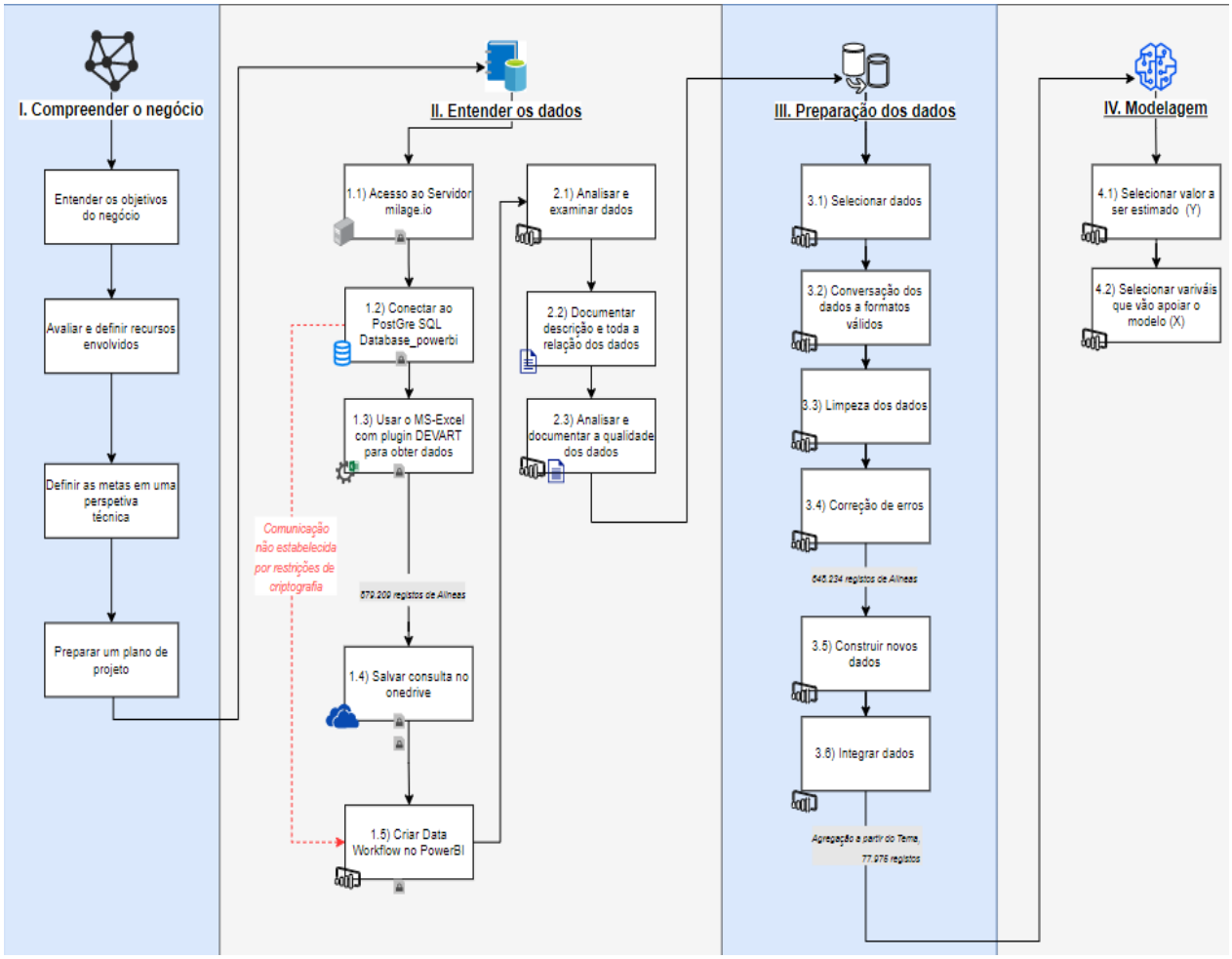


Figura 22 - Realização das fases a partir da Metodologia CRISP-DM (Felipe Fonseca, 2021).

Como apresentado neste capítulo e a partir do diagrama da figura 22, nota-se que até chegar à fase de Modelação há um conjunto vasto de ações que precisam ser realizadas para se conseguir estabelecer o melhor conjunto de dados possível.

4.6 Recursos utilizados

4.6.1 Power BI Auto ML

Toda a fase de Análise para compreender e preparar os dados foi realizada usando os recursos do Power BI, com licença premium, ligados diretamente à base de dados do MILAGE Aprender+. Nesta secção é apresentado como foram realizados os passos na ferramenta.

Com a Amostra III organizada e orientada dimensões de resultados das Alineas e Alunos e tratando-se de um grande volume de informação, esta ferramenta foi utilizada porque possibilita a apresentação visual dos dados. Com o Power BI temos a possibilidade de explorar

e analisar toda a amostra de dados, identificando inclusive valores discrepantes ou fora de um intervalo de valores esperados (*outliers*[14]).

Usando a ferramenta de transformação de dados do Power BI com recurso à abordagem de AUTO ML ou AM Automatizada, toda a fase de preparação dos dados foi realizada, criando um fluxo para garantir que as ações de limpeza, correção, construção e agregação de dados sejam um processo automático. A figura 23 apresenta a criação do fluxo de dados no Power BI, onde se identifica a obtenção dos dados, ou seja, toda a Amostra III do Milage Aprender+, sendo a partir deste fluxo que se realiza toda a configuração da rotina de transformação dos dados.

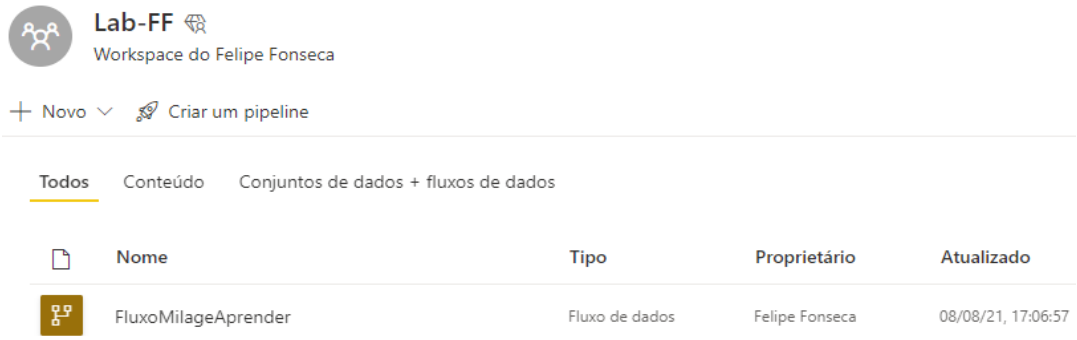


Figura 23 - Fluxo de dados no Power BI.

Conforme apresentado na figura 24 o processo dentro da ferramenta Power BI é totalmente guiado por instruções e conjunto de orientações para se usar a ferramenta.

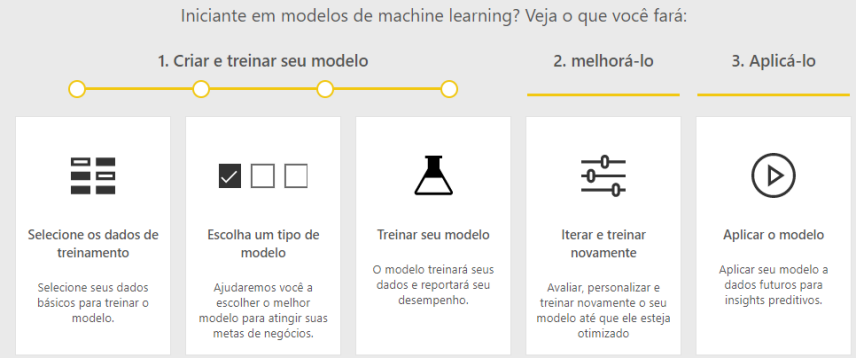


Figura 24 - Tutorial de apresentação do recurso AutoML do Power BI.

[14] Num conjunto de dados podem existir valores extremos que estão fora de um intervalo do que é esperado. Eles são chamados de *outliers* e, frequentemente, a Modelação de AM e a habilidade do modelo em geral podem ser melhoradas com a compreensão e até mesmo a remoção desses valores *outliers*. Um *outlier* é uma observação (um dado/valor) improvável num conjunto de dados. Pode ter uma de muitas causas para existirem, como por exemplo: Falha em um sensor de temperatura que obtém um registo de uma temperatura ambiente de 500º C em uma estação do Polo Norte.

Com o fluxo de dados pronto para utilização e a devida validação de qualidade dos atributos (colunas) e registos (linhas) foi criado um Modelo de AM. Na figura 25 é apresentado a opção de criação de um Modelo de AM a partir de um conjunto de dados.

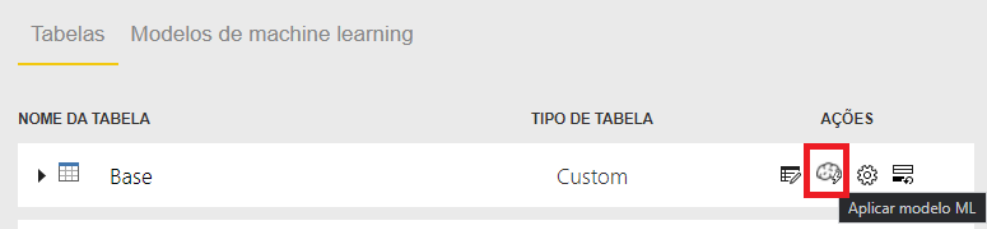


Figura 25 - Aplicar modelo ML ao conjunto de dados preparado.

O primeiro passo do processo é definir o conjunto de dados (tabela) onde se aplica o processo de AM e de seguida selecionar o atributo resposta (Y). Para o problema apresentado no âmbito deste trabalho, o atributo (coluna) que queremos prever é o atributo "*AlunoQtdPontosRealizadoTema*" e "classificação". A primeira variável reúne o aproveitamento que cada um dos alunos obteve para cada um dos temas. Ao usarmos esta coluna, o objetivo é criar um modelo que consiga prever, antes do aluno realizar determinado tema, o total de pontos que este aluno obterá. Para a segunda, Classificação, o objetivo é agrupar é prever os resultados por grupos de classificação.

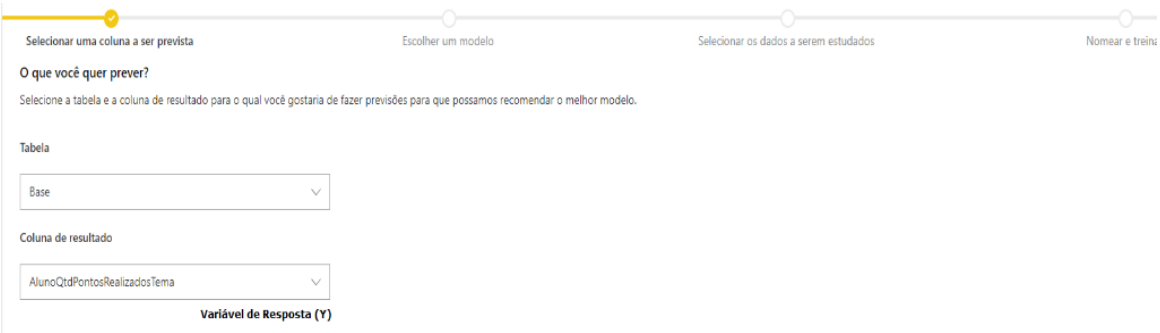


Figura 26 - Seleção da coluna de resultado (Y).

Para o conjunto de dados dos Temas, como apresentado na revisão da literatura, no capítulo 3, o atributo de resposta (Y) possui uma saída numérica, pois representa os pontos dos alunos. Logo para este tipo de problema o modelo recomendado é de regressão. A própria ferramenta do Power BI recomenda a aplicação deste tipo de modelo conforme se apresenta na figura 27. Assim como para o conjunto de dados Alunos e Escola, onde temos o atributo de resposta (Y) Classificação como categórico, o tipo de modelo recomendado pela ferramenta é o de classificação.

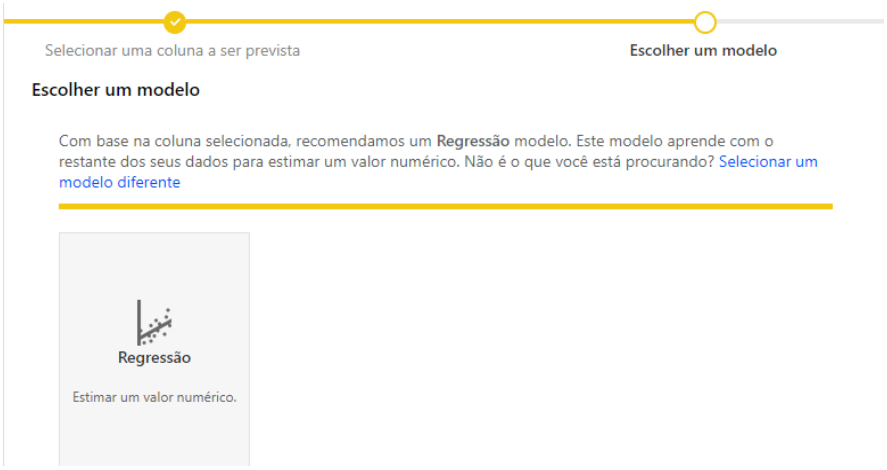


Figura 27 - Exemplo Escolha do modelo de regressão linear para estimar valor numérico.

Depois da escolha do modelo, os atributos independentes (X) são fundamentais para estabelecer uma função. A ação realizada nesta fase resume-se a selecionar os atributos que possibilitem a criação de um modelo com maior capacidade de generalização.

Todo o processo de seleção dos atributos independentes (X) também é realizado a partir da ferramenta. A escolha dos atributos (X) que serão utilizados no modelo envolve uma avaliação de correlação[15] para que o modelo possa ter um desempenho mais eficiente.

Embora a própria ferramenta realize uma validação de correlação no momento de escolha dos atributos, este é um processo muito importante para a execução do modelo. Realizar uma análise de correlação possibilitará o uso de atributos (X) que realmente tenham importância e relevância para o modelo de previsão (Y). Após a seleção dos atributos, será definido o limite de tempo para realizar o treino.

No Power BI o processo de divisão e separação dos dados é também realizado automaticamente. No conjunto de dados é realizada uma separação aleatória de 80% para treino. No final, os restantes 20% são utilizados para teste. Com os resultados dos testes será possível apresentar informação sobre a precisão do modelo.

A fase de divisão é essencial, a separação de dados permitirá a validação do modelo gerado e garantirá que o modelo realmente aprende a prever respostas e não apenas a memorizá-las.

[15] A correlação indica a interdependência entre duas ou mais variáveis. Também chamado de "coeficiente de correlação produto-momento" ou simplesmente de "ρ de Pearson" mede o grau da correlação (e a direção dessa correlação — se positiva ou negativa) entre duas variáveis. Este coeficiente, normalmente representado por ρ assume apenas valores entre -1 e 1. A fórmula desenvolvida por Karl Pearson, há mais de 120 anos, continua a ser a mais utilizada para o cálculo da correlação. É muito importante observar que mesmo uma correlação forte não implica que x "causa" y.

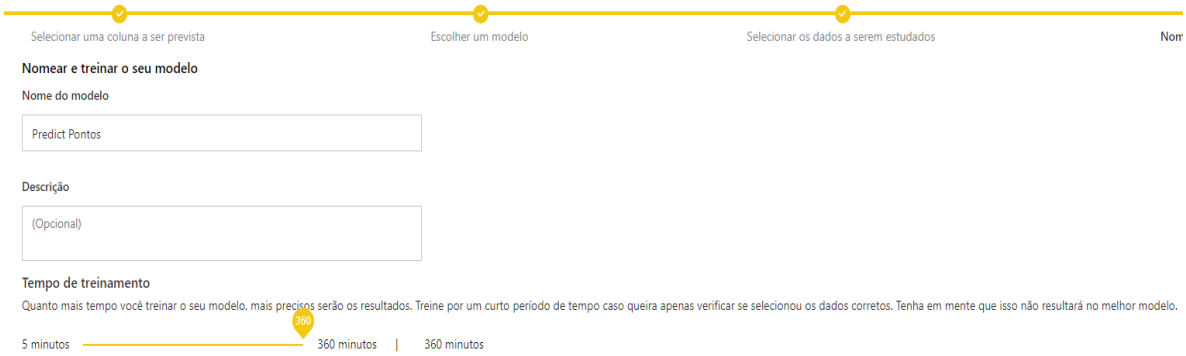


Figura 28 – Definindo o tempo de treinamento.

Com as definições concluídas, a ferramenta iniciará o processo de aprendizagem. A partir do tempo definido, a ferramenta irá realizar diversas iterações utilizando os atributos (X) e procurando a melhor generalização de um modelo e função para prever (Y).

O Power BI já reúne uma ampla biblioteca de Algoritmos de AM. Durante este período de processamento, vários algoritmos serão processados. Todos os algoritmos estão reunidos no ambiente Microsoft Azure, os dados serão processados no ambiente computacional da nuvem[16] a partir de inúmeros algoritmos já avaliados e já aplicados para diversos tipos de problemas, mas que, entretanto, poderão ser utilizados como ponto de partida para a resolução do problema apresentado neste trabalho. Os resultados das experiências realizadas no Power BI serão apresentados no Capítulo 5 como análise de resultados.

4.6.2 Matlab R2021b App Toolbox Machine Learning

O Matlab é reconhecido como um software interativo de alto desempenho para a realização de cálculos. O software permite a resolução de muitos problemas numéricos com baixo tempo de execução. O software dispõe também de algumas extensões, chamadas de caixas de ferramentas (Toolbox). São várias as extensões disponíveis, entre elas a *Statistics and Machine Learning Toolbox*, *Classification Learner*, traduzido para o português Estatísticas e AM e Aprendiz de Classificação. Esta extensão fornece funções e aplicações para descrever, analisar e modelar dados. Há algoritmos de regressão e classificação que permitem realizar inferências de dados e criar modelos preditivos iterativamente, usando as aplicações de

[16] A computação em nuvem (traduzido do inglês, *cloud computing*) é o fornecimento de serviços informáticos, incluindo servidores, armazenamento, bases de dados, rede, software, análises e inteligência, através da Internet ("a cloud") para disponibilizar mais rapidamente inovação, recursos flexíveis e poupanças no dimensionamento. Normalmente, paga apenas pelos serviços cloud que utiliza, o que o ajuda a reduzir os custos de funcionamento, executar a infraestrutura de forma mais eficaz e dimensionar à medida que a sua empresa precisa de mudar. (Microsoft, 2021)

classificação e regressão linear, ou programaticamente, usando AutoML. O Matlab e a ferramenta de AM possibilitam comparar vários algoritmos. (Mathworks, 2021).

O Matlab foi usado apenas na fase de criação do modelo. Todo as fases anteriores, foram realizadas, conforme apresentado anteriormente no Power BI. Com os dados já preparados, a ferramenta Matlab foi iniciada, usando em específico a extensão *Toolbox Machine Learning de AutoML* (figura 29).

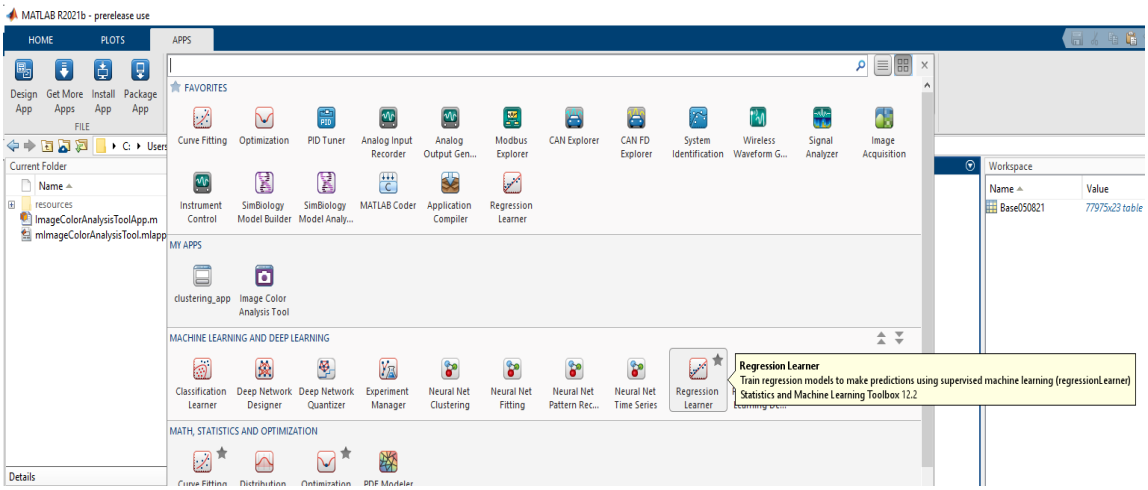


Figura 29 – Matlab regression learner Toolbox

Ao importar cada respetivo conjunto de dados Temas, Alunos e Escolas, já previamente preparado nas fases anteriores, há necessidade de conferir e definir o tipo de atributos (coluna X), indicando se são dados numéricos, texto, etc. Depois da importação do conjunto de dados, ao criar uma nova sessão dentro da aplicação *Regression Learner* ou Classification Learner selecionámos a coluna de resposta (Y).

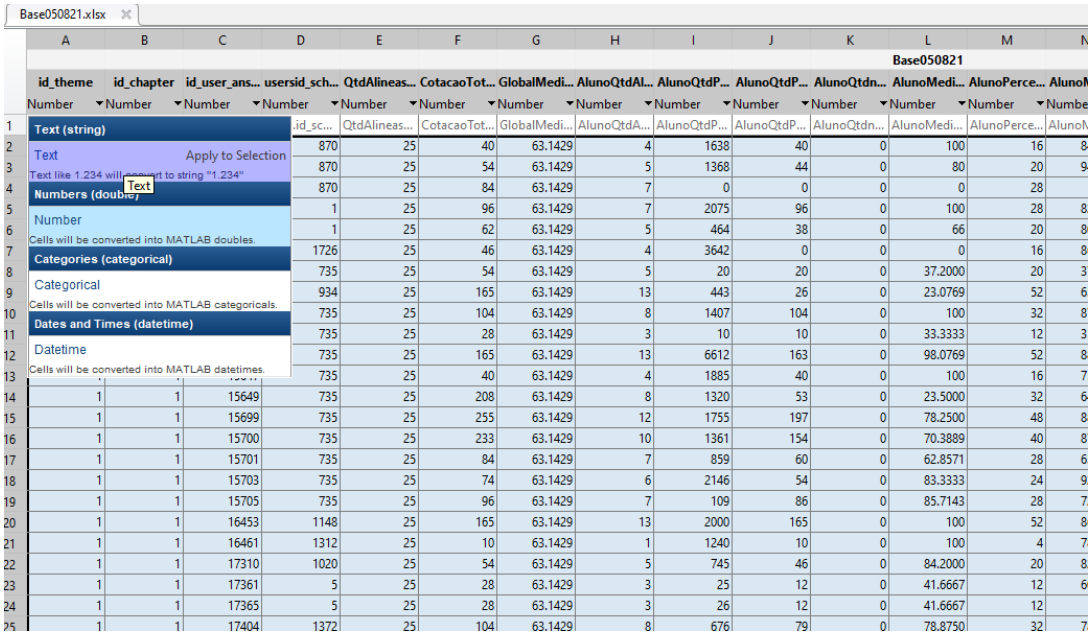


Figura 30 – Importando conjunto de dados para criar um modelo.

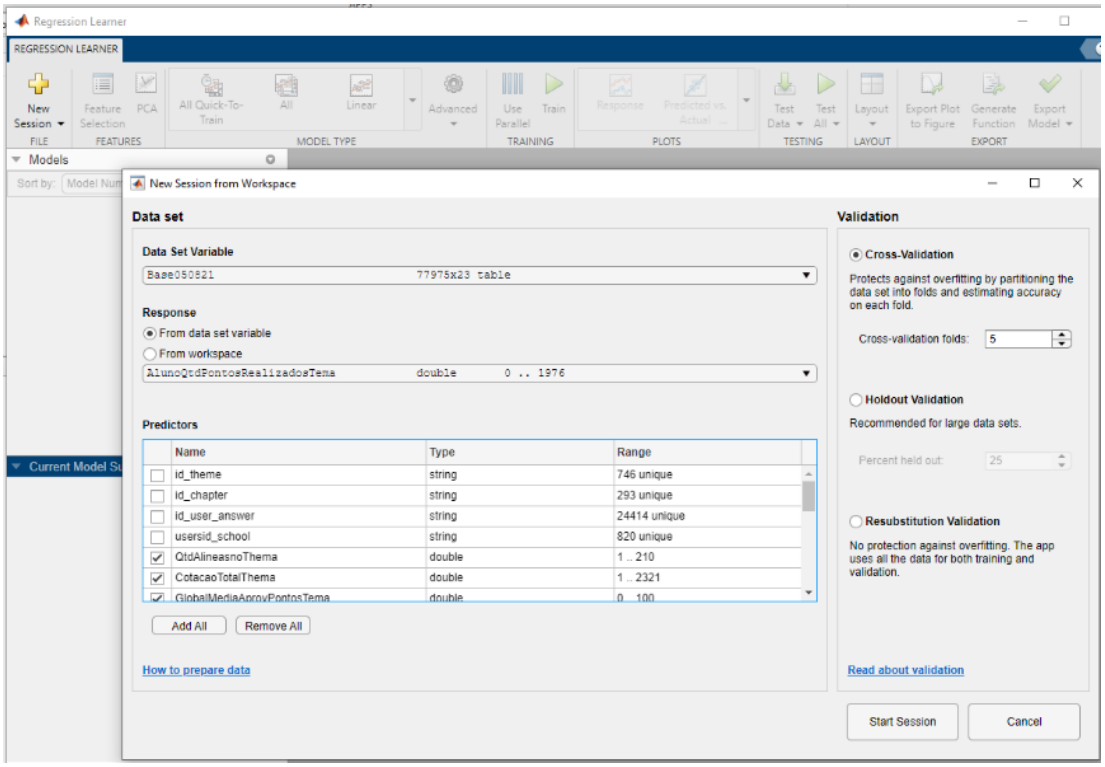


Figura 31 – Selecionando as colunas (X) no Matlab Regression Learner Toolbox para Regressão.

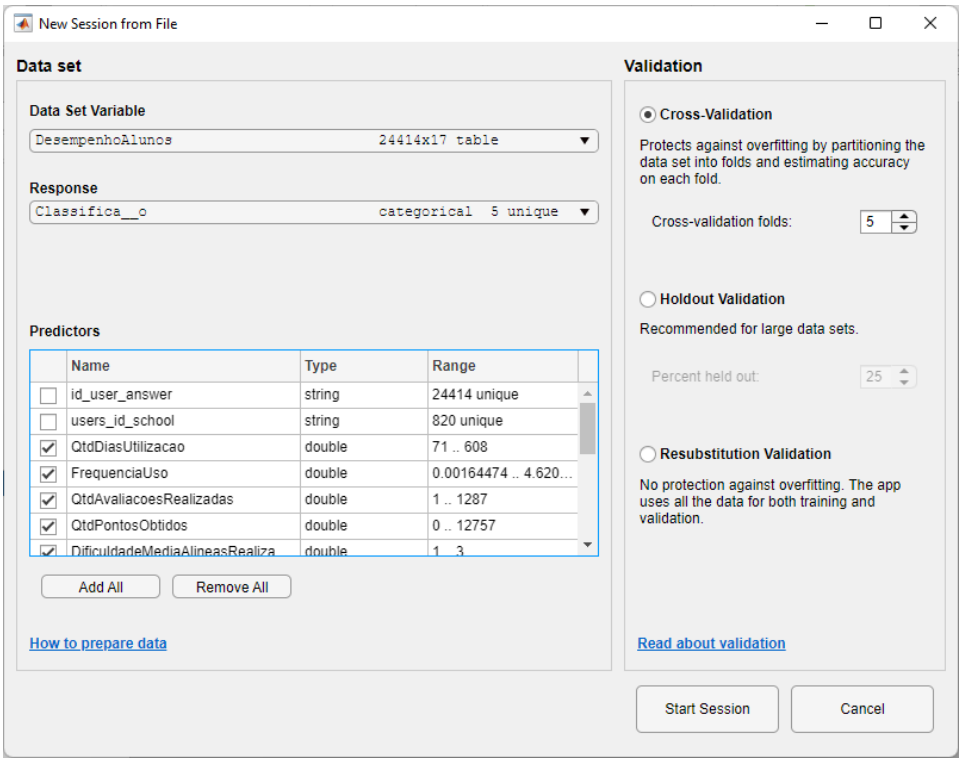


Figura 32 – Selecionando as colunas (X) no Matlab Regression Learner Toolbox para Classificação.

Na opção de Preditores escolhemos as colunas (x) que já foram preparadas para o conjunto de dados. A separação dos dados é realizada a partir da ativação da Validação Cruzada, fazendo com que o conjunto de dados seja separado em dados de treino e dados de teste. De

seguida configurou-se o Matlab para realizar o exercício de testar modelos. Esta opção permite testar todos os modelos disponíveis. É preciso considerar que quanto mais modelos são escolhidos para teste, maior será o tempo investido até à finalização de todo o processo. Os resultados da experiência realizada no Matlab serão apresentados no Capítulo 5 como análise de resultados.

4.6.3 RapidMiner Studio 9.9

O software RapidMiner Studio 9.9 também foi utilizado na fase de Modelação. Com todo o conjunto de dados preparado, a ferramenta foi utilizada com recurso ao Auto Model. O Auto Model é um recurso existente dentro do RapidMiner Studio que agiliza o processo de construção e validação de modelos. Com este recurso Auto Model foi possível também avaliar e preparar o conjunto de dados do MILAGE Aprender+ e criar os modelos a partir de modelos sugeridos pelo RapidMiner para a solução do problema. Por tratar-se de um recurso de construção de AM Automatizada, todo o processo é guiado e orientado pela própria ferramenta. A partir da entrada de dados, ligado à base de dados do MILAGE Aprender+, a ferramenta oferece a opção, além da análise e visualização dos dados, de criar um modelo automatizado (auto Model).

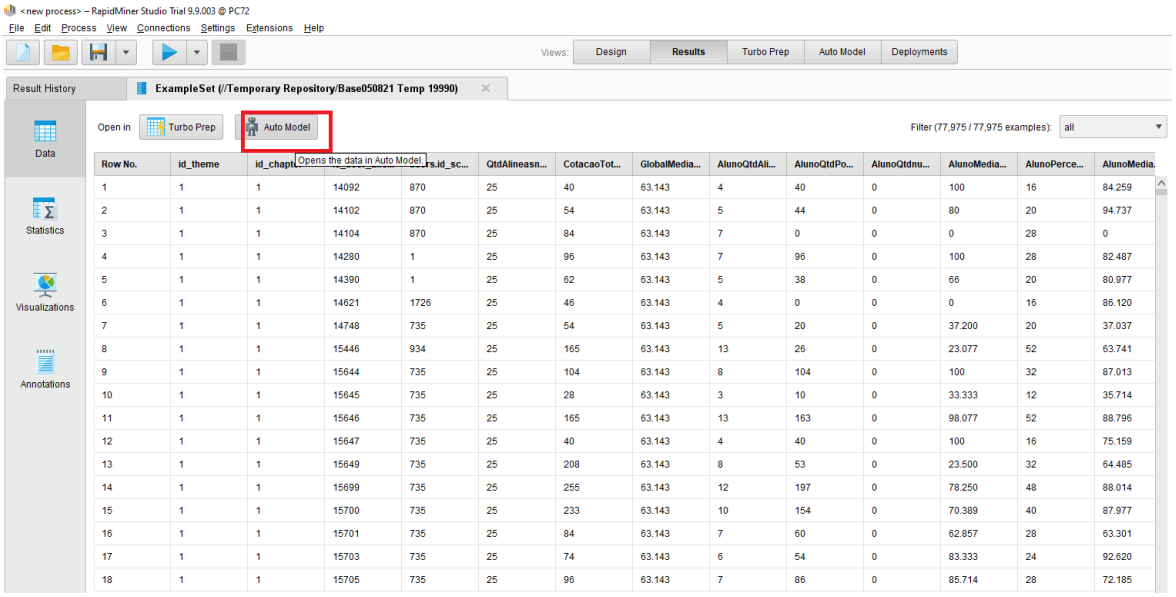


Figura 33 – Carregando conjunto de dados e iniciando Auto Model.

De seguida o utilizador pode optar por escolher o tipo de problema ou modelo a ser construído, selecionado também a coluna de resposta. Como apresentado na revisão da literatura, no capítulo 3, o atributo de resposta (Y) para a amostra de Tema possui uma saída

numérica, pois representa os pontos dos alunos. Logo para um problema com estas características o modelo recomendado é de regressão.

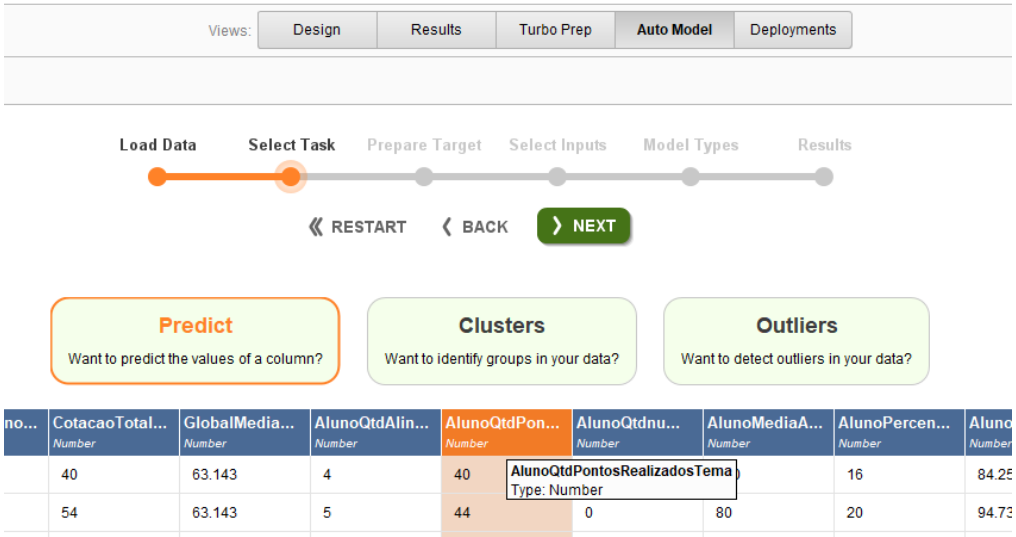


Figura 34 – Escolhendo o modelo Preditivo e coluna de resposta (Y).

Assim como nas ferramentas apresentadas anteriormente, após selecionar o tipo de modelo a construir e escolher a resposta (Y), "*AlunoQtdPontosRealizadoTema*" e "Classificação", há necessidade de selecionar as variáveis independentes (X). O RapidMiner oferece um recurso muito importante para a construção do modelo pois apresenta em detalhe a qualidade de cada atributo (X). Conforme apresentado na figura 35, a ferramenta disponibiliza ao utilizador a opção de escolher os atributos (X) e também já apresenta indicadores de qualidade, cujo significado se descreve na tabela 21.

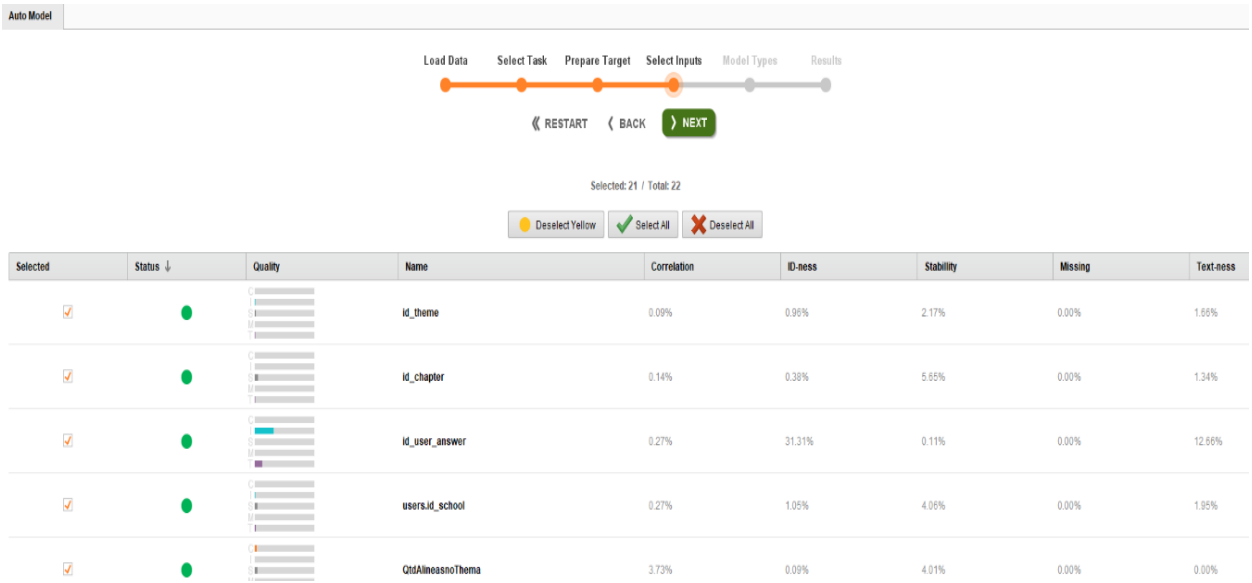


Figura 35 – Escolhendo colunas (X) e avaliando qualidade dos dados.

Critério de Qualidade	Significado
Correlação (C):	Mede a correlação linear entre a coluna de dados e a coluna de destino. Esta barra de qualidade só está disponível quando a tarefa é "Prever".
ID-ness (I):	Mede o grau em que este Atributo se assemelha a um ID. O número de valores diferentes para o Atributo dividido pelo número de linhas de dados.
Estabilidade (S):	Mede o quão estável ou constante esta coluna é. O número de linhas com o valor não omisso mais frequente dividido pelo número total de linhas de dados com valores não omissos.
Em falta (M):	O número de valores em falta nesta coluna como uma fração do número total de linhas de dados.
Text-ness (T):	Esta é a média do ID-ness, a fração de células contendo limitadores de token e uma pontuação baseada no comprimento do conteúdo da célula.

Tabela 21 – Significado dos indicadores de qualidade das colunas no RapidMiner.

Como último passo, apresentado na figura 35, antes de iniciar a execução da Aprendizagem Automatizada, podemos escolher qual ou quais modelos de aprendizagem serão executados para obter o melhor modelo preditivo para a coluna de resposta (Y), a partir do conjunto de dados e atributos (X). Todo o processamento é realizado localmente, usando os recursos da máquina de trabalho, em oposição ao Microsoft Power BI em que a execução ocorre num ambiente computacional da nuvem. O processamento local foi realizado num computador portátil com a configuração apresentada na tabela 22. Os resultados obtidos da experiência no RapidMiner serão apresentados no Capítulo 5 de análise de resultados.

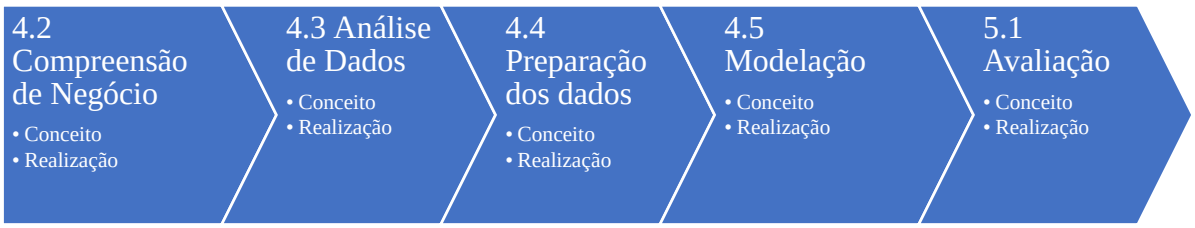
Processador	Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz 2.30 GHz
RAM instalada	16,0 GB (15,8 GB utilizável)
Tipo de Sistema	Sistema operativo de 64 bits, processador baseado em x64

Tabela 22 – Configuração de Hardware do computador portátil.

Capítulo 5: Análise de resultados

Neste capítulo serão apresentados os resultados da etapa de Avaliação. Após todas as etapas anteriores de Compreensão do Negócio, Análise de Dados, Preparação dos dados e Modelação, o capítulo 5 apresentará os resultados juntamente com o conceito da metodologia CRISP-DM.

5.1 Avaliação



5.1.1 Conceito

Utilizando o resultado da fase de Modelação, cada experiência produz um modelo para uma avaliação de classificação. A avaliação do modelo é realizada a partir dos resultados obtidos no conjunto de teste. A avaliação deve ter em consideração a precisão de cada modelo e comparação de qualidade com todos aos modelos criados. A precisão refere-se à capacidade de o modelo ser genérico. Será nesta fase de avaliação que observamos o grau com que o modelo responde aos objetivos definidos na primeira fase.

Nesta fase também será realizada uma avaliação da qualidade do modelo, verificando ponto a ponto todas as fases realizadas a fim de garantir que o modelo foi construído corretamente. Para fazer uma avaliação da precisão dos modelos, é fundamental estabelecer uma métrica que permita analisar as diferenças entre os valores estimados pelo modelo e os valores reais observados.

No final da avaliação será produzido um relatório com um resumo das experiências realizadas contendo os resultados de cada modelo.

5.1.2 Medidas para a avaliação da qualidade dos modelos

- O Desvio Médio Quadrático, do inglês *Root Mean Square desviation* (RMSD), é frequentemente utilizado para avaliar os desvios ou diferenças entre os resultados de um modelo com dados estimados e os resultados reais. Esta métrica ou o Erro Quadrático Médio, do inglês *Root Mean Square Error* (RMSE) são as medidas comuns mais aplicada.

Na aprendizagem máquina, é extremamente útil ter uma única métrica para avaliar o desempenho de um modelo, seja durante o treino, validação cruzada ou monotorização após a implementação (C3.ai, 2021).

Para calcular o RMSE é determinado o valor residual, ou seja, a diferença entre a previsão (valor estimado) e o valor real para cada dado. Depois é calculado a norma do resíduo ou erro para cada dado. Finalmente, é calculado a média dos resíduos e determinado a raiz quadrada dessa média. O RMSE é muito comum em aplicações de aprendizagem supervisionada, pois o RMSE usa e precisa de medições verdadeiras para cada dado previsto.

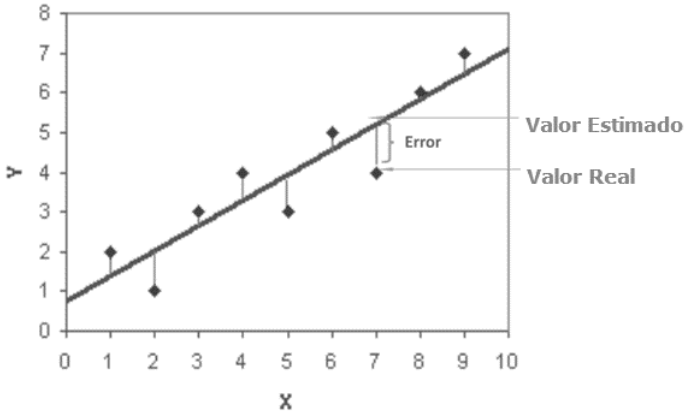


Figura 36 – Resíduo / Erro entre o valor estimado versus o valor real

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

Figura 37 - Representação da expressão RMSE

O Erro Quadrático Médio pode ser especificado pela fórmula apresentada na figura 38, sendo que N é o número de eventos de dados, y (i) é a i-ésima medição e y ̂(i) é a previsão do valor estimado correspondente. Quanto menor o valor do RMSE, maior é a capacidade de um determinado modelo se "ajustar" a um conjunto de dados. Uma forma de visualizar o desempenho de um modelo de classificação é a partir de uma matriz de confusão.

Valor real / Valor previsto	Positivo	Negativo
Positivo	Verdadeiro Positivo (TP)	Falso Positivo (FP)
Negativo	Falso Negativo (FN)	Verdadeiro negativo (TN)

Figura 38 - Exemplo Matriz de confusão

Esta matriz indica quantos exemplos existem em cada classe: falso positivo (FP), falso negativo (FN), verdadeiro positivo (TP) e verdadeiro negativo (TN). É essencial identificar o número de observações nas diferentes classes, tanto em números absolutos quanto em percentagens, já que o número de exemplos em cada classe pode variar. A matriz de confusão possibilita identificar quantos exemplos foram classificados de forma correta e também incorreta em cada uma das classes, também ajuda a perceber se o modelo está favorecendo uma classe em detrimento da outra. É importante, principalmente nas situações em que os erros possuem custos diferentes. Um exemplo, seria um modelo para classificar exames de diagnóstico de uma doença. Quando o erro for falso positivo (classificar um paciente como doente, quanto está saudável) seria inconveniente dado que ele demoraria mais a receber alta, o erro por falso negativo (classificar um paciente como saudável, quanto está doente) seria muito mais grave, dado que ele poderia receber alta e não ter o acompanhamento necessário.

Dado um modelo que atribui uma probabilidade para a classe positiva, é necessário definir um limiar de classificação. Acima deste limiar, um exemplo é classificado como positivo, caso contrário, é classificado como negativo. O limiar de classificação influencia o valor das métricas mencionadas anteriormente (precisão, etc), e a sua escolha deve ter em consideração o custo de cada erro. A curva ROC (traduzido do inglês *Receiver Operating Characteristic*) é utilizada para avaliar o desempenho de um classificador para diferentes limiares de classificação, medindo a Taxa de Falso Positivo (FPR — False Positive Rate) e também a Taxa de Verdadeiro Positivo (TPR — True Positive Rate) para cada limiar de classificação possível. A precisão pode ser calculada conforme a expressão abaixo. (Robert Tibshirani, Trevor Hastie, & Jerome Friedman, 2008)

$$\text{Precisão} = \frac{TP + TN}{TP + FP + TN + FN} = \frac{\text{Previsões corretas}}{\text{Todas as previsões}}$$

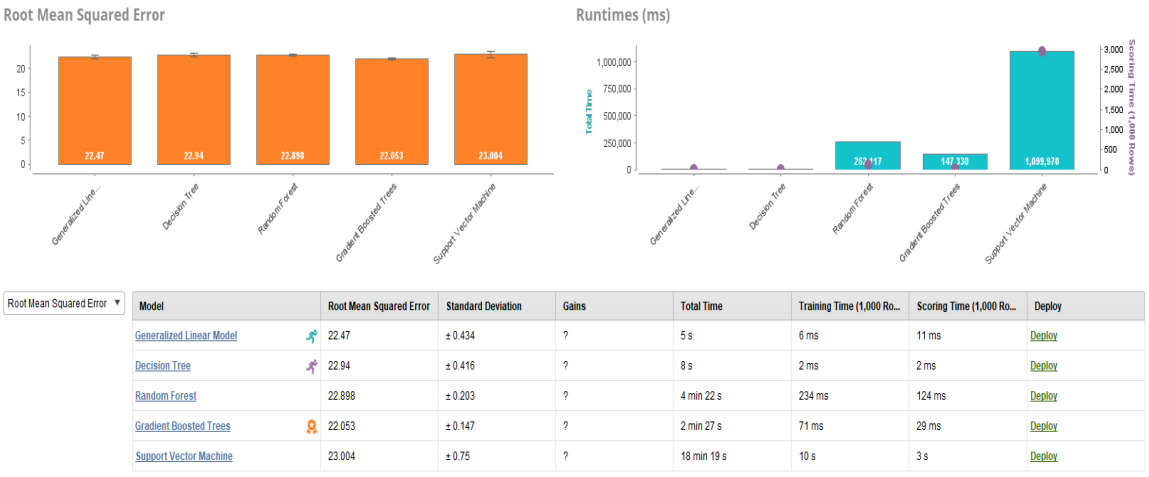
Figura 39 – Expressão para calculo de Precisão

5.1.3 Realização

5.1.3.1 Regressão

O resultado das experiências com os modelos usando os recursos de AUTO ML providenciados pelo software Matlab, RapidMiner e Power BI são apresentados de seguida, iniciando pela apresentação dos resultados de Regressão e posteriormente de Classificação.

Figura 40 – Apresentação dos resultados no Matlab , regressão para Pontuação.



Model	Root Mean Squared Error	Standard Deviation	Gains	Total Time	Training Time (1,000 Ro...	Scoring Time (1,000 Ro...	Deploy
Generalized Linear Model	22.47	± 0.434	?	5 s	6 ms	11 ms	Deploy
Decision Tree	22.94	± 0.416	?	8 s	2 ms	2 ms	Deploy
Random Forest	22.898	± 0.203	?	4 min 22 s	234 ms	124 ms	Deploy
Gradient Boosted Trees	22.053	± 0.147	?	2 min 27 s	71 ms	29 ms	Deploy
Support Vector Machine	23.004	± 0.75	?	18 min 19 s	10 s	3 s	Deploy

Figura 41 – Apresentação dos resultados no RapidMiner, regressão para Pontuação

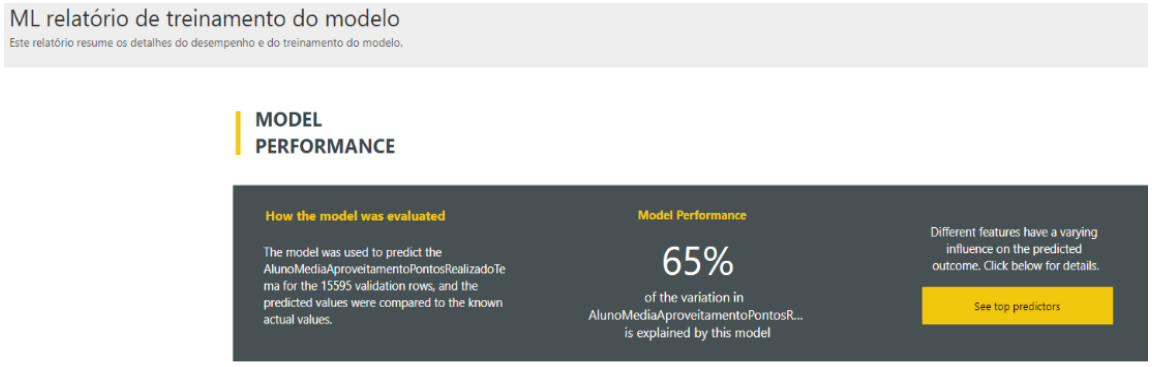


Figura 42 – Apresentação dos resultados no Power BI, regressão para Pontuação.

Os resultados apresentados nas figuras anteriores constituem evidência de como os resultados foram obtidos nas três ferramentas Matlab, RapidMiner e Power BI respetivamente. Nos resultados finais do Power BI, de acordo com a figura 44, os valores de RMSE não foram apresentados e, portanto, foi necessário realizar o cálculo manual (usando a expressão matemática de RMSE) usando o Power BI Desktop como ferramenta de cálculo.

Contagem de id_user_answer	id_chapter	Média de AlunoMediaAproveitamentoPontosRealizadoTema	Média de ML.RegressionResult	DIFF	RMSE
1	177	0,00	0,00	0,0000	0,00
1	144	38,00	38,00	0,0000	0,00
1	141	100,00	100,00	0,0000	0,00
1	199	100,00	100,00	0,0000	0,00
1	271	91,00	91,00	0,0000	0,00
1	270	0,00	0,00	0,0000	0,00
2	5	0,00	0,00	0,0000	0,00
1	269	0,00	0,00	0,0000	0,00
1	7	92,00	92,00	0,0000	0,00
1	140	100,00	100,00	0,0000	0,00
1	197	57,00	57,00	0,0000	0,00
1	8	98,00	98,00	0,0000	0,00
3	267	0,00	0,00	0,0000	0,00
77975		**58,77**	**58,73**		**20,98**

20,98
RMSE

Figura 43 – RMSE calculado a partir dos dados do Power BI

Com o intuito de facilitar a visualização e interpretação dos resultados de forma padronizada e usando o mesmo critério de avaliação, foi produzida a tabela 23.

RMSE por Ferramenta de AUTO ML									
Modelo	**Power BI**			**RapidMiner**			**Matlab**		
	Tempo (s)	Linhas (observações)	RMSE	Tempo (s)	Linhas (observações)	RMSE	Tempo (s)	Linhas (observações)	RMSE
Light GBM Regressor	2400	15595	20,98	-	-	-	-		-
Generalized Linear Model	-	-	-	5	1000	22,47			
Decision Tree	-	-	-	8	1000	22,94			
Random Forest	-	-	-	262	1000	22,90			
Gradient Boosted Trees	-	-	-	147	1000	22,05			
Suppor Vector Machine	-	-	-	1099	1000	23,00			
Linear Regression (Linear)								10000	22,38
Linear Regression (Interactions Linear)								10000	22,29
Linear Regression (Robust Linear)								10000	22,67
Stepwise Linear Regression								10000	22,30
Tree (Fine tree)								10000	27,69
Tree (Medium Tree)								10000	25,03
Tree (Coarse Tree)								10000	23,19
SVM (Linear SVM)								10000	22,58
SVM (Quadratic SVM)								10000	22,60
SVM (Cubic SVM)								10000	22,65
SVM (Fine Gaussian SVM)								10000	24,24
SVM (Medim Gaussian SVM)								10000	22,73
SVM (Coarse Gaussian SVM)								10000	22,58
Ensemble (Boosted Trees)								10000	22,36
Ensemble (Bagged Trees)								10000	22,17
Gaussian Process Regression (Squared Exponential GPR)								10000	22,30
Gaussian Process Regression (Matern 5/2 GPR)								10000	22,31
Gaussian Process Regression (Exponential GPR)								10000	21,91
Gaussian Process Regression (Rational Quadratic GPR)								10000	22,30

Tabela 23 – Resultados geral dos Modelos a partir do Power BI, RapidMiner e Matlab para Regressão

Pontuação.

A partir dos resultados apresentados na tabela 23, destaca-se o melhor resultado obtido no Power BI com o modelo Light GBM Regressor[17], consiste num modelo que se generaliza com o resultado RMSE de 20.98, após um total de 15595 observações no treino. Na apresentação padrão dos resultados, o Power BI apresenta um indicador de 65% em relação à performance do modelo apresentado.

Como também pode ser observado, o Power BI apresenta apenas o resultado final, não indicando os resultados detalhados para que seja possível calcular o RMSE dos restantes modelos avaliados. Apenas se apresenta a qualidade do modelo conforme o gráfico da figura 42. Observámos que não é possível parametrizar o modelo que é apresentado como resultado

[17] O modelo Light GBM é uma estrutura de aumento de gradiente que usa algoritmo de aprendizagem baseado em árvore de decisão. No Light GBM a árvore cresce verticalmente enquanto em outros algoritmos a árvore cresce horizontalmente, o que significa que no Light GBM a árvore cresce em forma de folha enquanto noutro algoritmo cresce ao nível da árvore. O Light GBM tem o prefixo 'Light' devido à sua alta velocidade. O GBM leve pode lidar com um grande volume de dados e requer menos memória para ser executado. Outra razão pela qual o Light GBM é popular é porque se concentra na precisão dos resultados. O LGBM também oferece suporte a aprendizagem por GPU e, portanto, os cientistas de dados usam amplamente o LGBM para o desenvolvimento de aplicações de ciência de dados (Mandot, 2017).

final. Também não se consegue exportar o modelo para execução noutro ambiente, tornando apenas possível a execução e apresentação dentro da própria ferramenta Power BI.

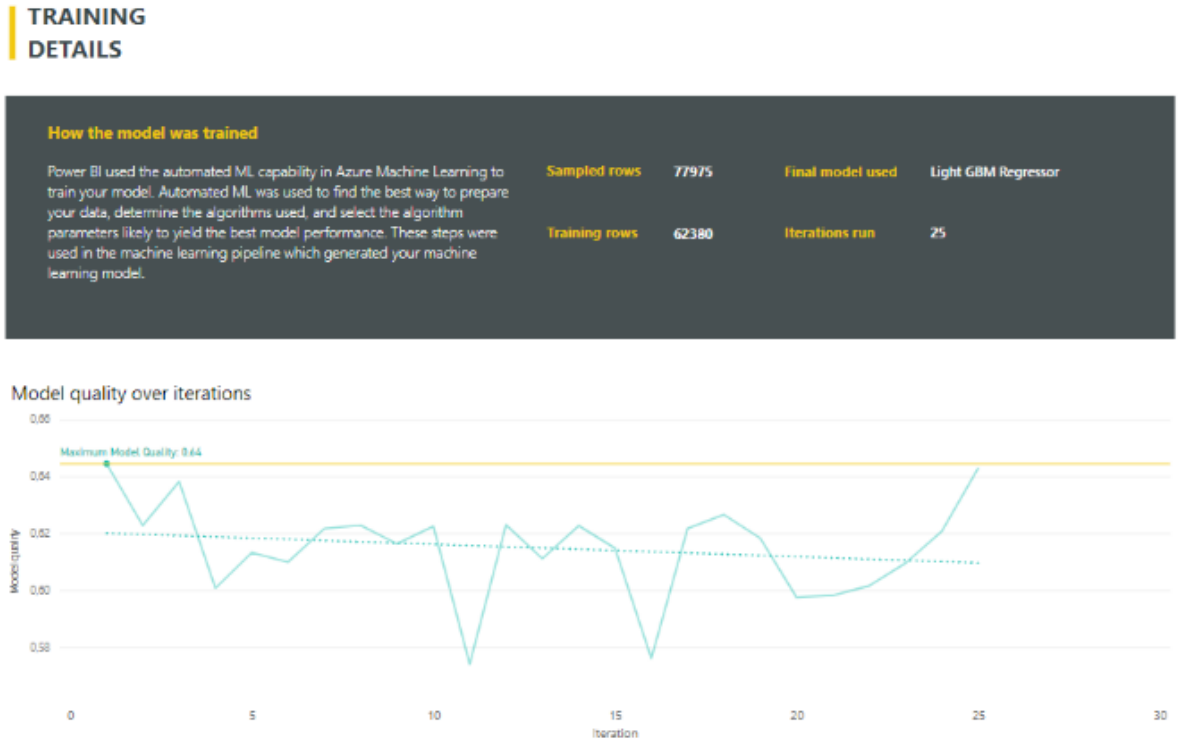


Figura 44 – Apresentação do desempenho dos modelos PowerBI.

Além do resultado do modelo, é também apresentado o resultado de influência dos atributos (x) utilizados na modelação. Como podemos perceber pela figura 45, o atributo de maior influência foi construído a partir da engenharia de recursos: Aluno – Média de Aproveitamento de Pontos e a Média de Aproveitamento de pontos Global.

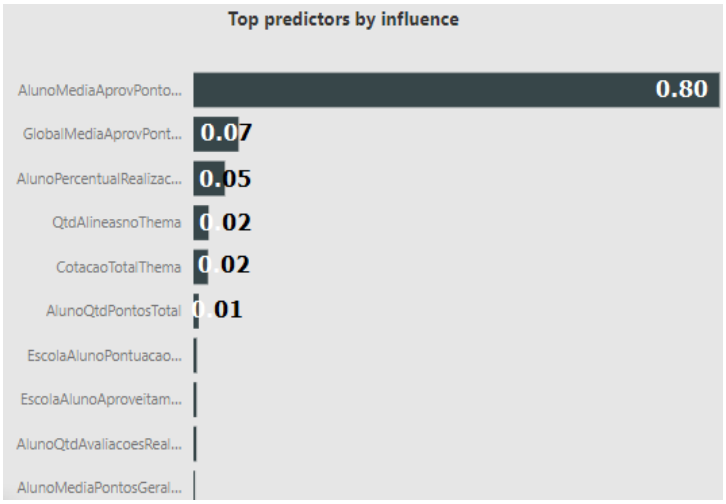


Figura 45 – Apresentação dos atributos com maior influência no Modelo.

Os resultados das outras duas ferramentas, RapidMiner e Matlab também são referidos e de forma detalhada, apresentando todos os modelos que as ferramentas usaram a partir da sua biblioteca de modelos.

Nota-se ainda que o RapidMiner, embora seja mais amigável e fácil de usar, o melhor resultado da ferramenta apresenta performance inferior quando comparado às outras ferramentas. Destaca-se também que o total de linhas utilizadas foi inferior no processo de treino quando comparado ao Matlab e Power BI, treinando com apenas 1000 linhas. A ferramenta RapidMiner possibilita, a exportação e desenvolvimento do modelo para que possa ser incorporado noutras plataformas com recurso a serviços web *(Webservices).*

No Matlab, existe uma ampla biblioteca de modelos, podendo o utilizador escolher os modelos a aplicar no treino. Contudo, o Matlab permite parametrizar ou customizar as configurações finais do modelo além de possibilitar exportar o modelo para outras plataformas.

5.1.3.2 Classificação

De seguida são apresentados os resultados para as experiências de Classificação para o conjunto de dados relacionado comos Alunos.

A figura 46 apresenta os resultados obtidos a partir do Rapidminer, destacando o Generalized Linear Model, como modelo com melhor desempenho.

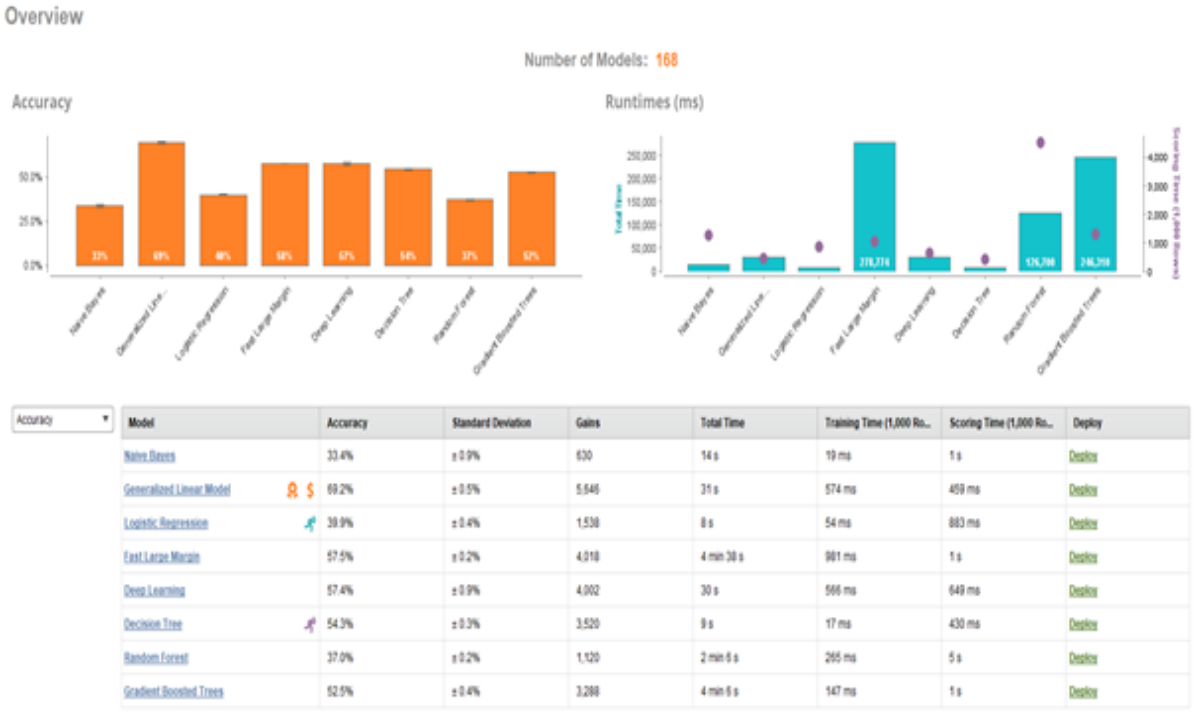


Figura 46 – Apresentação dos resultados no RapidMiner, Modelo Classificação para a variável Classificação.

A figura 47 apresenta os resultados obtidos a partir do Microsoft Power BI.

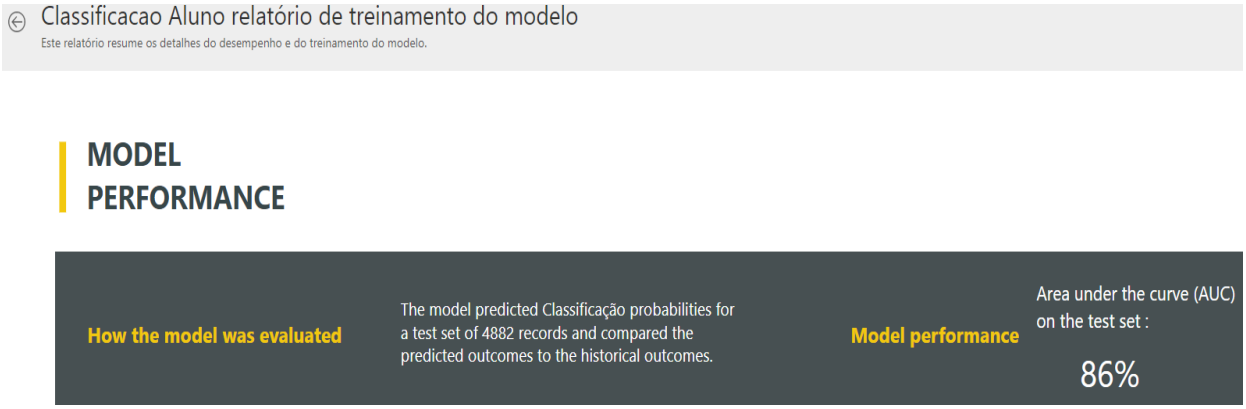


Figura 47 – Apresentação dos resultados no Power BI, Modelo *Classificação para* a variável *Classificação.*

A figura 48 apresenta os resultados obtidos a partir do Matlab onde o modelo avaliado com maior precisão foi o Support Vector Machine Quadratic (SVM).

Figura 48 – Apresentação dos resultados no Power BI, Modelo *Classificação para* a variável *Classificação.*

Com o intuito de facilitar a visualização e interpretação dos resultados de forma padronizada e usando o mesmo critério de avaliação, foi produzida a tabela 24.

Acurácia por Ferramenta de AUTO ML - Classification									
Modelo	**Power BI**			**RapidMiner**			**Matlab**		
	Tempo (s)	Linhas (observações)	Acc %	Tempo (s)	Linhas (observações)	Acc %	Tempo (s)	Linhas (observações)	Acc %
Pre-fitted Soft Voting Classifier	2400	14777	86,0%	-	-	-	-		-
Naive Bayes	-	-	-	19	1000	33,40			
Generalized Linear Model	-	-	-	574	1000	69,20			
Logistic Regression	-	-	-	54	1000	39,90			
Fast Large Margin	-	-	-	981	1000	57,50			
Deep Learning	-	-	-	566	1000	57,40			
Decision Tree				17	1000	54,30			
Random Forest				265	1000	37,00			
Gradient Boosted Trees				147	1000	52,50			
Tree Fine Tree							20,8	24414	80,60
Tree Medium Tree							2,89	24414	74,50
Tree Coarse Treee							1,73	24414	62,10
Linear Discriminant							2,4	24414	74,00
Quadratic Discriminant							1,9	24414	75,50
Naive Bayes - Gaussian Naive Bayes							2,49	24414	52,00
Naive Bayes - Kernel Naive Bayes							564,52	24414	52,70
SVM - Linear SVM							190	24414	86,20
SVM - Quadratic SVM							3224	24414	86,60
SVM - Cubic SVM							5141	24414	80,40
SVM - Fine Gaussian SVM							131,89	24414	81,40
SVM - Medium Gaussian SVM							51,74	24414	85,70
SVM - Coarse Gaussian SVM							68,59	24414	84,90
KNN - Fine KNN							12,21	24414	71,90
KNN - Medium KNN							15,32	24414	74,40
KNN - Coarse KNN							20	24414	70,30
KNN - Cosine KNN							12,735	24414	74,40
KNN - Cubic KNN							426,8	24414	74,20
KNN - Weighted KNN							14	24414	75,80
Ensemble - Boosted Trees							13,4	24414	78,60
Ensemble - Bagged Trees							16,47	24414	84,90
Ensemble - Subspace Discriminant							8,63	24414	70,70
Ensemble - RusBoosted Trees							21,2	24414	72,20
Kernel - SVM Kernel							94,9	24414	86,80
Kernel - Logistic Regression Kernel							34,34	24414	60,40

Tabela 24 – Resultados geral dos Modelos a partir do Power BI, RapidMiner e Matlab Modelo Classificação para a variável Classificação.

A partir dos resultados apresentados na tabela 24, destaca-se o melhor resultado obtido no Matlab com o modelo SVM – Quadratic SVM, com resultado de precisão de 86.60% na etapa de validação com a observação dos 24414 registos. De seguida é apresentada a matriz de confusão.

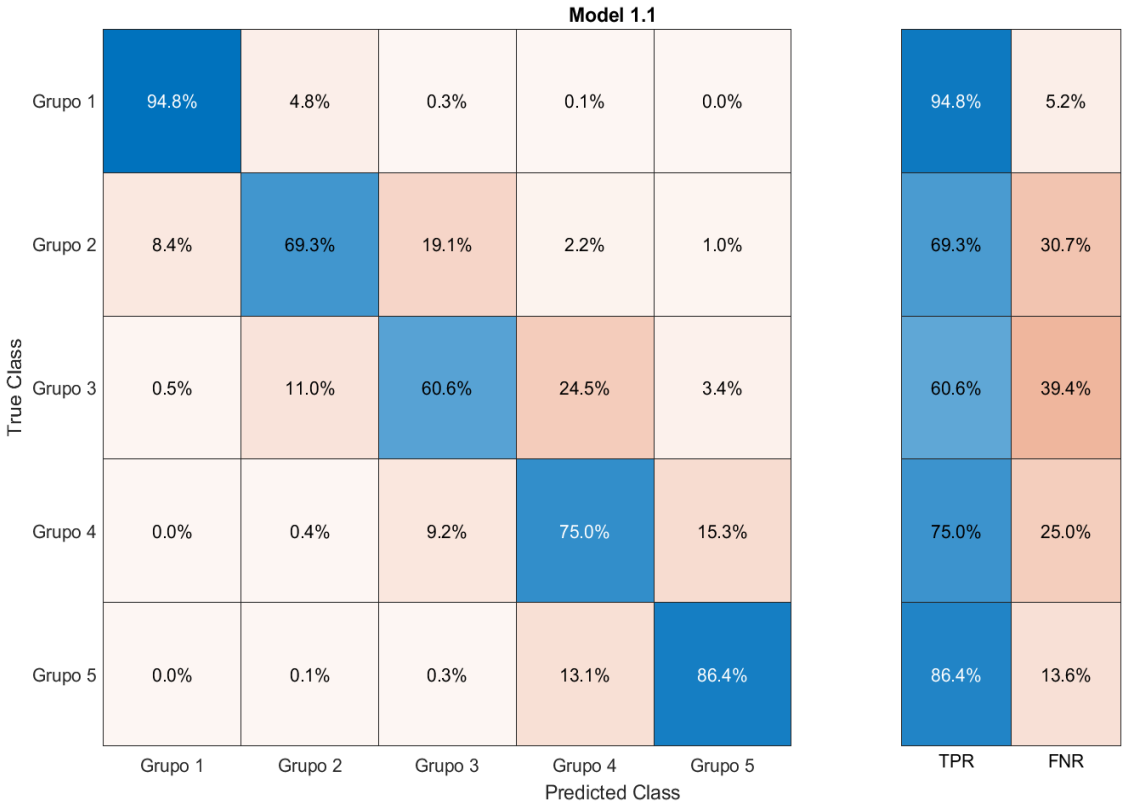


Figura 49 – Validação da Matriz Confusão – Modelo SVM Quadratic 86.6% de precisão.

Capítulo 6: Recomendações e Implementação

6.1 Recomendações e pressupostos para o MILAGE Aprender+

Conforme apresentado anteriormente no capítulo 4.3, na secção de análise dos dados, no MILAGE Aprender+ identificámos um conjunto de erros e respetivas oportunidades de correção cujas recomendações são apresentadas de seguida. Estas recomendações servem, principalmente, para viabilizar o desenvolvimento de modelos AM e potencializar trabalhos futuros no campo de análise de dados e AM. Adicionalmente são indicadas recomendações para que viabilizem a implementação de modelos AM para o MILAGE Aprender+.

I. É recomendado que o MILAGE Aprender+ propague o estado do atributo Active a todos os níveis de conteúdo. Ou seja, quando um Capítulo, Tema ou Ficha forem desativados, os demais assuntos associados e subsequentes, têm de ser desativados, caracterizando-os com o atributo Active = False. Evita-se assim que um conteúdo já desativado seja apresentado indevidamente ao utilizador.

II. Recomenda-se que o processo de avaliação, quando executado pelo Professor, seja devidamente identificado no registo. Ou seja, o utilizador professor seja identificado e a ação realizada seja devidamente registada. Assim todos os registos anteriores, de outros utilizadores, não serão apagados ou substituídos, garantindo assim a integridade da informação no sistema.

III. O campo de entrada de pontos deve ser definido para não aceitar valores negativos.

IV. O campo de entrada de pontos deve ser definido para não aceitar valores maiores que a cotação da questão.

V. Recomenda-se que numa Alínea já com registo de respostas não seja possível alterar a respetiva cotação, mitigando a possibilidade de existir divergência de pontuação anterior versus uma nova cotação para uma Alínea já respondida.

VI. Deve ser revista a parametrização para garantir que uma questão tem, no máximo, três avaliações. Sendo a auto-avaliação do próprio aluno, do par e do professor. Estas três avaliações devem ser possíveis de identificar separadamente por data, hora, utilizador e pontuação. Na hipótese de edição para correção, deverá existir a possibilidade de definir o estado (true/false) daquela avaliação, considerando sempre a última data de registo da avaliação. O histórico sempre deve ser preservado para que seja possível realizar a auditoria das ações modificadas.

VII. Recomenda-se que o MILAGE Aprender+ separe os seus ambientes em Desenvolvimento, Qualidade e Produção. Evitando assim que dados de teste e ou de desenvolvimento se misturem com os dados do ambiente de produção.

VIII. Recomenda-se garantir que toda a Alínea, inserida pelo aluno e ou com uma avaliação de par, incorra a necessidade de avaliação de um professor, estabelecendo assim um critério de fiabilidade para os pontos obtidos nos exercícios.

IX. Deve-se garantir que cada aluno responde apenas uma vez a cada Alínea. Caso seja necessário, por uma necessidade de correção do exercício, deverá ser evidenciado no sistema o motivo e a evidência da anulação dos pontos atribuídos anteriormente. O mesmo deverá acontecer na avaliação do par.

X. O campo de registo *"Time_stamp"* da realização da Alínea deverá sempre ser preenchido com a indicação de data e hora, mesmo que o aluno recuse fazer a auto-avaliação. Desta forma garantiremos os registos histórico de data e hora da realização das Alíneas.

XI. O "*time_stamp*" deverá ser padronizado com a entrada de dd/mm/aaaa hh:mm:ss.

XII. Todos os registos com conteúdos criados pelos utilizadores do tipo professor deverão conter a indicação da data de criação e ou edição.

XIII. Recomenda-se para todo o registo de resposta seja apresentado o tipo de utilizador Aluno, par ou Professor para que seja possível tipificar o tipo de autor da ação de resposta.

XIV. Toda a questão deverá ser obrigatoriamente identificada como uma questão de escolha múltipla ou como questão com resposta do tipo entrada livre de texto. As questões precisam ser devidamente categorizadas para que seja possível realizar análises por tipos de exercícios.

XV. Toda a Alínea respondida pelo aluno deverá ter obrigatoriamente uma avaliação do par e de um professor. Independentemente se a turma é formada por um grupo de total ímpar ou par. Assim nenhum aluno será penalizado pela falta da avaliação do par.

XVI. Toda questão deverá obrigatoriamente ter a indicação se contém ou não um tutorial.

XVII. Recomenda-se ainda a adoção de critério de pesos ao invés de somente contabilizar-se pontos. Conforme exemplificado na tabela abaixo, a partir dos pesos indicados pelos respetivos professores autores de cada questão, a pontuação de cada utilizador de avaliação será quantificada a partir do cálculo de pesos resultando no aproveitamento da questão e pontos finais balanceados.

Exemplo Alínea_1 *(Input):*	18	0	20
Avaliação	*Avaliação_autoavaliação*	*Avaliação_Par*	*Avaliação_Professor*
Peso da avaliação do avaliador	*17*	*3*	*80*
Nível de Dificuldade (Cotação)	20		
Total de Pontos obtidos	19,06		
Aproveitamento obtido	95		

Tabela 25 – Exemplo de recomendação de pontos e pesos para avaliação

XVIII. Recomenda-se, para efeito de qualidade dos dados, a realização de um questionário de avaliação geral do Tema e não somente avaliação nas alineas e fichas.

XIX. Recomenda-se que seja registado o tempo de realização da atividade. Desde o início da atividade até a sua conclusão. Em casos de abandonos da sessão, também permitir identificar o abandono e o tempo total decorrido.

XX. Recomenda-se a inclusão do extrato de pontos obtidos pelo aluno. O aluno deverá ser capaz de observar o seu progresso de pontos, consultando os pontos obtidos após login na plataforma, por avaliação realizada, por autoavaliação, avaliação do par e respetivo professor.

XXI. Recomendo que seja corrigido o texto (copy) onde indica que o aluno recebe 10 pontos por cada resposta dada. Verificamos sistematicamente que o aluno somente recebe pontos se a sua avaliação por superior a Zero (0).

XXII. Com o objetivo de angariar mais dados e que estes agreguem mais informação para contribuir na construção ou evolução de modelos AM, recomendo que o MILAGE Aprender+ capture mais informação durante a realização das atividades, tais como:

- Ao concluir um tema verificar a aprendizagem geral do tema registando cada resultado obtido no tema. Este recurso é adotado pela ferramenta Duolingo.
- Solicitar ao aluno qualificar a dificuldade do tema na sua conclusão e antes de ter a correção do seu professor.
- Antes de iniciar o MILAGE Aprender+ o aluno deverá realizar um teste de verificação geral de aprendizagem. Desta forma será possível observar o nível de conhecimentos do aluno em seu ponto de início.

- Para verificar o nível de conhecimento do aluno, recomenda-se obter mais informação sobre o perfil do aluno, relacionado ao seu currículo e histórico escolar.
- Mesmo as atividades por avaliar, pelos pares ou professores, devem aparecer na tabela de respostas realizadas e devidamente identificado o seu estado (*status*). Estes dados históricos poderão ser utilizados no desenvolvimento do modelo.

XXIII. Uma vez implementados os modelos recomendados, será necessário um conjunto de instruções para que os conteúdos propostos ao aluno tenham em consideração:

- Não recomendar conteúdo que o aluno já tenha realizado.
- A recomendação de conteúdo deverá estar relacionada com o Tema trabalhado pelo aluno, permitindo adequação ao conteúdo que o aluno está a aprender.
- A apresentação do conteúdo deverá incluir a opção de sugerir conteúdo relacionado com a escola do aluno. Desta forma garantimos que o conteúdo também será adequado ao grupo a que o aluno pertence.
- Ao detetar que o aluno apresenta uma provisão inferior de pontos ou aproveitamento, o MILAGE Aprender+ deverá recomendar exercícios que outros alunos da mesma escola tenham realizado e com desempenho superior.

6.2 Implementação

A partir das recomendações da secção 6.1, nesta secção são especificadas as recomendações para a implementação dos modelos AM destacados na análise de resultados.

A partir da compreensão dos objetivos do projeto, o estudo realizado sugere a implementação de dois modelos de Aprendizagem Máquina, off-line, para a recomendação de conteúdos, um a partir da previsão de pontos e outro a partir da previsão de desempenho do aluno.

Embora ambos os modelos sejam orientados para a previsão de resultados, um de classificação qualitativa e outro com previsão de pontos, ambos têm como objetivo apresentar

ao aluno exercícios adequados ao âmbito pretendido. O aluno será orientado a realizar conteúdos que favoreçam pontuar mais ou melhorar o seu desempenho no contexto da "aula".

Figura 50 – Exemplo de ecrã para apresentação da recomendação de conteúdo

6.2.1 Recomendação a partir da previsão de pontos

Na figura 51, a opção "Acelere sua pontuação no jogo" permitirá ao MILAGE Aprender+ recomendar Temas (Sub-Capítulos) a partir do atual desempenho do utilizador para conseguir obter mais pontos. A recomendação terá em consideração os Temas que ainda não foram realizados pelo aluno, assim como os Temas propostos pela escola do aluno. O objetivo é apresentar exercícios em que seja previsível o aluno ter sucesso obtendo mais pontos. A estimativa será de acordo com a função associada ao modelo AM, baseada nas variáveis apresentadas anteriormente na tabela 18, que prevê os pontos do aluno tendo em consideração os pontos obtidos até ao momento. Ou seja, de acordo com o resultado atual do aluno, a função de previsão de pontos do modelo AM irá recomendar Temas, com base na previsão de pontos para o aluno.

O modelo de previsão de pontos, com RMSE de 20,98, será capaz de recomendar exercícios ao aluno, observando o total de pontos do tema e a probabilidade de acerto (pontuação estimada do aluno) para determinado Tema recomendado. Fora do ambiente de aula, a adoção de AM na aplicação MILAGE Aprender+ poderá motivar o aluno a trabalhar para obter uma pontuação melhor no ranking dos utilizadores.

O diagrama que se segue tem como objetivo exemplificar a aplicação do modelo no MILAGE Aprender+. A função de AM poderá ser exportada do Microsoft Power BI, ambiente de desenvolvimento do modelo. No MILAGE Aprender+, a implementação deverá ser complementada com a seleção de conteúdos ainda não realizados relacionados com o Capítulo e Escola.

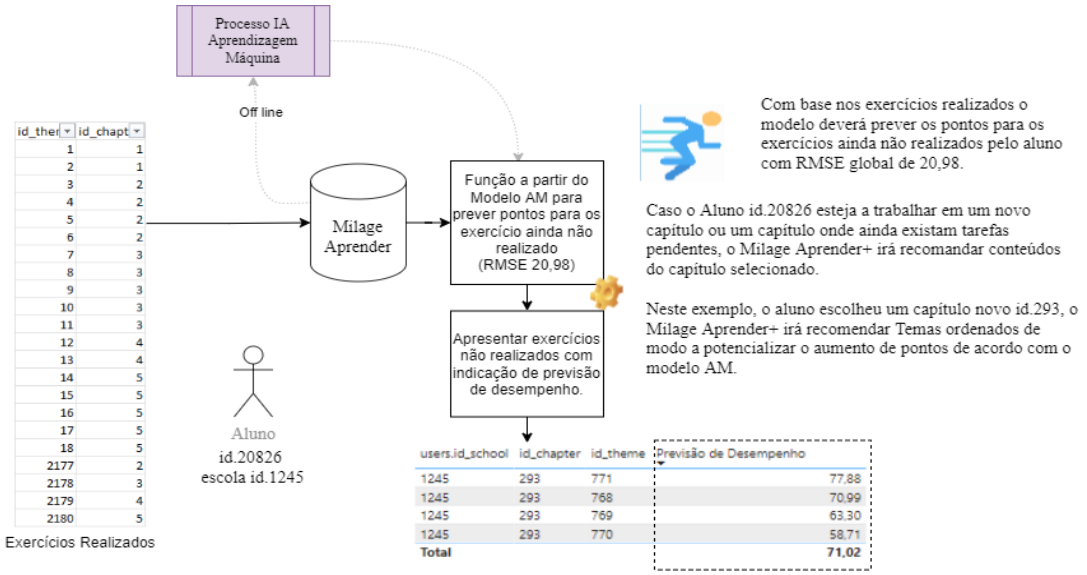


Figura 51 – Exemplo da aplicação do modelo de regressão

O pseudocódigo para implementar a função de gamificação na aplicação MILAGE Aprender+ será:

Início

Aluno Escolhe Capítulo

Para *id = 1:n, i++*

Se

Tema(id) estado = ativo ***e***

Tema(id) condição = pendente para o aluno ***e***

Tema(id) escola = realizados na escola do aluno

Então

Pontos(id)= Predict Aproveitamento Pontos

(QtdAlineasnoThema, CotacaoTotalThema, GlobalMediaAprovPontosTema, AlunoQtdAlineasRealizadasTema, AlunoPercentualRealizacaodoTema, AlunoMediaAprovPontosGeral, AlunoDiferencaAprovMediaGeral, AlunoFrequenciaUso, AlunoQtdAvaliacoesRealizadas, EscolaAlunoPontuacaoMedia, EscolaAlunoAproveitamentoMedia, EscolaAlunoPontuacaoMinima, EscolaAlunoPontucaoMaxima)

Fim Se

fim Para

Ordenar por Previsão Pontos

Apresentar temas relacionados

Fim

6.2.2 Recomendação a partir da classificação do grupo de desempenho

Na figura 51, a opção "Melhore o seu aproveitamento" permitirá ao MILAGE Aprender+ recomendar Temas (Sub-Capítulos) a partir do atual desempenho do aluno com base no histórico do próprio. Este modelo também terá em consideração os exercícios que o aluno ainda não realizou e escola relacionada. A função do modelo AM, a partir dos dados de desempenho atual do aluno, conforme atributos indicados na tabela 18, realizará a previsão de classificação do aluno para cada Tema.

O MILAGE Aprender+ irá apresentar o conteúdo dos Temas não realizados e com a respetiva previsão de classificação superior a atual do aluno.

A classificação dos alunos foi organizada em 5 grupos de desempenho, com precisão do modelo de classificação de 86,60.

Caso a recomendação de classificação seja para o mesmo e atual grupo de classificação do aluno, o MILAGE Aprender+ poderá recomendar Temas de alunos classificados no grupo superior. Neste passo de recomendação não terá relação com modelos AM, trata-se apenas de uma maneira de evitar a recomendação desnecessária para conteúdo do mesmo grupo de classificação. O objetivo é apresentar ao aluno uma referência de conteúdos que o orientem na resolução de exercícios, melhorando o seu desempenho e possibilitando a classificação num grupo superior ao previsto, motivando e priorizando ao aluno a realizar atividades que potencializem o seu desempenho.

A função do modelo AM poderá ser exportada do Matlab, ambiente de desenvolvimento do modelo. No MILAGE Aprender+, a implementação deverá ser complementada com a seleção de conteúdos ainda não realizados e filtros relacionados com o Capítulo e a Escola.

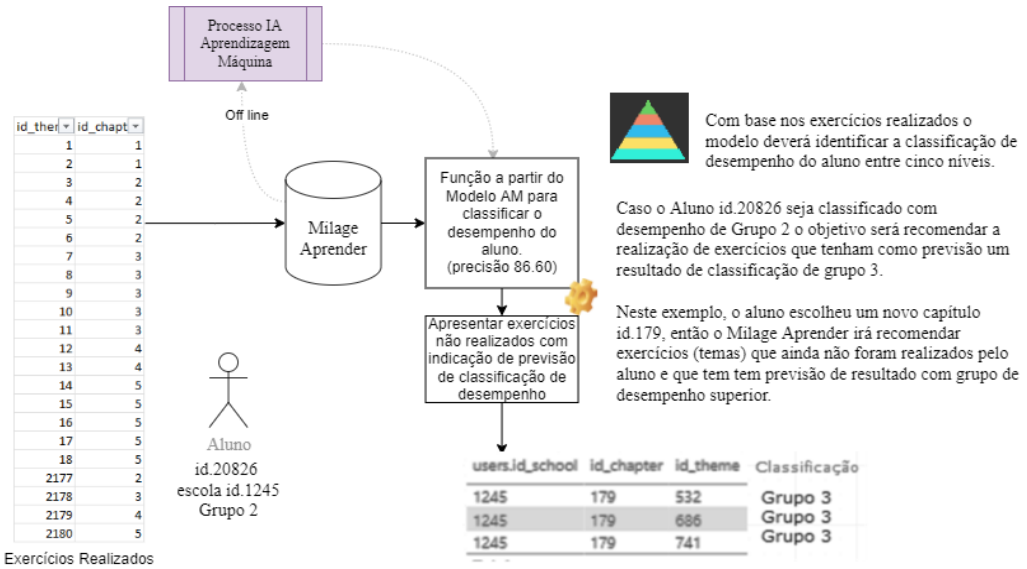


Figura 52 – Exemplo de aplicação do modelo de classificação.

O pseudocódigo para implementar a função para melhorar o desempenho do aluno na aplicação MILAGE Aprender+ será:

```
Início
  Aluno Escolhe Capítulo
        Para id = 1:n, i++
          Se
            Tema(id) estado = ativo e
            Tema(id) condição = pendente para o aluno e
            Tema(id) escola = realizados na escola do aluno
                Então
                  Desempenho(id) = Predict Classificação
                        (QtdAlineasnoThema, CotacaoTotalThema, GlobalMediaAprovPontosTema,
                        AlunoQtdAlineasRealizadasTema, AlunoPercentualRealizacaodoTema,
                        AlunoMediaAprovPontosGeral, AlunoDiferencaAprovMediaGeral, AlunoFrequenciaUso,
                        AlunoQtdAvaliacoesRealizadas, EscolaAlunoPontuacaoMedia,
                        EscolaAlunoAproveitamentoMedia, EscolaAlunoPontuacaoMinima,
                        EscolaAlunoPontucaoMaxima)

                  Se Total Desempenho = Atual Desempenho do Aluno
                        Então
                              Apresentar dados de referência de outro aluno com grupo de classificação acima.
                        Senão
                              Ordenar por desempenho
                              Apresentar temas relacionados com Desempenho acima da atual.

                  fim Se
          fim Se
        fim Para
fim
```

Na figura 53 apresenta-se um diagrama com o modelo de implementação para as duas funcionalidades sugeridas no MILAGE Aprender+ com recurso a técnicas de AM.

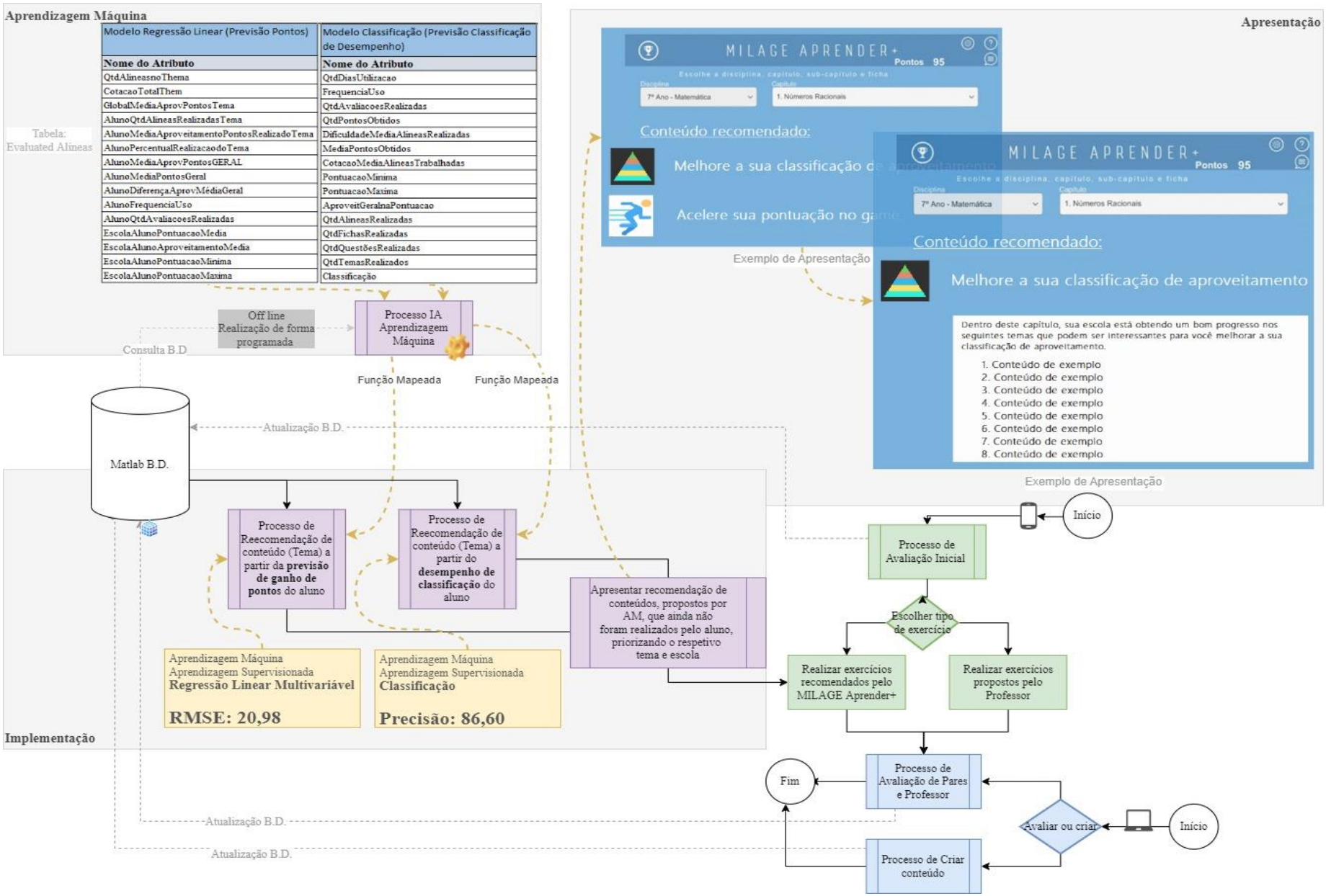


Figura 53 – Modelo para Implementação

Capítulo 7: Conclusão e Trabalho futuro

7.1 Conclusão

Neste capítulo são apresentadas as conclusões do trabalho realizado e de acordo com os objetivos propostos na secção 1.6 desta dissertação.

Objetivo i. Investigar o MILAGE Aprender+ a fim de identificar a atual estrutura de dados e modelo de funcionamento direcionando para a aplicação de modelos de AM para oferecer ao aluno e/ou professor recursos que possibilitem um percurso de aprendizagem personalizada a partir da recomendação de conteúdo.

O desenvolvimento deste projeto aprofundou amplamente a análise funcional e toda a estrutura de dados da aplicação Milage Aprender+. Adotando a metodologia CRISP-DM para guiar e suportar a realização do projeto, ao longo do estudo foram identificados alguns pontos de atenção e constrangimentos em relação à qualidade dos dados conforme evidenciado no capítulo 4.3. Algumas anomalias nos dados e falhas no funcionamento foram detetadas e como consequência limitaram alcançar resultados de maior qualidade (evidenciados na análise de resultados). Observou-se também que o atual modelo de dados e funcionalidades podem evoluir e ser otimizados, conforme é apresentado na secção 5.2. As recomendações indicadas permitirão e viabilizarão a evolução do MILAGE Aprender+ com integração de modelos AM com menor taxa de erro. Entretanto, embora com os obstáculos identificados no capítulo 4.3, concluiu-se que é possível utilizar modelos de AM para propor ao aluno atividades que o motivem no processo de aprendizagem seguindo uma abordagem de gamificação e no contexto da aula.

Com a adoção dos Modelos AM apresentados neste trabalho, o MILAGE Aprender+ poderá evoluir com uma nova característica de recomendação de Temas. Os alunos poderão utilizar os Temas recomendados automaticamente pelo MILAGE Aprender+, aumentando assim a sua probabilidade de sucesso na resolução de exercícios propostos.

Objetivo ii. Investigar se o MILAGE Aprender+ atende e/ou preenche os requisitos para a implementação de recursos de AM.

A metodologia CRISP-DM não apresenta uma ficha de pré-requisitos ou relação de pré-critérios para avaliar se um problema está apto ou não aos requisitos para implementação de um modelo de AM. Não se observou um mapa que pré-avaliasse se um problema é ou não, e ou se pode ou não ser trabalhado para a criação de um modelo AM. Obviamente, como a adoção

de um modelo de AM é totalmente dependente dos dados, há um amplo trabalho nas etapas de Conhecimento Negócio e Domínio e Problema, Análise e preparação dos dados. Neste sentido, este trabalho permitiu concluir que o problema proposto inicialmente neste estudo é totalmente pertinente e com possibilidade de adotar AM para resolução do problema já exposto no capítulo 1.

O estudo ainda constatou que o processo de desenvolvimento dum modelo de AM é contínuo. Ou seja, os resultados inicialmente obtidos podem retroalimentar o processo ciclicamente, permitindo assim que o problema e a solução sejam continuamente revistos, avaliados e melhorados.

O resultado apresentado neste estudo permitirá, além de agir sobre as recomendações geradas, servir como um novo ponto de partida para futuras melhorias na aplicação de modelos AM.

Objetivo iii. Estudar a viabilidade de utilizar um modelo de AM. Para o ensino adaptativo, pretende-se investigar a viabilidade de um modelo, algoritmo, com a adoção de AM, que possa, a partir dos dados já existentes no MILAGE Aprender+, estimar o aproveitamento dos alunos e assim, a partir de grupos de desempenho e aproveitamento gerar informação para tomar decisões, por exemplo a recomendação de conteúdos específicos.

Este projeto apresentou como resultado o modelo de AM *Light GBM Regressor* e SVM - Quadratic SVM que viabilizaram no MILAGE Aprender+ a possibilidade de recomendação de conteúdo para os alunos. Com o modelo de regressão elaborado a partir da ferramenta MS-Power BI e o modelo de classificação no Matlab, com o recurso de AUTO-ML, usando a amostra de dados históricos dos alunos do 9º ano do MILAGE Aprender+, os resultados evidenciaram uma precisão de RMSE 20.98 para o modelo de regressão para previsão de pontos e uma precisão de 86.6% para o modelo de classificação para desempenho. Como mais-valia, estes modelos de recomendação, desenvolvidos de forma off-line, proporcionarão aos alunos a opção de realizar atividades que permitirão aumentar os seus pontos e melhorar a sua posição no ranking dentro da ferramenta e/ou também melhorar o seu aproveitamento no contexto da aula.

Os professores também poderão ser beneficiados com a adoção dos modelos pois poderão ter uma visão antecipada a partir da previsão de pontuação e de classificação de um aluno, oferecendo assim a possibilidade de ações preventivas junto ao aluno e não somente no fim do percurso de aprendizagem do aluno.

Objetivo iv. Identificar no MILAGE Aprender+, informação atual e conjunto de dados, que possuam correlação forte, para que sejam aplicados nos modelos de AM.

Conforme apresentado na subsecção 5.1.3, no estudo concluiu-se que o conjunto de dados possui erros e com pouca informação relacionada ao detalhamento e categorização dos exercícios. O resultando final para os dois modelos AM foram apresentados com RMSE de 20.98 para a previsão de pontos e para a previsão de classificação com precisão de 86.60%.

Este estudo permitiu constatar que atualmente o MILAGE Aprender+ concentra os seus dados especificamente no registo dos pontos obtidos nos exercícios realizados pelos alunos e a respetiva data da realização da atividade. Estes dados, por sua vez, não são suficientes para recomendar conteúdos ao aluno. O projeto, na etapa de engenharia de recursos, apresentou um conjunto de atributos construídos através de combinações para potencializar e enriquecer o modelo.

Objetivo v. Apresentar indicações e propostas que viabilizem a implementação de modelos de AM no MILAGE Aprender+.

A partir deste projeto foi possível identificar um conjunto de recomendações, conforme apresentado na secção 5.2, que poderão ser adotadas para que o MILAGE Aprender+ angarie mais dados e informação que possa enriquecer e assim favorecer os modelos de AM, reduzindo a taxa de erro e tornando-os mais eficientes. As recomendações, em geral, centraram-se sobretudo no atual modelo de cálculo de pontos dos alunos, na normalização e padronização da entrada de dados (pontos) do aluno, da revisão pelos pares e também do professor, além da correção de alguns erros sistémicos.

É salutar finalizar indicando o contributo deste projeto para o MILAGE Aprender+, sendo este o primeiro projeto de investigação na aplicação relacionado a AM. Além da apresentação das recomendações, este poderá ser o ponto de partida para estudos futuros sobre a aplicação de AM no MILAGE Aprender+.

Realço também o contributo deste projeto no meu desenvolvimento como aluno e profissional da área de tecnologia da informação. O projeto permitiu o desenvolvimento de novas competências e sobretudo ampliou a minha capacidade para a realização de projetos relacionados com a análise de dados e AM. Seguramente servirá como suporte e base para trabalhos futuros.

Por fim, finalizo comentando os resultados deste projeto, sobretudo sobre o potencial que a tecnologia de AM representou neste projeto, ainda que com deteção de constrangimentos

e obstáculos. Concluindo, a AM permite resolver problemas complexos como o apresentado neste projeto. Sobretudo, foca a importância que os dados têm nos modelos de AM e respetiva influência na geração de conhecimento para tomar decisões.

7.2 Trabalho futuro

A partir deste projeto identificaram-se oportunidades para avançar com trabalho futuro, dentro do MILAGE Aprender+, depois da resolução dos problemas mencionados anteriormente e de implementação das recomendações relacionadas com o modelo de pontuação, possibilitando a realização de novas experiências com modelos de AM. Por outro lado, será importante como trabalho futuro, explorar a possibilidade de aplicar modelos de AM por Reforço, muito comuns em jogos, já que podemos classificar o MILAGE Aprender+ como um sistema de gamificação.

Referências Bibliográficas

Alzubi, J., Nayyar, A. & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. Journal of Physics: Conference Series.

Azevedo P, J. M. (2020). SARS-CoV-2 e COVID-19: *Os Aspetos Virológicos de uma Pandemia*, Revista Portuguesa de Farmacoterapia.

Brownlee, J. (2016). *Machine Learning Mastery, Discover How They Work and Implement them from scratch*. http://MachineLearningMastery.com.

C3.ai. (2021). Erro de raiz quadrada média, *Glossário*, https://c3.ai/glossary/data-science/root-mean-square-error-rmse/

Casatti, D. (2020). Ensino remoto na pandemia pode transformar educação, *Jornal da USP*. https://jornal.usp.br/universidade/ensino-remoto-na-pandemia-pode-transformar-educacao/

Chandrashekar, A., Amat, F., Basilico, J., & Jebara, T. (2020). *Netflix Research*, https://research.netflix.com/business-area/personalization-and-search

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000). CRISP-DM 1.0, Step-by-step data mining guide.

Cortez, M. B. (2017). AI in Education Will Grow Exponentially by 2021. *Edtech Magazine*, https://edtechmagazine.com/k12/article/2017/07/ai-education-will-grow-exponentially-2021

Daniel Johnson. (2021). Aprendizagem por reforço: o que é, algoritmos, aplicativos, exemplo. https://www.guru99.com/reinforcement-learning-tutorial.html

Departamento de Educação dos Estados Unidos da América (2013). Expandindo abordagens de evidências para a aprendizagem em um mundo digital.

Figueiredo, M., Bidarra, J, Godejord B., Perez A. G. (2018). Breaking Barriers in learning Math Architecture of the MILAGE Learn+ App. (A. International, Ed.)

Figueiredo, M., Godejord, B., Rodrigues, J. (2016). The development of an interactive mathematics app for mobile learning.

Frank Hutter, R. C. (2015). Auto ML, Jornal Conference, https://sites.google.com/site/automlwsicml15/

Google Trends. (2020). Google Trends Search: https://trends.google.pt/trends/explore?date=today%205-y&geo=PT&q=%2Fm%2F02h32

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O. (2018) When Will AI Exceed Human Performance? When Will AI Exceed Human Performance?

Hongjing Wu, Geert-Jan Houben, Paul De Bra. (2000). A Reference Model to Support Adaptive Hypermedia Authoring. A Reference Model to Support Adaptive Hypermedia Authoring.

Hosch, W. L. (2009). Aprendizagem de Máquina, https://www.britannica.com/technology/machine-learning

Hotz, N. J., & Saltz, D. J. (2021). O que é CRISP DM? https://www.datascience-pm.com/crisp-dm-2/#crisp-dm-for-data-science

IBM Cloud Education. (2020). Aprendizagem Supervisionada https://www.ibm.com/cloud/learn/supervised-learning

Mandot, P. (17 de Agosto de 2017). What is LightGBM, How to implement it? How to fine tune the parameters? https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

MathWorks. (2021). Machine Learning, https://www.mathworks.com/discovery/machine-learning.html#machine-learning-with-matlab

McCarthy, J. (1963). Programs with common sense. Proceedings of the Symposium on the Mechanization of Thought Processes.

Microsoft. (2020). Feature engineering in machine learning. https://docs.microsoft.com/en-us/azure/architecture/data-science-process/create-features

Microsoft. (2021). IA responsável e confiável. https://docs.microsoft.com/en-us/azure/cloud-adoption-framework/innovate/best-practices/trusted-ai

Microsoft. (2021). O que é a computação na nuvem ? https://azure.microsoft.com/pt-pt/overview/what-is-cloud-computing/#cloud-deployment-types

Microsoft. (15 de 08 de 2021). O que é a computação na nuvem ? https://azure.microsoft.com/pt-pt/overview/what-is-cloud-computing/#benefits

Mitchell, T. (1997). Machine Learning, Ed. McGraw-Hill Science/Engineering/Math.

Michael K., Cave R., Foden M., Stend M. (2016). Personalized Education: from curriculum to career with cognitive systems. Personalized Education.

Piatetsky, G. (2014). KDnuggets. https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Pimentel M., Denise Filippo; , Thiago Marcondes dos Santos (2020). Design Science Research: pesquisa científica atrelada ao design de artefatos.

Posner, Z. (2017, 01 11). Mheducation, https://www.mheducation.com/ideas/what-is-adaptive-learning.html

Puterman, M. L. (2014). Markov Decision Processes, Discrete Stochastic Dynamic Programming. Ed. Wiley.

Rapidminer. (2021). Aprendizagem máquina automatizada, https://rapidminer.com/glossary/automated-machine-learning/

Redman, T. C. (2017). Aproveitando a oportunidade em qualidade de dados, https://sloanreview.mit.edu/article/seizing-opportunity-in-data-quality/

Samuel, A. (1959). Some studies in machine learning using the game of checkers. IBM Journal of research and development. IBM Journal of research and development.

Sichman, J. S. (2021). Inteligência Artificial e sociedade: avanços e riscos. https://doi.org/10.1590/s0103-4014.2021.35101.004

Soares, Tiago (2020). A aplicação de telemóvel que torna a matemática um jogo viciante nasceu no Algarve: conheça o Milage Aprender+. *Jornal Expresso*, https://expresso.pt/sociedade/2020-06-18-A-aplicacao-de-telemovel-que-torna-a-matematica-um-jogo-viciante-nasceu-no-Algarve-conheca-o-Milage-Aprender-

Tibshirani, R., Hastie, T., & Friedman, J. (2008). The Elements of Statistical Learning. Ed. Springer.

Tocaria, M. (2020). Agência Brasil, https://agenciabrasil.ebc.com.br/educacao/noticia/2020-05/brasil-tem-48-milhoes-de-criancas-e-adolescentes-sem-internet-em-casa

UNESCO. (2020). 1.37 billion students now home as COVID-19 school closures expand, ministers scale up multimedia approaches to ensure learning continuity, https://en.unesco.org/news/137-billion-students-now-home-covid-19-school-closures-expand-ministers-scale-multimedia

US Education Sector (2018). *Artificial Intelligence Market in the US Education Sector 2018-2022*. Ed: TechNavio (Infiniti Research Ltd.).

Wong, W., Oxman, S. (2014). White Paper: Adaptive Learning Systems. Ed. Integrated Education Solutions.