

PREDIÇÃO DO NÚMERO DE ESTUDANTES NO ANO LETIVO 2021/22 NO ISLA SANTARÉM

PREDICTION OF THE NUMBER OF STUDENTS IN THE YEAR 2021/22 AT ISLA SANTARÉM

Nelson Duarte¹; Ricardo Vardasca²

¹ISLA Santarém; ²ISLA Santarém

nelson.duarte@islasantarem.pt; ricardo.vardasca@islasantarem.pt

Resumo

Introdução: Determinar o número de estudantes para o próximo ano letivo de uma instituição de ensino superior é sempre um fator crítico de gestão. Este estudo tem como objetivo prever o número de alunos no ano letivo de 2021/2022 na instituição de Ensino Superior ISLA Santarém.

Método: Para a realização deste estudo procurou-se desenvolver um modelo de predição em Data Mining para a obtenção de resultados usando a linguagem de programação Python. inicialmente foi criada uma base de dados em MS SQL Server, implementou-se no MS Visual Studio o Integration Services Project para efetuar o ETL dos dados extraídos de um ficheiro MS Excel e procedeu-se a criação de um cubo MOLAP onde se criaram as medidas a serem exportadas para o MS Power BI para a criação dos Dashboards. Para a predição exportaram-se os dados para um ficheiro csv do cubo OLAP para serem processados numa script Python.

Resultados: Através da predição conseguiu-se verificar que o número de alunos para o próximo ano letivo, no ISLA Santarém, vai aumentar.

Discussão: Tratou-se de um dos estudos pioneiros na tentativa de prever o número de estudantes do ano seguinte com base no histórico de anos anteriores, sendo por isso difícil realizar comparações com outros estudos.

Conclusão: Os resultados deste estudo podem ajudar no processo de gestão de recursos da instituição no que respeita a preparação do ano letivo seguinte.

Palavras-chave: ISLA santarém; número de alunos; predição.

Abstract

Introduction: Determining the number of students for the next academic year at a higher education institution is always a critical management factor. This study aims to predict the number of students in the academic year 2021/2022 in the Higher Education Institution ISLA Santarém.

Methods: To carry out this study it was sought to develop a prediction model in Data Mining to obtain results in Python programming language. initially a database was created in MS SQL Server, implemented in MS Visual Studio the Integration Services Project to perform the ETL of the data extracted from an MS Excel file and proceeded to create a MOLAP cube where they created the measures to be exported to MS Power BI for the creation of Dashboards. For the prediction, the data was exported from the OLAP cube to a csv file and processed by a Python script.

Results: Through the prediction, it was possible to verify that the number of students for the next academic year, in ISLA Santarém, will increase.

Discussion: This was one of the pioneering studies to predict the number of students for the following year based on the history of previous years, so it is difficult to make comparisons with other studies.

Conclusion: The results of this study can help in the institution's resource management process with regard to the preparation of the following academic year.

Keywords: number of students; ISLA Santarém.

INTRODUÇÃO

Prever o número de alunos para o próximo ano letivo em qualquer instituição de ensino é uma meta a alcançar por qualquer instituição, justificada na preparação atempada de toda a estratégia académica que lhe está inerente. Essa informação poderá ter um impacto bastante significativo para professores, funcionários e administração de uma instituição permitindo-lhes criar estratégias diminuindo a pressão imposta num início de ano letivo.

Para efetuar este estudo foi utilizado um **dataset** em Excel fornecido pela instituição ISLA Santarém com informações sobre o número de alunos e algumas das suas características desde 2013 a 2020, utilizou-se o MS SQL server para criar a base de dados e o MS Visual Studio para fazer a conversão de dados do MS Excel para o sistema gestor de base de dados, assim como a criação do cubo com a ferramenta Analysis Services Dimensional and Data Mining Project. Os dados gerados foram importados para o MS Power BI, onde foram criados os dashboards com a informação relevante para este estudo.

Uma solução multidimensional do Analysis Services usa estruturas de cubo para analisar dados de negócios em várias dimensões. O modo multidimensional é o modo de servidor padrão do Analysis Services. Inclui um mecanismo de consulta e cálculo para dados OLAP, com modos de armazenamento MOLAP, ROLAP e HOLAP para equilibrar o desempenho com requisitos de dados escalonáveis. O mecanismo OLAP do Analysis Services é um servidor OLAP líder do setor que funciona bem com uma ampla variedade de ferramentas de BI. A maioria das implantações do Analysis Services são instaladas como servidores OLAP clássicos (Paul et al., 2000).

O Data Mining é uma metodologia de análise de dados utilizada para identificar padrões ocultos num grande conjunto de dados.

Tem sido utilizado com sucesso em diferentes áreas, incluindo o ambiente educacional. A mineração de dados educacionais é uma área de pesquisa interessante que extrai padrões desconhecidos de bases de dados educacionais para melhor compreensão, melhor desempenho educacional e avaliação do processo de aprendizagem do aluno. Avaliar o aumento do número de alunos é uma questão complexa, que não pode ser restrita a classificação. Razões de aumento ou diminuição de alunos pertencem aos principais interesses dos professores, porque eles podem planear e personalizar a sua forma de ensino, com base no feedback. A mineração de dados é uma das abordagens, que pode fornecer uma assistência eficaz revelando relações complexas (Hung et al., 2020).

Em estatística, a regressão linear é uma abordagem linear para modelar a relação entre uma resposta escalar e uma ou mais variáveis explicativas (também conhecidas como variáveis dependentes e independentes). O caso de uma variável explicativa é chamado de regressão linear simples para mais de um, o processo é chamado de regressão linear múltipla. Este termo é diferente da regressão linear multivariada, onde várias variáveis dependentes correlacionadas são previstas, em vez de uma única variável escalar (Liu et al., 1997).

Na regressão linear, os relacionamentos são modelados usando funções de preditor linear cujos parâmetros de modelo desconhecidos são estimados a partir dos dados. Esses modelos são chamados de modelos lineares (Aldrich, 1999). A média condicional da resposta dada os valores das variáveis explicativas (ou preditores) é assumida como uma função afim desses valores, menos comumente, a mediana condicional ou algum outro quantil é utilizado. Como todas as formas de análise de regressão, a regressão linear concentra-se na distribuição de probabilidade condicional da resposta dados os valores dos preditores, ao invés da

distribuição de probabilidade conjunta de todas essas variáveis, que é o domínio da análise multivariada.

A regressão linear foi o primeiro tipo de análise de regressão a ser estudado rigorosamente e amplamente usado em aplicações práticas. Isso ocorre porque os modelos que dependem linearmente dos seus parâmetros desconhecidos são mais fáceis de ajustar do que os modelos não linearmente relacionados aos seus parâmetros e porque as propriedades estatísticas dos estimadores resultantes são mais fáceis de determinar (Uyanık & Güler, 2013).

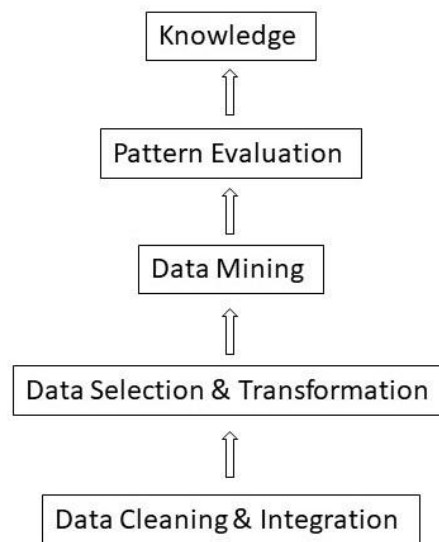


Figura 1 – Etapas do processo de Data Mining

Fonte – elaboração própria

ESTADO DA ARTE

Existem alguns estudos recorrendo a técnicas de Data Mining para obter resultados sobre estudantes numa instituição de ensino, mas focam-se sobretudo na taxa de abandono dos alunos nas instituições ou na taxa de sucesso dos mesmos (Kondo et al., 2017). Relativamente a estudos realizados sobre a predição do número de alunos para o próximo ano letivo foram encontrados poucos estudos académicos que estejam disponíveis para consulta pública.

Existe um estudo sobre a previsão do número de alunos em Portugal por regiões, desde 2013 a 2019 que inclui alunos do 1º e 2º ciclo de ensino básico (Canto Castro et al., 2013).

Para a instituição ISLA Santarém em particular, foi realizado um estudo sobre a previsão de abandono de alunos numa instituição de ensino superior utilizando os algoritmos de classificação Two class logistic regression, Two class boosted decision, Two class neural network e Two class support vector (Sobreiro & Martinho, n.d.).

METODOLOGIA

A metodologia implementada neste estudo baseia-se em quatro fases onde é definida em cada uma delas os processos e técnicas utilizados para a realização deste estudo.

Fase 1 – Definição do objetivo do estudo.

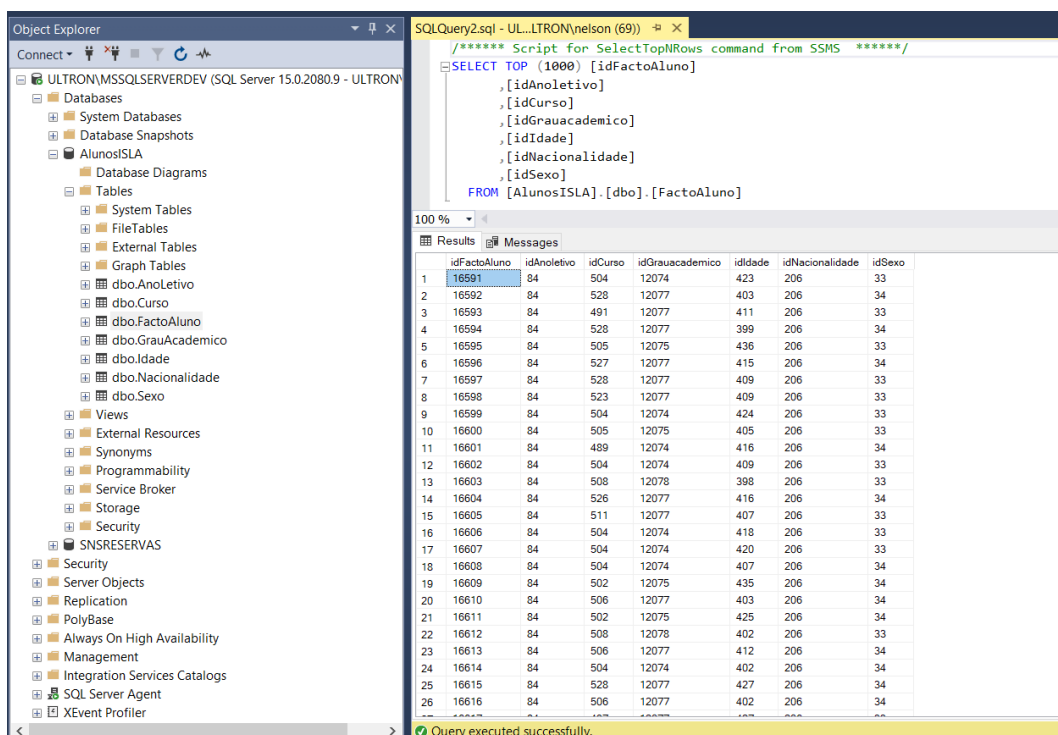
Foi realizada uma prospeção da necessidade de prever o número de alunos na instituição ISLA Santarém com o objetivo de se saber o número de alunos para o próximo ano letivo da instituição.

Fase 2 – Reunião de informação para a realização do estudo.

A instituição disponibilizou um dataset em excel com informação sobre o ano letivo, os cursos, o grau académico, o número de alunos, a naturalidade, a nacionalidade, a idade e o sexo.

Fase 3 – Transformação e carregamento dos dados através do processo de ETL

Os dados foram carregados numa base de dados criada no MS SQL Server através do Integration Services Provider do MS Visual Studio 2019 onde foram convertidos e inseridos para de seguida ser criado o cubo no Analysis Services Dimensional and Data Mining Project, onde foram criadas as medidas e preparados os dados para serem importados para o MS PowerBI da Microsoft, onde foram criados os dashboards com os gráficos referentes os dados processados.



The screenshot displays the Microsoft SQL Server Enterprise Manager interface. On the left, the 'Object Explorer' pane shows the database structure for 'ULTRON\MSSQLSERVERDEV (SQL Server 15.0.2080.9 - ULTRON)'. The 'Databases' folder is expanded, showing 'AlunosISLA' and its tables: 'dbo.AnoLetivo', 'dbo.Curso', 'dbo.FactoAluno', 'dbo.GrauAcademico', 'dbo.Idade', 'dbo.Nacionalidade', and 'dbo.Sexo'. The 'dbo.FactoAluno' table is selected. The right pane shows the 'SQLQuery2.sql' script, which is a 'SELECT TOP (1000) [idFactoAluno], [idAnoLetivo], [idCurso], [idGrauAcademico], [idIdade], [idNacionalidade], [idSexo] FROM [AlunosISLA].[dbo].[FactoAluno]' query. Below the script, the 'Results' pane displays the first 26 rows of the query output. The status bar at the bottom indicates 'Query executed successfully.'

	idFactoAluno	idAnoLetivo	idCurso	idGrauAcademico	idIdade	idNacionalidade	idSexo
1	16591	84	504	12074	423	206	33
2	16592	84	528	12077	403	206	34
3	16593	84	491	12077	411	206	33
4	16594	84	528	12077	399	206	34
5	16595	84	505	12075	436	206	33
6	16596	84	527	12077	415	206	34
7	16597	84	528	12077	409	206	33
8	16598	84	523	12077	409	206	33
9	16599	84	504	12074	424	206	33
10	16600	84	505	12075	405	206	33
11	16601	84	489	12074	416	206	34
12	16602	84	504	12074	409	206	33
13	16603	84	508	12078	398	206	33
14	16604	84	526	12077	416	206	34
15	16605	84	511	12077	407	206	33
16	16606	84	504	12074	418	206	33
17	16607	84	504	12074	420	206	33
18	16608	84	504	12074	407	206	34
19	16609	84	502	12075	435	206	34
20	16610	84	506	12077	403	206	34
21	16611	84	502	12075	425	206	34
22	16612	84	508	12078	402	206	33
23	16613	84	506	12077	412	206	34
24	16614	84	504	12074	402	206	34
25	16615	84	528	12077	427	206	34
26	16616	84	506	12077	402	206	34

Figura 2 - Base de Dados AlunosISLA no MS SQL Server

Fonte - Elaboração própria

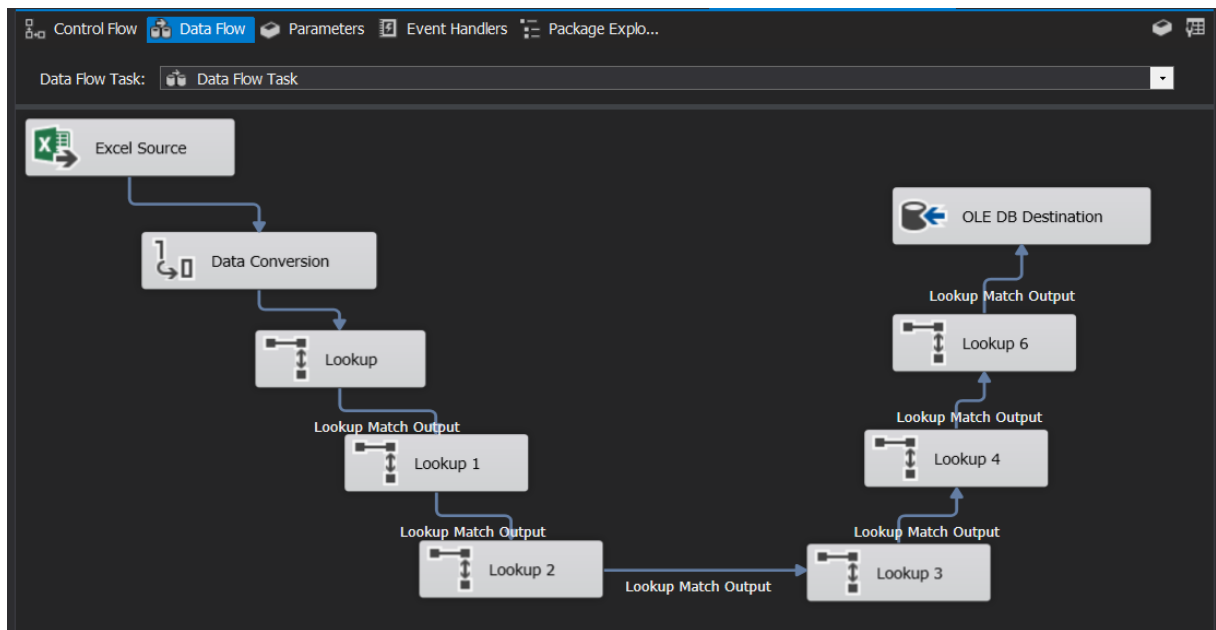


Figura 3 - Integration Service Project

Fonte - Elaboração própria

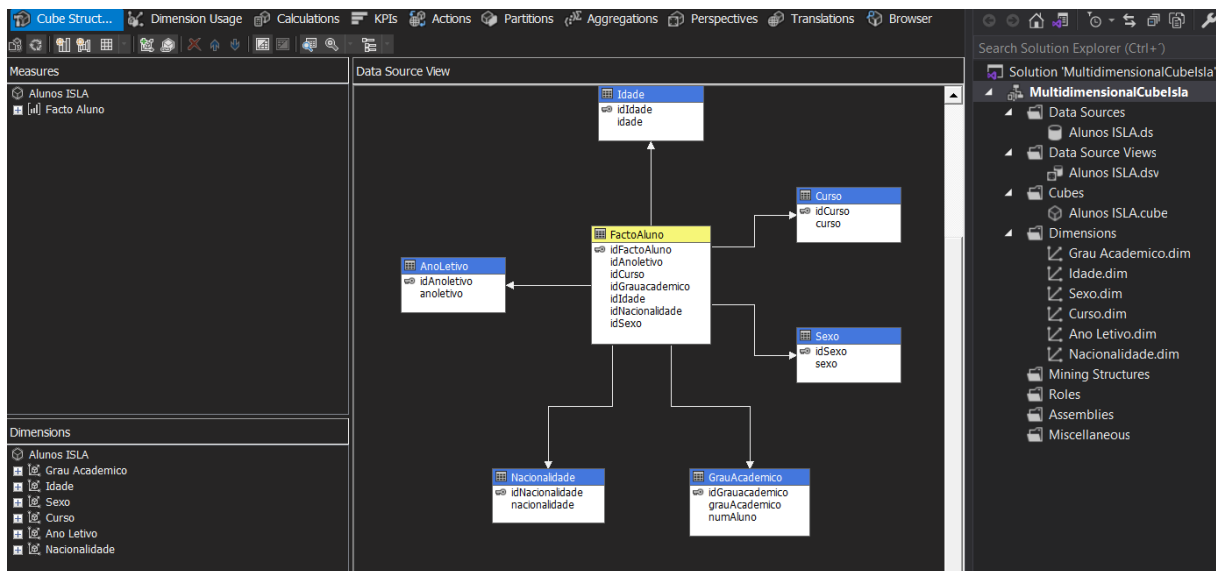


Figura 4 - Criação do Cubo no Analysis Multidimensions

Fonte - Elaboração própria

Dados de Alunos por Ano Letivo, Grau Acadêmico, Nacionalidade e Sexo

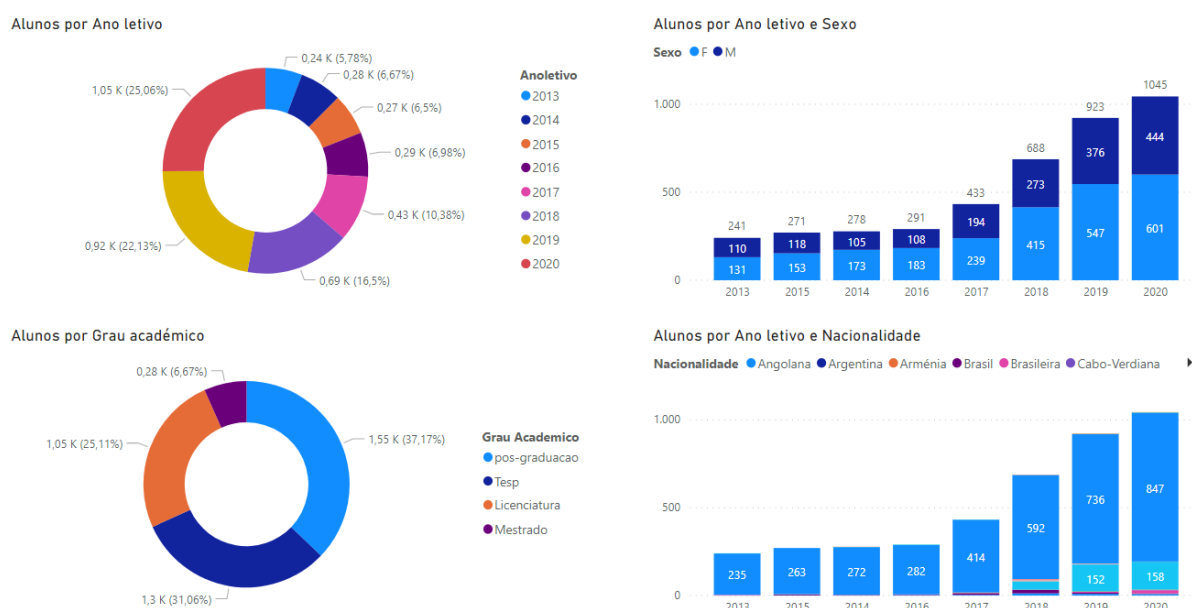


Figura 5 - Gráficos com dados Alunos no PowerBI

Fonte - Elaboração própria

Fase 4 – Predição do número de alunos para o próximo ano letivo através de Data Mining.

Para fazer a predição deste estudo foram exportados do cubo Molap os dados para um ficheiro csv e trabalhados em script realizada em Python utilizando a regressão linear para se obterem os resultados pretendidos.

RESULTADOS E DISCUSSÃO

Principais Resultados

Como principais resultados do estudo efetuado podemos verificar com base na figura 4 que o número de alunos no ISLA Santarém tem vindo a crescer desde 2013 até 2020, nos cursos com dois semestres as licenciaturas são as que têm mais alunos, os alunos do sexo feminino são em número superior aos do sexo masculino mantendo essa tendência desde 2013 até 2020, a predominância da nacionalidade de alunos no ISLA é portuguesa mas desde 2018 que alunos de nacionalidade guineense têm aumentado de número e em relação à idade, os alunos do isla ao longo dos anos de 2013 a 2020 tem vindo a ter alunos de idade mais jovem sendo a sua maioria entre os 20 e os 27 anos.

Pretendeu-se estimar o número total de estudantes, o número total de estudantes por grau (Mestrado, Pós-Graduação, Licenciatura e TeSP), o número total de estudantes por sexo, o número total de estudantes estrangeiros (também por grau e sexo), e o total de estudantes do curso de Licenciatura de Gestão de Recursos Humanos do ISLA Santarém para o ano letivo 2021-22, os resultados do teste estão apresentados na tabela 1. Na figura 6 é apresentado o gráfico com a evolução dos estudante da instituição desde 2013 e o valor previsto para o

próximo ano académico. Enquanto na figura 7 é mostrada a mesma evolução para o curso de Licenciatura de Gestão de Recursos Humanos.

Tabela 1. Totais para o ISLA Santarém em anos anteriores e previstos para o ano 2021-22(*).

	2013-14	2019-20	2020-21	2021-22
Total geral de estudantes	241	923	1045	1069*
Total de Mestrado	28	45	58	49*
Total de Pós-Graduação	132	307	355	345*
Total de Licenciatura	61	224	241	257*
Total de TeSP	20	347	391	417*
Total sexo feminino	131	547	601	626*
Total sexo masculino	110	376	444	443*
Total de estrangeiros	6	187	198	201*
Estrangeiros em Mestrado	1	8	10	10*
Estrangeiros em Pós-Graduação	2	5	10	11*
Estrangeiros em Licenciatura	2	48	46	45*
Estrangeiros em TeSP	1	126	132	133*
Estrangeiros de sexo feminino	4	106	113	112*
Estrangeiros de sexo masculino	2	81	85	88*
Total da Licenciatura em GRH	36	99	103	107*



Figura 6 – Evolução do total de estudantes do ISLA Santarém

Fonte - Elaboração própria

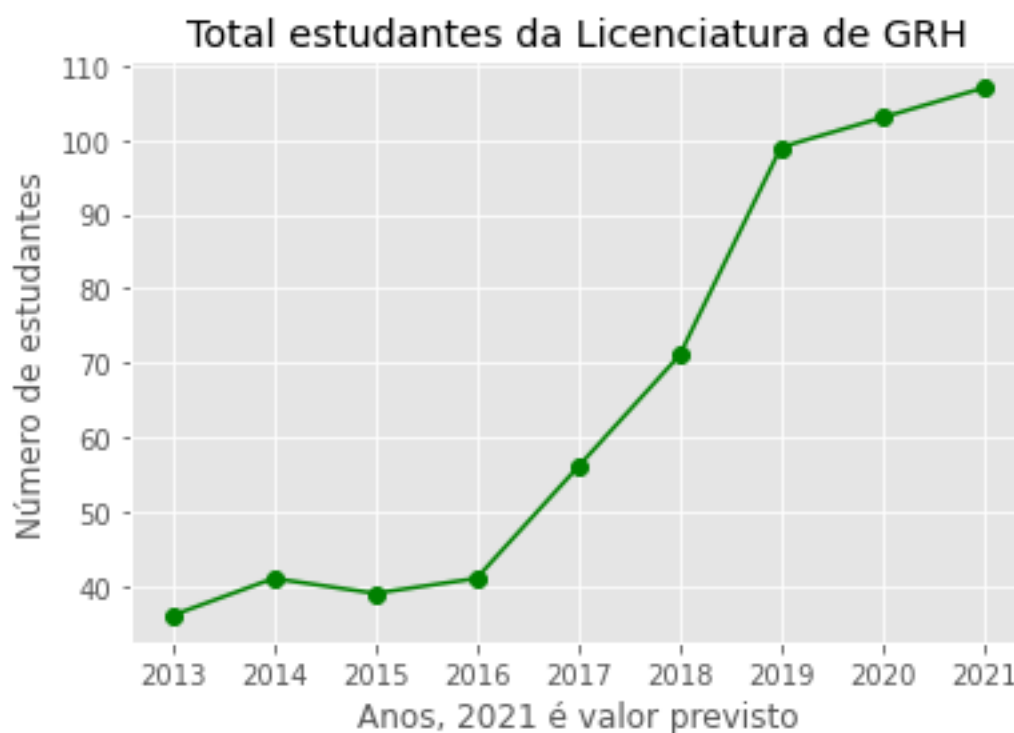


Figura 7 – Evolução do total de estudantes do curso de Licenciatura de Gestão de Recursos Humanos.

Fonte - Elaboração própria

Em relação aos anos anteriores verifica-se um aumento generalizado do número previsto de estudantes para o ano de 2021, no entanto o método previu uma redução nos cursos de grau de Mestrado e Pós-Graduação assim como do sexo masculino, isto deve-se a ser uma aproximação matemática da função de regressão linear que tem como entradas o histórico de anos anteriores o que pode influenciar os resultados futuros.

CONCLUSÕES

A instituição ISLA Santarém, pelos dados apresentados, tem vindo a ter uma evolução positiva na aquisição de alunos para os seus cursos. Usando o resultado de um sistema como o aqui apresentado pode ajudar no processo de gestão e organização do próximo ano letivo. Este estudo poderá ser melhorado fazendo a predição do aumento de alunos por curso, por sexo e por nacionalidade, algo a efetuar futuramente noutro estudo.

REFERÊNCIAS

- Aldrich, J. (1999). Determinacy in the linear model: Gauss to Bose and Koopmans. *International Statistical Review*, 67(2), 211–219. <https://doi.org/10.1111/j.1751-5823.1999.tb00427.x>
- Canto Castro, L., Duarte, J., & Nóvoas Pedro Martins, D. (2013). *Previsão do número de alunos em Portugal por regiões* Previsão do número de alunos em Portugal por regiões, 2013-2019. 2013–2019. [http://www.dgeec.mec.pt/np4/PREVISOES/%7B\\$clientServletPath%7D/?newsId=65&fileName=Modelo_de_previs_o_do_n_mero_de_alunos_e.pdf](http://www.dgeec.mec.pt/np4/PREVISOES/%7B$clientServletPath%7D/?newsId=65&fileName=Modelo_de_previs_o_do_n_mero_de_alunos_e.pdf)
- Hung, H. C., Liu, I. F., Liang, C. T., & Su, Y. S. (2020). Applying educational data mining to explore students' learning patterns in the flipped learning approach for coding education. *Symmetry*, 12(2), 1–14. <https://doi.org/10.3390/sym12020213>
- Kondo, N., Okubo, M., & Hatanaka, T. (2017). Early Detection of At-Risk Students Using Machine Learning Based on LMS Log Data. *Proceedings - 2017 6th IIAI International Congress on Advanced Applied Informatics, IIAI-AAI 2017*, 198–201. <https://doi.org/10.1109/IIAI-AAI.2017.51>
- Liu, J., Wu, S., & Zidek, J. V. (1997). On segmented multivariate regression. *Statistica Sinica*, 7(2), 497–525.
- Paul, S., Guatam, N., & Balint, R. (2000). *Seth Paul Nitin Guatam Raymond Balint*.
- Sobreiro, P., & Martinho, D. (n.d.). *SUPERIOR PREDICTING STUDENT DROPOUT IN A HIGHER EDUCATION* Resumo de poucos estudos para prever o abandono de alunos (Bogard , Helbig , Huff , & James , 2011), on (Aulck et al . , 2016) adeq Ferreira & Fernandes , 2015). tem sido aproveitado (Aulck , . 149–154.
- Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, 106, 234–240. <https://doi.org/10.1016/j.sbspro.2013.12.027>

PERFIL ACADÉMICO E PROFISSIONAL DOS AUTORES

Nelson Duarte

Técnico informático no ISLA Santarém há 10 anos, formador nas áreas de tecnologia e informática, docente, licenciado em Informática de Gestão pelo ISLA Santarém, aluno do mestrado em Engenharia de Tecnologias e Sistemas Web pelo ISLA Santarém e ISLA Gaia.

Ricardo Vardasca

Professor coordenador de ciência informáticas no ISLA Santarém, investigador integrado do INEGI-LAETA, investigador visitante da University of South Wales (Reino Unido) e professor externo da Universidad de Valencia (Espanha). Na ISLA Santarém é o diretor do Mestrado em Engenharia de Tecnologias e Serviços Web e o coordenador do Gabinete de Apoio a Projectos. Membro do conselho editorial das revistas: *Thermology International*, *The Imaging Science Journal*, *Educational Sciences* e *International Journal of E-Health and Medical Communications*.

Endereço postal dos autores.

ISLA Santarém
Largo Cândido dos Reis
2000-241 Santarém, Portugal