PIAES Learning Analytics

Resumo

As instituições de ensino superior registam uma quantidade significativa de dados sobre os alunos, representando estes, um potencial considerável de geração de informação, conhecimento e monitorização do percurso académico dos alunos. O conceito de Learning Analytics (LA) refere-se ao uso inteligente desses dados para extrair informação relevante para a tomada de decisões, incluindo a previsão de desempenho académico. O projeto aqui apresentado consiste numa ferramenta de LA que foi chamada de "PIAES Learning Analytics" a qual auxilia na definição de estratégias e mecanismos de melhoria por parte dos órgãos de gestão (pedagógico, coordenações de curso, entre outros). Agrega dados provenientes de diversas fontes e faz o processamento dos mesmos através do uso de técnicas de ciência dos dados e aprendizagem máquina para efetuar um conjunto de análises que incluem a caracterização dos alunos, ingressos, empregabilidade, abandono, percurso académico e previsão de sucesso, disponibilizando um conjunto de painéis de controlo que permitem visualizar a informação destas análises. Neste momento, a ferramenta encontra-se em utilização e a fornecer informação para o Grupo de Trabalho do Tutorado do Politécnico de Portalegre, no que diz respeito a quatro turmas do primeiro ano (quatro cursos) abrangidas pelo apoio deste grupo.

1. Desenvolvimento

A maioria das implementações de LA centra-se na utilização apenas de dados registados nas plataformas de ensino das instituições de ensino e abordam essencialmente a questão da previsão do sucesso ou desempenho. O projeto aqui apresentado, complementa a previsão do sucesso (análise preditiva) com uma análise descritiva utilizando também dados externos o que, no seu conjunto, possibilita uma visão mais alargada da instituição e do percurso dos alunos.

1.1 Objetivo

O principal objetivo do projeto consiste na disponibilização de informação aos vários órgãos, de forma a auxiliar na tomada de decisão relacionada com a definição de estratégias e mecanismos de melhoria em termos de atratividade, retenção e desempenho dos alunos.

1.2 Soluções tecnológicas

A Figura 1 ilustra a arquitetura do "PIAES Learning Analytics" a qual pretende fornecer uma visão global das soluções tecnológicas utilizadas e que irá ser detalhada de seguida.

Fontes de dados

O conjunto de dados inclui informação no momento do ingresso do aluno, o percurso académico, dados demográficos, socioeconómicos e macroeconómicos. As fontes de dados utilizadas, são constituídas por dados internos e externos ao Politécnico de Portalegre, e incluem dados provenientes de:

- Sistema de gestão académico do Politécnico de Portalegre (DIGITALIS);
- Sistema de apoio à atividade letiva do Politécnico de Portalegre (PAE);

- Dados anuais da DGES referentes ao ingresso por via do Concurso Nacional de Acesso ao Ensino Superior (CNAES);
- Estudo da DGEEC de 2018 sobre "PERCURSOS NO ENSINO SUPERIOR - Situação após quatro anos dos alunos inscritos em licenciaturas de três anos";
- Estudo da DGEEC de junho de 2020 relativo à "Caracterização dos desempregados registados com habilitação superior";
- PORDATA - Base de Dados Portugal Contemporâneo;

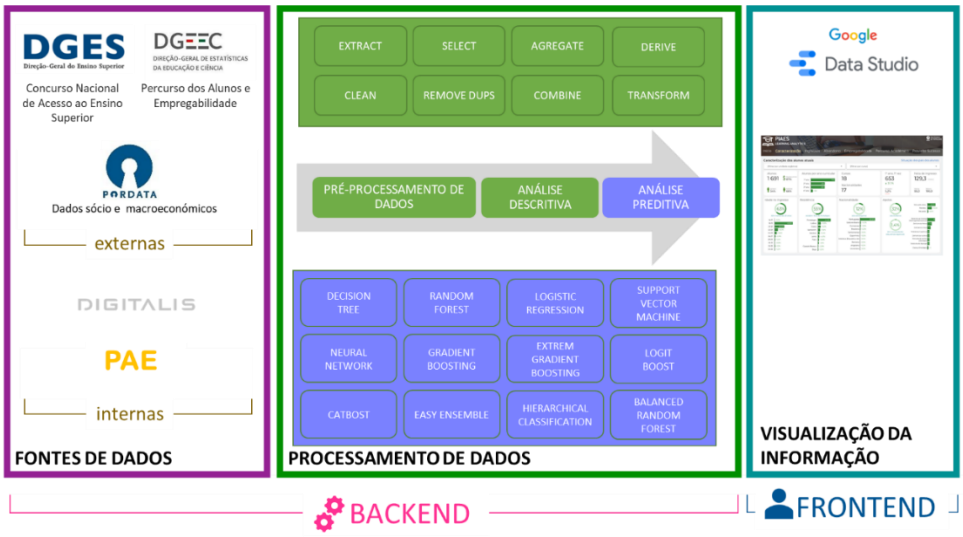


Figura 1. Arquitetura do sistema

A Tabela 1 ilustra os dados recolhidos os quais podem ser categorizados em dados demográficos, dados socioeconómicos, dados macroeconómicos, dados académicos no ingresso e dados académicos no fim de cada semestre.

Tipo de dado	Dado
Dados demográficos	Idade no ano de ingresso Género Estado civil Nacionalidade Concelho da morada oficial Distrito da morada oficial Aluno deslocado (S/N) Aluno estrangeiro (S/N)
Dados socioeconómicos	Habilitação literária do pai Habilitação literária da mãe Grupo profissional do pai Grupo profissional da mãe Aluno com propinas em dia (S/N) Aluno bolseiro (S/N) Aluno devedor (S/N) Aluno com necessidades especiais (S/N)
Dados macroeconómicos	Taxa de crescimento real do PIB Taxa de desemprego Taxa de inflação
Dados académicos no ingresso	Forma de ingresso Curso Nota de ingresso Regime de frequência (diurno/noturno) Nota da habilitação anterior Escola de realização dos exames de acesso

	Concelho dos exames de acesso Distrito dos exames de acesso Número de exames realizados no acesso Média dos exames no concurso de acesso
Dados académicos no fim de cada semestre	Média das notas semestres anteriores Número de unidades curriculares aprovadas Percentagem de unidades aprovadas Número de unidades não aprovadas Percentagem de unidades não aprovadas Número de creditações Número de avaliações efetuadas Número de aprovações

Tabela 1. Categorização dos dados recolhidos.

Os dados internos são recolhidos através de exportação para ficheiros em formato comma-separeted values (CSV) a partir do sistema de gestão académica (DIGITALIS), e dizem respeito a 15 anos letivos (deste a introdução do processo de Bolonha), a 50 cursos (entre Cursos Técnicos Superiores Profissionais, Licenciaturas e Mestrados), e contemplam 10.296 alunos e mais de 470.000 avaliações.

Todos os dados são anonimizados encontrando-se assegurado o cumprimento da Política de Privacidade e de Tratamento de Dados Pessoais do Politécnico de Portalegre, disponível na sua página de internet, em https://pae.ipportalegre.pt/policy/rgpd.

Processamento de dados

O processamento de dados permite preparar os dados para ser feita a análise descritiva e preditiva e todo o processamento foi feito em Python a correr sobre o sistema operativo Ubuntu num computador NVIDIA DGX Station com 2 CPU Intel Xeon E5-2698V4 20 core 2.2Ghz, 256 GB de memória e 4 GPU NVIDIA Tesla V100.

Análise descritiva

A análise descritiva dos dados permite criar um conjunto de análises que possibilitam efetuar:

- a caracterização dos alunos;
- uma análise dos ingressos dos alunos;
- uma análise de empregabilidade dos cursos;
- uma análise do abandono;
- uma análise do percurso académico.

Análise preditiva

O problema de previsão do sucesso foi abordado apenas para os estudantes de licenciatura, e formulado como uma tarefa de classificação de três categorias. As categorias consideradas dizem respeito à situação académica dos estudantes ao fim dos N anos curriculares do curso (3, ou 4, no caso da licenciatura em Enfermagem):

- **Diplomado** nos N anos curriculares do curso
- Ainda não diplomado mas **inscrito** no curso
- **Abandono** quer por via da anulação da matrícula, quer por via da não renovação da inscrição

A distribuição destas classes mostrou um forte desequilíbrio, pelo que foi necessário testar diferentes estratégias para a promoção do balanceamento: ao nível dos dados e ao nível dos algoritmos de aprendizagem máquina que incorporam passos de balanceamento.

No primeiro caso, foram testados métodos de sobreamostragem sintética, nomeadamente SMOTE, ADASYN, e SVM-SMOTE, ao qual se seguiu o treino de modelos com algoritmos de aprendizagem máquina que não incorporam balanceamento. Entre este, salienta-se a Regressão Logística, Máquinas de Suporte Vetorial, Redes Neuronais, Árvores de Decisão, Florestas Aleatórias, Random Boost, Extreme Gradient Boost, RusBoost, CatBoost, e Classificação Hierárquica. Testados os modelos, concluiu-se que a estratégia SVM-SMOTE seguida de treino com Florestas Aleatórias conduz a melhores resultados (valores de pontuação F1 mais elevadas).

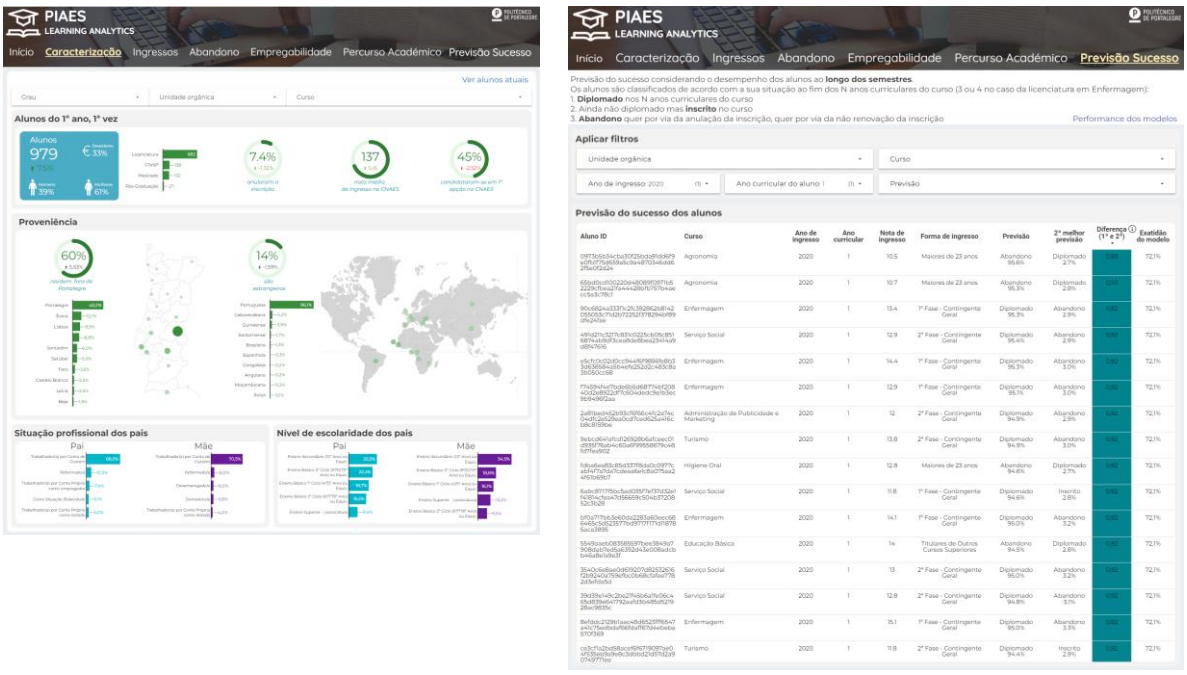
No segundo caso, os dados foram usados com algoritmos que incorporam estratégias para balanceamento, como Balanced Random Forest e Easy Ensemble Classifier, tendo o primeiro revelado uma performance superior, tanto no geral quanto para as classes individuais.

Foram treinados e testados modelos tendo como base um conjunto de 22 variáveis que representam a informação conhecida sobre o estudante no ingresso no ensino superior, e modelos com a informação académica adicional disponível no final do primeiro e do segundo semestre. Os modelos foram treinados considerando os dados referentes a 4.424 alunos correspondente a todos os que ingressaram em ciclos de estudo de 1º ciclo desde o ano letivo 2008/09.

Visualização de informação

A componente de visualização de informação foi feita recorrendo ao Google Data Studio o qual disponibiliza todo o conjunto de objetos gráficos que facilitam a criação das interfaces com o utilizador.

A Figura 2a) ilustra um dos écrans de caracterização dos alunos disponibilizados, enquanto a Figura 2b) o écran de previsão do sucesso. Mais écrans disponíveis [nesta ligação](#).



a) b)

Figura 2. Écran de (a) caracterização dos alunos (1º ano / 1ª vez) e (b) previsão do sucesso.

1.3 Atividades

O projeto, com uma duração inicial de 16 meses, acabou por se prolongar por 24 meses devido à situação pandémica que fez atrasar alguns dos desenvolvimentos. As atividades realizadas encontram-se apresentadas na Tabela 2.

Atividades
1. Caracterização do problema, dos objetivos e do impacto esperado
2. Análise e preparação de dados
3. Implementação dos modelos de análise e processamento (ciência dos dados e inteligência artificial)
4. Prototipagem do sistema de implementação dos modelos
5. Testes de funcionalidade do protótipo
6. Apresentação de resultados e divulgação
7. Procedimento de aquisição de equipamento
8. Acompanhamento e gestão do projeto

Tabela 2. Atividades do projeto.

1.4 Recursos utilizados

Na equipa do projeto fizeram parte 4 investigadores com perfil técnico na área de informática e 3 técnicos com perfil funcional e experiência na área dos serviços académicos.

Como já referido, foi utilizado um computador NVIDIA DGX Station com 2 CPU Intel Xeon E5-2698V4 20 core 2.2Ghz, 256 GB de memória e 4 GPU NVIDIA Tesla V100, para efetuar todo o processamento de dados.

2. Conclusões

A plataforma desenvolvida permite a valorização dos dados recolhidos todos os anos pelo Politécnico de Portalegre relativos aos seus estudantes, e espera-se que possa contribuir para tomadas de decisão estratégicas mais informadas. Atualmente, a plataforma foi disponibilizada a um grupo restrito de utilizadores. A componente preditiva está a ser utilizada para apoio à seleção de estudantes acompanhados pelo programa de tutorado do Instituto Politécnico de Portalegre (4 turmas do 1º ano abrangendo um total de 234 alunos).

Prevê-se que a experiência ganha no desenvolvimento da plataforma, o feedback da sua utilização e o natural aumento da informação registada anualmente no sistema permitam a sua validação no curto prazo, nomeadamente no que diz respeito à componente preditiva.

3. Resultados

- Ferramenta “PIAES Learning Analytics” atualmente em utilização e a fornecer informação para o Grupo de Trabalho do Tutorado do Politécnico de Portalegre.

- Conjunto de dados de treino dos modelos disponibilizados em *open access* na UCI e no Zenodo.

Valentim Realinho, Jorge Machado, Luís Baptista, & Mónica Martins. (2021). **Predict Students' Dropout and Academic Success** (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5777340

- Artigos científicos e apresentação em conferências com parte dos resultados alcançados:

  Martins M.V., Tolledo D., Machado J., Baptista L.M.T., Realinho V. (2021) **Early Prediction of Student's Performance in Higher Education: A Case Study**. Trends and Applications in Information Systems and Technologies. WorldCIST 2021. Advances in Intelligent Systems and Computing, vol 1365. Springer, Cham. https://doi.org/10.1007/978-3-030-72657-7_16 (SJR 2019: 0.1840 Q3)

  Realinho V., Martins M.V., Baptista L.M.T., Machado J. (2021) **Intelligent Use of Data for Monitoring the Academic Pathway in Higher Education**. International Congress of 21st Century Literacies, 15-16 july, Portalegre, Portugal

- Tolledo D. (2020) **Estudo e Avaliação de Algoritmos de Aprendizagem Máquina Aplicados ao Insucesso Escolar**. Trabalho Final de Curso em Engenharia Informática, Politécnico de Portalegre

- Pedido de patente apresentado

Financiamento

Este trabalho foi financiado pelo programa SATDAP - Capacitação da Administração Pública através do projeto POCI-05-5762-FSE-000191