Relatório da Unidade Curricular

Análise de Dados em Informática

Ana Maria Dias Madureira Pereira

Universidade de Trás-os-Montes e Alto Douro

Escola de Ciências e Tecnologia
Departamento de Engenharias

Julho de 2021

Unidade Curricular

Análise de Dados em Informática - ANADI

Relatório da unidade curricular de Análise de Dados em Informática (ANADI) apresentado no âmbito das Provas de Agregação em Informática, em cumprimento do disposto na alínea b), do número 2, do artigo 8º do Decreto-Lei nº239/2007 de 19 de Junho de 2007.

Ana Maria Dias Madureira Pereira.

Julho de 2021

Índice

Índice de Figuras

Índice de Tabelas

1. Introdução

Apesar de se constatar ser a Estatística frequentemente uma das unidades curriculares mais complexas nos ciclos de estudo (CE) em engenharia, ciências sociais, ciências biológicas, entre outras, é cada vez mais consensual considerar-se esta área do conhecimento como uma ferramenta fundamental para a análise e interpretação de dados, mas em particular para o suporte de conclusões fundamentadas a partir da análise desses dados [2].

A sociedade atual é dominada pela geração de grandes volumes de dados e a necessidade de extração de conhecimento a partir deles. O próprio conceito de dados tem vindo a mudar ao longo dos tempos. Já não temos apenas números para tratar. Os dados agora consistem também em texto e imagens, que apresentam vários desafios para a respetiva triagem, organização e tratamento. D. Howell [1] referia que "*Statistics is not really about numbers; it is about understanding our world*" e H.G. Well que já no séc. IX profetizou que "*Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write*" [2].

A unidade curricular (UC) **Análise de Dados em Informática** (ANADI) surgiu como resultado da reforma do plano de estudos (Tabela 2) da Licenciatura em Engenharia Informática (LEI), do Instituto Superior de Engenharia do Porto (ISEP/P.PORTO), concretizada no ano letivo de 2015/2016. A candidata é regente da UC desde então, tendo-lhe sida proposta a dinamização da UC **Análise de Dados em Informática** incorporando algumas sugestões de auditores externos, fruto dos processos de avaliação e acreditação dos ciclos de estudo a funcionar no Departamento em Engenharia Informática (DEI). A sua génese foi pensada como resultando da integração de competências de duas áreas científicas [3]: Ciências e Tecnologias da Especialidade/Engenharia Informática e Ciências de Base no sentido de compensar a lacuna do CE na área da Estatística e principalmente na área de Análise de Dados. Dotar os alunos de competências no sentido da exploração e aplicação de ferramentas estatísticas na análise de dados. Mais do que conhecer os fundamentos matemáticos dos métodos/técnicas estatísticas considerou-se relevante no âmbito de ANADI, que os alunos perante um determinado *dataset* e os dados que o compõem, sejam capazes de identificar qual o gráfico mais adequado para a sua visualização ou o método mais adequado para a análise de inferência, assim como a sua interpretação. É dada particular ênfase à componente de seleção da técnica ou método a usar, em determinado cenário, e dependendo das variáveis em estudo, mas principalmente à análise e discussão dos resultados obtidos no suporte à extração de informação dos dados.

Sendo uma UC do 6º semestre da LEI, pretende-se que seja um suporte adicional na construção do relatório de fim de curso no âmbito da UC Projeto/Estágio (PESTI), particularmente no capítulo de Teste e Validação.

De forma a acompanhar os desenvolvimentos permanentes que se verificam na área, é não só natural, mas também indispensável, a atualização regular de conteúdos e de métodos de ensino de uma UC que pretende orientar-se para métodos de análise e processamento de dados. Além disso, dado que parte das atividades de investigação da candidata se tem centrado neste tópico, a

experiência resultante desta atividade terá, sem dúvida, uma contribuição significativa na lecionação da unidade curricular.

Assim sendo, a reformulação de ANADI proposta surge como o resultado de um processo de aprendizagem e de experiências adquiridas ao longo do período que a UC tem sido lecionada, da experiência acumulada pela sua responsável, nomeadamente em projetos de investigação aplicada na área de otimização e particularmente na análise de desempenho de algoritmos de otimização e de aprendizagem automática na resolução de problemas complexos. Assim como, das tendências de investigação que atualmente se verificam e dos desafios colocados à comunidade científica nesta área.

O relatório inicia-se com uma breve Introdução, contextualizando-se no capítulo 2 a unidade curricular de **Análise de Dados em Informática** como uma das áreas de suporte a duas áreas em evidência na atual conjuntura técnico-científica – Ciência dos Dados (*Data Science)* e *Big Data* - resultantes da combinação dos esforços conjuntos em Engenharia Informática, Estatística e Inteligência Artificial. Identificam-se os seus objetivos, como a sistematização do estado da arte na área de *Data Science* e análise de dados, as fronteiras do conhecimento e os desafios atuais e futuros existentes na área, bem como os resultados de aprendizagem esperados. Ainda neste capítulo, são apresentadas as estratégias de ensino e aprendizagem preconizadas, o resumo e o programa detalhado da unidade curricular e a respetiva planificação das sessões, os recursos pedagógicos, a bibliografia essencial e recomendada, e os métodos de avaliação de conhecimentos e competências. Os anexos incluem um exemplar dos enunciados dos trabalhos práticos (edição 2015/2016 e 2019/2020). São depois definidas as competências e o modelo letivo que se julga adequado para alcançar os objetivos propostos.

O capítulo 3 é dedicado ao ensino de análise de Dados em Informática, no contexto do espírito de Bolonha (1º ciclo de estudos), tendo como referência as principais universidades a nível nacional. São também analisados a nível internacional alguns CE que orientam os seus objetivos para a Análise de Dados. Perspetivam-se ainda neste capítulo, as dependências da atual e da futura UC, com as restantes unidades curriculares lecionadas na LEI e no Mestrado em Engenharia Informática (MEI) ou no Mestrado em Engenharia de Inteligência Artificial (MEIA), no âmbito da oferta letiva do ISEP/P.PORTO. Conclui-se com: a identificação dos tópicos a serem atualizados e introduzidos na nova unidade curricular; as metodologias de ensino e avaliação da proposta de reformulação da unidade curricular; e identificada a bibliografia principal.

Finalmente, no capítulo 4 são efetuadas algumas considerações finais. O relatório inclui ainda um conjunto de Anexos dedicados à apresentação dos enunciados das Fichas Teórico-Práticas (Anexo A), e dos enunciados dos Trabalhos Práticos (Anexo B).

2. Análise de Dados em Informática

A análise estatística é uma área bem definida no âmbito da Matemática e as ferramentas a que esta recorre têm evoluído de uma forma significativa nos últimos anos. A unidade curricular de **Análise de Dados em Informática** (ANADI), sobre a qual incide este relatório para apreciação no âmbito das provas conducentes à obtenção do título académico de Agregado, está inserida no 2º semestre do 3º ano da Licenciatura em Engenharia Informática (LEI), do Departamento em Engenharia Informática (DEI), do Instituto Superior de Engenharia do Porto (ISEP/P.PORTO).

A LEI inclui como principais objetivos a conceção e desenvolvimento de software, sistemas de informação, a administração de redes e segurança, e o tratamento de dados. A sua estrutura curricular [3] é baseada nas melhores práticas internacionais (ACM, IEEE-CS e CDIO Initiative), tendo sida distinguida com a certificação de qualidade EUR-ACE da Ordem dos Engenheiros. O ciclo de estudos da LEI apresenta como principais objetivos, em consonância com a missão do ISEP/P.PORTO:

- A formação integral de profissionais de engenharia informática que sejam capazes de se constituir como agentes de progresso, da inovação e da intervenção cultural e social;
- O desenvolvimento da investigação aplicada e da disponibilização social do conhecimento adquirido;
- O estímulo das capacidades que permitam aos graduados da LEI compreender e intervir globalmente no desenho do futuro da humanidade.

Neste capítulo, pretende-se descrever o funcionamento da Unidade Curricular de Análise de Dados em Informática. Serão apresentados o plano de estudos e a estrutura curricular, os objetivos da UC e as competências esperadas no final da UC, os conteúdos programáticos, a metodologia de ensino e de avaliação e a bibliografia e ferramentas de ensino aconselhadas. Serão ainda analisados os resultados finais nos anos letivos 2015/2016 a 2019/2020. O enquadramento de ANADI ao nível do Ensino Superior Politécnico e com as melhores práticas internacionais (ACM, IEEE-CS e CDIO Initiative) será realizado no capítulo 3.

2.1 Plano de Estudos e Estrutura Curricular

As unidades curriculares da LEI estão organizadas em 4 áreas científicas: Ciências de Base; Informática de Base; Ciências Complementares; e Ciências e Tecnologias da Especialidade/ Engenharia Informática (Tabela 1).

Tabela 1 - Áreas científicas da LEI [3]

Área Científica	Sigla	Créditos	
		Obrigatórios	Optativos
Ciências de Base	CB	44	-
Informática de Base	IB	52	-
Ciências Complementares	CC	21	-
Ciências e Tecnologias da Especialidade/Engenharia Informática	CTE-EI	63	-
Total		**180**	-

A unidade curricular ANADI surgiu como resultado da reforma do plano de estudos (Tabela 2) da Licenciatura em Engenharia Informática (LEI) concretizada no ano letivo de 2015/2016 [3]. A candidata é regente da UC desde então, tendo-lhe sida proposta a dinamização da UC "Análise de Dados em Informática" incorporando algumas sugestões de auditores externos fruto dos processos de avaliação e acreditação dos ciclos de estudo a funcionar no DEI. A sua génese foi pensada como resultando da integração de competências de duas áreas científicas: Ciências de Base e Ciências e Tecnologias da Especialidade/Engenharia Informática no sentido de compensar a lacuna da LEI na área da Estatística.

Tabela 2 - Plano de estudos da LEI [3]

		Unidades Curriculares	Áreas Científicas	Tipo	Horas de trabalho		Créditos
					Total	Contacto	
1º ano	1º Semestre	Álgebra Linear e Geometria Analítica	CB	Semestral	140	T - 12; TP - 48	5
		Algoritmia e Programação	CB	Semestral	168	T - 12; TP - 12; PL- 48	6
		Análise Matemática	CB	Semestral	140	T - 12; TP - 48	5
		Laboratório / Projeto I	CB/CC	Semestral	224	TP - 30; PL - 18; OT - 48	CB:5 CC:3
		Princípios da Computação	CB	Semestral	168	T - 12; TP - 12; PL - 48	6
	2º Semestre	Engenharia de Software	IB	Semestral	168	T- 12; TP - 12; PL - 48	6
		Laboratório / Projeto II	IB/CC	Semestral	224	TP - 30; PL - 12; OT - 48	IB:5 CC:3
		Matemática Computacional	CB	Semestral	140	T - 12; TP - 24; PL - 24	5
		Matemática Discreta	CB	Semestral	140	T - 12; TP - 48	5
		Paradigmas da Programação	IB	Semestral	168	T - 12; TP - 12; PL - 48	6
2º ano	1º Semestre	Arquitetura de Computadores	IB	Semestral	140	T - 12; TP - 12; PL - 48	5
		Bases de Dados	IB	Semestral	168	T - 12; TP - 12; PL - 48	6
		Estruturas de Informação	IB	Semestral	168	T - 12; TP - 12; PL - 48	6
		Física Aplicada	CB	Semestral	140	T - 12; TP - 12; PL - 24	6
		Laboratório / Projeto III	CTE-EI/ CC	Semestral	224	TP - 22,5; PL - 18; OT - 48	CTE-EI:5 CC: 3
	2º Semestre	Engenharia de Aplicações	CTE-EI	Semestral	168	T - 12; TP - 12; PL - 48	6
		Laboratório / Projeto IV	CTE-EI/ CC	Semestral	168	TP - 18; PL - 18; OT - 48	CTE-EI:4 CC: 2
		Linguagens e Programação	IB	Semestral	168	T - 12; TP - 12; PL - 48	6
		Redes de Computadores	IB	Semestral	168	T - 12; TP - 12; PL - 48	6
		Sistemas de Computadores	IB	Semestral	168	T - 12; TP - 12; PL - 48	6
3º ano	1º Semestre	Algoritmia Avançada	CTE-EI	Semestral	140	T - 12; TP - 12; PL - 36	5
		Arquitetura de Sistemas	CTE-EI	Semestral	140	T - 12; TP - 12; PL - 36	5
		Administração de Sistemas	CTE-EI	Semestral	140	T - 12; TP - 12; PL - 36	5
		Gestão	CC	Semestral	112	T - 24; TP - 30	4
		Sistemas Gráficos e Interação	CTE-EI	Semestral	140	T - 12; TP - 12; PL - 36	5
		Laboratório/Projeto V	CTE-EI/ CC	Semestral	168	TP - 15; PL - 18; OT - 42	CTE-EI:4 CC: 2
	2º Semestre	Informática nas Organizações	CTE-EI	Semestral	112	T - 15; PL - 30	4
		Comportamento Organizacional	CC	Semestral	112	T - 15; PL - 30	4
		Análise de Dados em Informática	CTE-EI/CB	Semestral	112	T - 15; TP - 30	CTE-EI:2 CB:2
		Projeto/Estágio	CTE-EI	Semestral	504	OT - 96	18

A 1ª edição da UC em 2015/2016 foi definida no sentido de dar conhecimentos estruturantes no planeamento do estudo estatístico no âmbito da Análise Exploratória de Dados (AED) para problemas da área da Informática, particularmente relacionados com o processo de análise de dados no suporte à tomada de decisão nas áreas de planeamento e gestão.

A UC requeria conhecimentos específicos prévios sobre análise de desempenho de algoritmos, Sistemas Gráficos, Análise de Sistemas, Engenharia de Software e conceitos básicos de Estatística.

Na edição 2019/2020 foram introduzidas algumas alterações no sentido da reformulação da UC que visa dar conhecimentos estruturantes no planeamento do estudo estatístico no âmbito da Análise Exploratória de Dados (AED) para problemas reais no suporte à tomada de decisão. A UC visa fornecer ao aluno conhecimentos e competências na aplicação de tópicos clássicos (classificação e regressão) e tópicos avançados de Aprendizagem Automática no tratamento de grandes volumes de dados - *Big Data* - numa perspetiva de Ciência dos Dados.

2.2 Objetivos

Como objetivos específicos considerou-se que no final da UC o aluno deveria ser capaz de, mediante conhecimentos sólidos sobre técnicas de estatística descritiva e inferencial, implementar a AED e:

- Discutir as técnicas básicas para a conceção de experiências no domínio da Engenharia Informática.
- Ser capaz de analisar e organizar dados de uma diversidade de fontes;
- Discutir as diferentes técnicas de estatística descritiva e inferencial para implementar a AED;
- Selecionar e usar ferramentas estatísticas no suporte ao processo de AED;
- Especificar o processo de análise de usabilidade, de análise da fiabilidade de sistemas, e análise de desempenho de algoritmos;
- Desenvolver trabalhos em grupo e produzir relatórios técnicos e artigos científicos.

Na edição 2019/2020, a UC foi reformulada acompanhando também a reformulação de Matemática Computacional (MATCP) com a introdução de tópicos na área da estatística. Neste sentido, foram definidos como resultados expectáveis (*outcomes*), em conformidade com os critérios EUR-ACE, no final da UC, o aluno deveria ser capaz de:

- CO1 - Discutir as técnicas básicas de AED e Aprendizagem Automática (AA) para a conceção de experiências no domínio de *Data Science*;
- C02 - A analisar e organizar dados de uma diversidade de fontes;
- CO3 - Discutir as diferentes técnicas de estatística descritiva e inferencial para implementar a AED;
- CO4 - Identificar, selecionar e usar ferramentas de AA adequadas no suporte ao processo de *Data Science*;
- CO5 - Formular os problemas reais no contexto de terminologia de AA e escolher a abordagem mais adequada para a sua resolução;
- CO6 - Implementar e otimizar os modelos relevantes dos dados e avaliar o desempenho e a comparação dos modelos;
- CO7 - Especificar e otimizar o processo de análise de desempenho dos modelos;
- CO8 - Desenvolver trabalhos em grupo e produzir relatórios técnicos e artigos científicos, e comunicação oral em Português/Inglês.

2.3 Competências

Como resultados expectáveis (*outcomes*), definidos em conformidade com os critérios EUR-ACE foram considerados, na edição 2015/2016 e seguintes:

- **Knowledge and Understanding (#KU)**
  - Discutir as técnicas básicas para a conceção de experiências no domínio da Engenharia Informática
  - Discutir as diferentes técnicas de estatística descritiva e inferencial para implementar a AED
- **Engineering Analysis (#EA)**
  - Ser capaz de analisar e organizar dados de uma diversidade de fontes

- **Investigations (#IN)**
  - Caracterizar o estado da arte das técnicas estatísticas/ferramentas para aplicação em AED e sua potencial aplicação em engenharia e ciências aplicadas
- **Engineering Practice (#EP)**
  - Selecionar ferramentas estatísticas para suporte ao processo de AED
- **Making Judgements (#MJ)**
  - Especificação e acompanhamento da análise de usabilidade
  - Especificação e acompanhamento da análise de fiabilidade
  - Especificação e acompanhamento do processo de análise de desempenho de algoritmos
- **Communication and Team-working (#CT)**
  - Desenvolver trabalhos em grupo e produzir relatórios técnicos e artigos científicos

Como resultados expectáveis (outcomes), definidos em conformidade com os critérios EUR-ACE, foram considerados, na edição 2019/2020 e 2020/2021, os seguintes:

- **Knowledge and Understanding (#KU)**
  - Discutir as técnicas básicas de AED e AA para a conceção de experiências no domínio de Data Science
  - Discutir as diferentes técnicas de estatística descritiva e inferencial para implementar a AED
- **Engineering Analysis (#EA)**
  - Ser capaz de analisar e organizar dados de uma diversidade de dados
- **Investigations (#IN)**
  - Caracterizar o estado da arte das técnicas estatísticas/ferramentas para AED e AA e sua potencial aplicação em engenharia e ciências aplicadas
- **Engineering Practice (#EP)**
  - Identificar, selecionar e usar ferramentas de AA adequadas no suporte ao processo de Data Science
- **Making Judgements (#MJ)**
  - Formulação de problemas reais no contexto de terminologia de AA e identificação da abordagem mais adequada para a resolução do problema
  - Especificação e acompanhamento do processo de construção e otimização de modelos relevantes dos dados
- **Communication and Team-working (#CT)**
  - Desenvolver trabalho em grupo e produzir relatórios técnicos e artigos científicos

2.4 Conteúdos Programáticos

Os conteúdos programáticos de ANADI pretendem responder às necessidades formativas relativas ao desenvolvimento e utilização eficaz de aplicações/ferramentas estatísticas que facilitem e agilizem a análise de dados. Inserindo-se numa área científica em constante evolução, os conteúdos programáticos procuram ser suficientemente abertos e as atividades letivas e o trabalho autónomo contemplam frequentemente pesquisa e revisão do estado da arte, de forma a dar resposta às necessidades de atualização tecnológica. Os conteúdos programáticos da UC "Análise de Dados em Informática" nas edições 2015/2016 até 2018/2019, foram os seguintes:

1 - Análise Exploratória de Dados (4 horas - 70%T+30%TP)

- Fases e Princípios Básicos do planeamento do estudo estatístico

- Dados determinísticos e aleatórios; População, Amostra e Estatísticas
- Probabilidades e Distribuições (varáveis discretas e continuas)
- Apresentação e resumo dos dados: Medidas de localização, dispersão, forma e medidas de Associação para variáveis contínuas, ordinais e nominais.

2 - Inferência Estatística (10 horas - 50%T+50%TP)

- Testes de hipóteses paramétricos: Inferência para uma, duas ou mais populações
- Testes de hipóteses não paramétricos: Inferência para uma, duas ou mais populações

3 - Regressão e Correlação (6 horas - 50%T+50%TP)

- Regressão linear simples
- Regressão Múltipla

4 - AED na área da Informática (22 horas - 20%T+80%TP). Os casos de estudo serão tratados de forma transversal com os tópicos anteriores. De referir: Análise de Usabilidade; Análise de Fiabilidade; Análise de Desempenho de algoritmos.

O conteúdo programático da UC foi revisto em 2019/2020 para se ajustar à evolução tecnológica, e às necessidades e desafios do mercado de trabalho, particularmente na área de iniciação à Ciência dos Dados (*Data Science*) e Técnicas de Aprendizagem Automática. A UC de Matemática Computacional (MATCP) tem vindo progressivamente a incluir tópicos da área da estatística, o que nos permitiu evoluir numa fase inicial para a componente aplicacional dos conceitos de estatística descritiva, probabilidade e testes de hipóteses, particularmente nos trabalhos práticos. Os conteúdos programáticos da UC "Análise de Dados em Informática" desde a edição 2019/2020, são os seguintes:

1- Data Science/BigData/Data Analytics (2h - 30%T+70%TP)

2- Análise Exploratória de Dados (4h - 30%T+70%TP)

- Estatística Descritiva
- Organização dos dados, análise e discussão dos resultados

3- Inferência Estatística (10 h - 40%T+60%TP)

- Testes de hipóteses paramétricos: Inferência para uma, duas ou mais populações
- Testes de hipóteses não paramétricos: Inferência para uma, duas ou mais populações

4- Regressão e Correlação (6h - 40%T+60%TP)

- Regressão linear simples
- Regressão Múltipla

5- Técnicas de Aprendizagem Automática (26h - 20%T+80%TP).

- Noções e aplicabilidade de técnicas de AA
- Regressão e Classificação (Aprendizagem supervisionada, Aprendizagem não-supervisionada, Aprendizagem por reforço, Deep Learning)

2.5 Plano Semanal das aulas

A UC é assegurada por 2 grupos de docentes conforme definido no plano de estudos, de acordo com as áreas científicas definidas: Ciências e Tecnologias da Especialidade/Engenharia Informática (CTE-EI) e Ciências de Base (CB). Até à edição 2018/2019, enquanto regente da UC fui responsável pelas aulas teóricas (T). As aulas Teórico-Práticas (TP) foram asseguradas por docentes do departamento de Matemática (da área da Estatística). O plano das aulas no ano letivo 2015/2016, para os diferentes tipos de aulas é apresentado na Tabela 3.

Tabela 3 - Plano das aulas no ano letivo 2015/2016

Semana	T – Teórica	T/P–Teórico-Prática
1ª 22-02 a 27-02	Conversação com os alunos quanto à metodologia de ensino/aprendizagem a ser seguida: objetivos, conteúdos programáticos, bibliografia e método de avaliação. Análise de dados: Definição, contexto, e fases. Teamwork Project 1 Assignment	Análise Exploratória de Dados Estatística Descritiva: Conceitos Básicos, População, Amostra e Estatísticas, Tipos de variáveis. Apresentação e resumo dos dados: Medidas de localização, dispersão e forma; Medidas de Associação para variáveis contínuas, ordinais e nominais Organização dos Dados com o R.
2ª 29-02 a 05-03	Análise Exploratória de Dados. Fases e Princípios Básicos do planeamento do estudo estatístico Análise de Usabilidade: conceito, estrutura e fases do processo de avaliação.	Estatística Descritiva: Exemplos e discussão dos resultados com o R.
3ª 07-03 a 12-03	Probabilidades e Distribuições (variáveis discretas e contínuas)	Análise de Usabilidade: recolha e organização dos dados. Seleção de medidas estatísticas para apresentação e resumo dos dados. Uso do R.
4ª 14-03 a 19-03	Estimação de parâmetros. Intervalos de Confiança.	Uso do R para Inferência estatística. Intervalos de confiança. Practical Work 1 Delivery
21-03 a 28-03	Férias da Páscoa	
5ª 29-03 a 02-04 (2ªf de Páscoa)	Não há aulas na 2ª feira	Uso do R para Inferência estatística. Intervalos de confiança. Análise e discussão dos resultados.
6ª 04-04 a 09-04	Inferência estatística: Testes de Hipóteses Paramétricos.	Uso do R para Inferência estatística. Testes de hipóteses paramétricos. Análise e discussão dos resultados.
7ª 11-04 a 16-04	Teamwork Project 2 Assignment Análise de Fiabilidade: definição, caracterização e análise dos dados.	Uso do R para Inferência estatística. Testes de hipóteses paramétricos. Análise e discussão dos resultados.
8ª 18-04 a 23-04	Inferência estatística: Testes Não Paramétricos.	Análise de Fiabilidade: definição, caracterização e análise dos dados. Análise e discussão dos resultados.
9ª 25-04 a 30-04 Feriado a 5.04	Não há aulas T e algumas TP	Uso do R para Inferência estatística. Testes Não Paramétricos. Análise e discussão dos resultados.
02-05 a 07-05	Queima das Fitas	
10ª 09-05 a 14-05	Inferência estatística: Testes Não Paramétricos. (Continuação Teamwork Project 3 Assignment	Practical Work 2 Delivery Uso do R para Inferência estatística. Testes Não Paramétricos. Análise e discussão dos resultados
11ª 16-05 a 21-05	Análise de desempenho de algoritmos: definição, caracterização e Plano de testes.	Uso do R para Inferência estatística. Testes Não Paramétricos. Análise e discussão dos resultados.
12ª 23-05 a 28-05	Análise de desempenho de algoritmos: Análise e discussão dos resultados.	Análise de desempenho de algoritmos: Análise e discussão dos resultados. Uso do R para suporte ao estudo estatístico.
13ª 30-05 a 04-06	Correlação e regressão. Regressão linear simples.	Uso do R para Correlação e regressão. Regressão linear simples.
14ª 06-06 a 11-06	Correlação e regressão. Regressão Múltipla.	Uso do R para Correlação e regressão. Regressão Múltipla Practical Work 3 Delivery
15ª 13-06 a 18-06	Análise de Desempenho: Trabalho Prático 3 Defesas e apresentações	Análise de Desempenho: Trabalho Prático 3 Defesas e apresentações

Desde a edição de 2019/2020, enquanto regente da UC fui responsável pelas aulas teóricas**.** As aulas TP das primeiras semanas foram asseguradas por docentes do departamento de Matemática (da área da Estatística). As aulas TP do 2º módulo – Aprendizagem Automática – foram asseguradas por docentes do DEI. O plano das aulas no ano letivo 2019/2020, para os diferentes tipos de aulas é apresentado na Tabela 4.

Tabela 4 - Plano das aulas no ano letivo 2019/2020

Semana	T – Teórica	T/P–Teórico-Prática
1ª 17-02 a 22-02	Conversação com os alunos quanto à metodologia de ensino/aprendizagem a ser seguida: objetivos, conteúdos programáticos, bibliografia e método de avaliação. Análise de dados: Definição, contexto, e fases. Análise Exploratória de Dados.	Estatística Descritiva: Exemplos e discussão dos resultados com o R.
2ª 24-02 a 29-02 Férias do Carnaval	Lançamento Enunciado do Trabalho Prático 1 Férias do Carnaval (2ª e 3ª)	Uso do R para Inferência estatística. Testes de hipóteses paramétricos. Análise e discussão dos resultados.
3ª 02-03 a 7-03	Inferência estatística: Testes de Hipóteses Paramétricos.	Uso do R para Inferência estatística. Testes de hipóteses paramétricos. Análise e discussão dos resultados.
4ª 9-03 a 14-03 Aulas de 2ª a 4ª	Inferência estatística: Testes de Hipóteses Paramétricos. Testes de Hipóteses não Paramétricos.	Uso do R para Inferência estatística. Testes de hipóteses não paramétricos. Análise e discussão dos resultados.
16-03 a 21-03	Suspensão das atividades letivas - Pandemia	
5ª 23-03 a 28-03	Testes de Hipóteses não Paramétricos. Correlação.	Uso do R para Testes de hipóteses não paramétricos e Correlação. Análise e discussão dos resultados.
6ª 30-03 a 04-04	Correlação e regressão. Regressão linear simples.	Uso do R para regressão. Regressão linear simples. Análise e discussão dos resultados.
06-04 a 11-04	Férias da Páscoa	
7ª 13-04 a 18-04	Correlação e regressão. Regressão múltipla. Entrega do TP1	Uso do R para regressão. Regressão linear múltipla. Análise e discussão dos resultados.
8ª 20-04 a 25-04	Avaliação e Defesa do TP1	Avaliação e Defesa do TP1
9ª 27-04 a 02-05 Feriado no dia 1-05 6ª Feira	Data Science/BigData/Data Analytics. Lançamento Enunciado do Trabalho Prático 2.	Modelos de Regressão Linear e Árvores de Regressão. Avaliação dos modelos: holdout e cross-validation (ligeiramente). Métricas de erros: MAE e RMSE.
10ª 04-05 a 09-05	Algoritmos de aprendizagem automática: conceitos, componentes e classificação. Noções e aplicabilidade de técnicas de AA.	Classificação: Árvores de Decisão – rpart -Modelo classificação -Matriz de confusão, accuracy
11ª 11-05 a 16-05	Aprendizagem supervisionada – classificação/regressão. Árvores de decisão.	Classificação: Árvores de Decisão -Introduzir as medidas: recall, precision, F1 -Treino/teste: holdout/validação cruzada
12ª 18-05 a 23-05	Aprendizagem supervisionada NN e KNN.	Classificação - Redes Neuronais -Normalização
13ª 25-05 a 30-05	Aprendizagem não-supervisionada – Clustering, k-medias.	Classificação - K-vizinhos Melhor K.
14ª 01-06 a 06-06	Aprendizagem por reforço (Reinforcement learning).	Apoio ao trabalho ao trabalho prático 2
15ª 08-06 a 13-06 Feriado 4ª e 5ª	Deep Learning. Entrega do TP2	Apoio ao trabalho ao trabalho prático 2
16ª 15-06 a 20-06	Avaliação e Defesa do TP2	Avaliação e Defesa do TP2

2.6 Metodologias de Ensino

Nas aulas Teóricas (T) são usados o método expositivo e interrogativo, e sempre que adequado, serão usadas técnicas do método ativo.

Nas aulas teórico-práticas (TP) são usadas preferencialmente técnicas do método ativo (trabalho de grupo, estudo de casos e aprendizagem baseada em problemas). As TP têm uma forte componente tecnológica, com uso da linguagem de programação R, para apoiar a dinâmica e objetivos da aula. As aulas TP suportam-se na utilização da ferramenta **RStudio** [10] na realização e concretização dos objetivos propostos para cada aula.

Usa-se a linguagem de programação Open-Source R por se considerar, que esta permite rentabilizar os objetivos com um menor esforço em termos de tempo de aprendizagem. A linguagem de programação open-source R [10], é muito popular entre os académicos e os investigadores da área da Ciência dos Dados, disponibilizando uma ampla variedade de bibliotecas e ferramentas para: a limpeza e pré-processamento de dados; a criação de gráficos e visualizações; e a especificação, treino e avaliação de técnicas de aprendizagem automática. O R foi desenvolvido especificamente para realizar a análise estatística e, consequentemente, tem uma maior diversidade de packages de análise estatística. Os recursos de visualização de dados do R são também mais ricos, mais sofisticados que os do Python e geralmente mais fáceis de gerar. A linguagem R é comumente utilizada no RStudio, um ambiente de desenvolvimento integrado (IDE) para análise estatística, visualização e reporting. As aplicações R podem ser usadas direta e interativamente na web via Shiny[1].

As estratégias ou metodologias de ensino usadas visam ajudar o aluno a assumir uma atitude de aprendizagem ativa, colaborativa e responsável, trabalho persistente e de aplicação de espírito crítico na análise e resolução de problemas.

Considerando que o principal objetivo da UC é proporcionar aos alunos competências para analisar e organizar os dados a partir de uma diversidade de fontes, o uso de técnicas de estatística descritiva e inferencial para implementar a AED e o desenho/especificação de testes/experiências para avaliar a análise de usabilidade, análise de fiabilidade e análise de desempenho, considerou-se que a forma mais eficaz para avaliar as competências adquiridas seria através de trabalhos práticos com dinamização do trabalho em equipa.

O uso do sistema de controle de versões *git* e *Bitbucket* poderá ser considerado nas atividades extra-aulas (ficheiros de dados, gráficos, script do R e documentos) e os *commits* poderão ser utilizados para avaliar a contribuição individual dos alunos.

As aulas T têm como principal foco a sistematização de conceitos teóricos relacionados com técnicas de estatística descritiva e inferencial na AED de dados oriundos de problemas da Engenharia Informática. As aulas teórico-práticas (TP) terão uma forte componente tecnológica, com uso intensivo da ferramenta estatística R para apoiar a dinâmica e objetivos da aula, definidos em Fichas TP's com a proposta de problemas (Anexo 1). Preferencialmente, as aulas devem ser lecionadas em laboratórios com PC's e portáteis.

2.7 Metodologia de Avaliação

A avaliação da UC é baseada preferencialmente em trabalhos práticos (projeto) com apresentação e defesa. Os critérios de avaliação são definidos no enunciado de modo que o aluno tenha conhecimento, à partida, dos critérios de êxito esperados e respetiva ponderação na nota global do

[1] Shiny é um package do RStudio que facilita o processo de construção de aplicações web interativas com R (https://shiny.rstudio.com/).

trabalho. A avaliação por trabalhos práticos considerou-se, dada a especificidade da UC, a forma mais adequada, tornando-se mais motivadora e menos exposta ao insucesso.

2.7.1 Ano Letivo 2015/2016

A nota da avaliação obtida durante o período letivo (FREQ) tem um peso de 100% na nota final da UC. A avaliação consiste na realização de 3 Trabalhos Práticos em grupo (máximo 3 alunos), de realização extra-aulas (submetidos no Moodle), com defesa/apresentação obrigatória, em grupo e individual.

Tabela 5 - Metodologia de Avaliação 2015/2016

Componente	Cotação	Nota Mínima
Trabalho Prático 1 /Projeto 1	25%	
Trabalho Prático 2 /Projeto 2	30%	7,5 valores
Trabalho Prático 3 /Projeto 3	45%	7,5 valores

Como o 1º trabalho Prático tem uma ponderação na nota final inferior a 30%, não é permitida a definição de nota mínima, de acordo com o Regulamento de Avaliação em vigor à data (Tabela 5). Como instrumentos de avaliação refere-se o desenvolvimento de 3 trabalhos práticos seguindo a metodologia descrita abaixo:

- **TP1 - Trabalho Prático 1 - Análise de Usabilidade**

Na realização deste trabalho pretende-se que os alunos desenvolvam o processo de Análise de Usabilidade, com o objetivo de testar e avaliar a usabilidade de sistemas, portais e sites. Sugerem-se os seguintes: Site das Finanças: e-fatura; Portal do ISEP; Moodle do ISEP; Página web do DEI; Ebay; Facebook, entre outros (ver enunciado completo no Anexo B1.1). Na Tabela 6 descreve-se a implementação do Trabalho Prático 1 (TP1).

Tabela 6 - Trabalho Prático 1 - Análise de Usabilidade

TÍTULO	**Análise de Usabilidade (Trabalho Prático 1)**
TIPO	Projeto
IMPLEMENTAÇÃO	2 semanas, realização extra-aulas (submetido no Moodle), realizado em grupos (max. 3 alunos da mesma turma TP), estimadas 10 horas de esforço por aluno, com uma apresentação individual obrigatória.
DESCRIÇÃO	O objetivo do projeto é a prática do aluno, as competências no uso de estatística descritiva para análise de usabilidade. Deve ser produzido um relatório com toda a documentação, o plano de testes, o inquérito elaborado, os resultados recolhidos e a análise e discussão dos resultados.
CRITÉRIOS DE AVALIAÇÃO, E RESPETIVAS COTAÇÕES	Existe uma grelha definida para a avaliação da componente técnica e o relatório cujo *template* é fornecido. O trabalho será avaliado em grupo e individualmente. O trabalho prático 1 tem peso de 25% na nota final.

- **TP2 - Trabalho Prático 2 - Análise de Fiabilidade**

Na realização deste trabalho prático pretendeu-se que os alunos desenvolvam o processo de Análise de Fiabilidade, com o objetivo de avaliar a fiabilidade da rede do **DEIspace** e da rede sem fios **EDUROAM** no DEI. No relatório final deverão ser documentadas todas as fases da Análise de Fiabilidade realizadas, contextualização do tema, recolha dos dados, organização do estudo estatístico, análise e discussão dos resultados, conclusões e anexos (ver enunciado completo no Anexo B1.2). Na Tabela 7, é descrita a implementação do Trabalho Prático 2 (TP2).

Tabela 7 - Trabalho Prático 2 - Análise de Fiabilidade

TÍTULO	**Análise de Fiabilidade (Trabalho Prático 2)**
TIPO	Projeto
IMPLEMENTAÇÃO	2 semanas, realização extra-aulas (submetido no Moodle), realizado em grupos (max. 3 alunos da mesma turma TP), estimadas 10 horas de esforço por aluno, com uma apresentação individual obrigatória.
DESCRIÇÃO	O objetivo do projeto é a prática do aluno, as competências no uso de estatística descritiva para análise de fiabilidade. Deve ser produzido um relatório com toda a documentação, o plano de testes, os resultados recolhidos e a análise e discussão dos resultados.
CRITÉRIOS DE AVALIAÇÃO, E RESPETIVAS COTAÇÕES	Existe uma grelha definida para a avaliação da componente técnica e o relatório cujo *template* é fornecido. O trabalho será avaliado em grupo e individualmente. O trabalho prático 2 tem peso de 30% na nota final. O aluno deverá atingir nota mínima de 7,5 valores.

- **TP3 - Trabalho Prático 3 - Análise de Desempenho**

O objetivo deste trabalho é praticar as competências na utilização de Análise Exploratória de Dados (Estatística Descritiva, Análise de Inferência e Correlação) e na Análise de Desempenho de algoritmos de otimização na resolução de problemas de escalonamento. Deve ser produzido um artigo científico (português ou inglês), conforme *template* fornecido, com o estado da arte sobre análise de desempenho de algoritmos, os dados em análise, a análise e discussão dos resultados e conclusões (ver enunciado completo no Anexo B1.3). Na Tabela 8, é descrita a implementação do Trabalho Prático 3 (TP3).

Tabela 8 - Trabalho Prático 3 - Análise de Desempenho

TÍTULO	**Análise de Desempenho (Trabalho Prático 3)**
TIPO	Projeto
IMPLEMENTAÇÃO	4 semanas, realização extra-aulas (submetido no Moodle), realizado em grupos (max. 3 alunos da mesma turma TP), estimadas 20 horas de esforço por aluno, com uma apresentação individual obrigatória.
DESCRIÇÃO	O objetivo do projeto é que o aluno pratique as competências na utilização da Análise Exploratória de Dados (Estatística Descritiva, Análise de Inferência e Correlação) para Análise de Desempenho de Algoritmos. Deve ser produzido um artigo científico com a definição do problema, plano de testes, resultados obtidos, análise e discussão dos resultados e conclusões.
CRITÉRIOS DE AVALIAÇÃO, E RESPETIVAS COTAÇÕES	Existe uma grelha definida para a avaliação da componente técnica e o artigo científico cujo template é fornecido. O trabalho será avaliado em grupo e individualmente. O trabalho prático 3 tem peso de 45% na nota final. O aluno deverá atingir nota mínima de 7,5 valores.

A nota final da UC é calculada seguindo a fórmula:

Classificação Final = x*TP1+y*TP2+z*TP3

onde:

- x=25%
- y=30% Nota mínima TP2 = 7,5 valores
- z=45% Nota mínima TP3 = 7,5 valores

É dada a possibilidade aos estudantes de poderem recuperar a nota dos Trabalhos práticos, desde que a respetiva classificação final seja <10, na época de recurso e na época especial.

Para melhoria de nota é dada a possibilidade aos estudantes de poderem melhorar a nota dos trabalhos práticos, desde que a classificação final seja >=10, na época de recurso ou na época especial.

2.7.2 Ano Letivo 2019/2020

Os estudantes têm a possibilidade de realizar toda a avaliação durante o período letivo (FREQ). Caso obtenham aprovação, estão dispensados da realização de avaliação final. Se isso não acontecer e os mínimos das componentes repetíveis não forem atingidos, têm ainda a possibilidade de realizar avaliação final, nas épocas de recurso e especial (Tabela 9).

A nota da avaliação obtida durante o período letivo (FREQ) tem um peso de 100% na nota final da UC. A avaliação consiste na realização de 2 Trabalhos Práticos em grupo (máximo 3 alunos), de realização extra-aulas (submetidos no Moodle), com defesa/apresentação obrigatória, em grupo e individual.

Tabela 9 - Metodologia de Avaliação 2019/2020

Componente	Cotação	Nota Mínima
Trabalho Prático 1 /Projeto 1	40%	7,5 valores
Trabalho Prático 2 /Projeto 2	60%	7,5 valores

Como instrumentos de avaliação refere-se o desenvolvimento de 2 trabalhos práticos seguindo a metodologia seguinte:

- **TP1 - Trabalho Prático 1 - Análise de Usabilidade**

O desenvolvimento de aplicações tem por base um trabalho de estudo sobre qual o conceito de design a adotar, a usabilidade (UI – interface do utilizador) e a experiência do utilizador (UX). Pretende-se com este trabalho realizar a análise de usabilidade do modelo de interação, de business apps de modo a avaliar a satisfação dos utilizadores. Sugere-se a análise de usabilidade dos sites de comércio eletrónico:

- Continente Online
- Auchan Online

O enunciado completo pode ser consultado no Anexo B2.1. Na Tabela 10, descreve-se a implementação do Trabalho Prático 1 (TP1).

Tabela 10 - Trabalho Prático 1 - Análise de Usabilidade

TÍTULO	**Análise de Usabilidade (Trabalho Prático 1)**
TIPO	Projeto
IMPLEMENTAÇÃO	3 semanas, realização extra-aulas (submetido no Moodle), em grupos (max. 3 alunos da mesma turma TP), estimadas 15 horas de esforço por aluno, com uma apresentação individual obrigatória.
DESCRIÇÃO	O objetivo do projeto é a prática do aluno, as competências no uso de estatística descritiva para análise de usabilidade. Deve ser produzido um relatório com toda a documentação, o plano de testes, o inquérito elaborado, os resultados recolhidos e a análise e discussão dos resultados.
CRITÉRIOS DE AVALIAÇÃO, E RESPETIVAS COTAÇÕES	Existe uma grelha definida para a avaliação da componente técnica e o relatório cujo template é fornecido. O trabalho será avaliado em grupo e individualmente. O trabalho prático 1 tem peso de 40% na nota final.

- **TP2 - Trabalho Prático 2 - Análise de Desempenho**

Pretende-se, no âmbito deste trabalho, praticar as competências na utilização de Análise Exploratória de Dados (Estatística Descritiva, Análise de Inferência e Correlação e Regressão) na Análise de Desempenho de algoritmos Aprendizagem automática. O objetivo principal deste trabalho consiste na aplicação de algoritmos de aprendizagem automática na exploração de dados e respetiva comparação, usando os testes estatísticos mais adequados. Deverá ser realizada a análise salarial da população masculina de uma dada região, com base nos atributos descritos, através de modelos de

classificação/regressão usando os algoritmos estudados: regressão linear, árvores de decisão, k-vizinhos-mais-próximos e redes neuronais. Deve ser produzido um artigo científico (português ou inglês), conforme *template* indicado, com o estado da arte sobre os diferentes algoritmos, os modelos desenvolvidos, os resultados obtidos, a análise e discussão dos resultados e as conclusões (ver enunciado completo no Anexo B2.2). Na Tabela 11, descreve-se a implementação do Trabalho Prático 2 (TP2).

Tabela 11 - Trabalho Prático 2 - Análise de Desempenho

TÍTULO	**Análise de Desempenho (Trabalho Prático 2)**
TIPO	Projeto
IMPLEMENTAÇÃO	4 semanas, realização extra-aulas (submetido no Moodle), em grupos (max. 3 alunos da mesma turma TP), estimadas 20 horas de esforço por aluno, com uma apresentação individual obrigatória.
DESCRIÇÃO	O objetivo do projeto é que o aluno pratique as competências na utilização da Análise Exploratória de Dados (Estatística Descritiva, Análise de Inferência e Correlação) para Análise de Desempenho de Algoritmos de Aprendizagem automática. Deve ser produzido um artigo científico com a definição do problema, plano de testes, resultados obtidos, análise e discussão dos resultados e conclusões.
CRITÉRIOS DE AVALIAÇÃO, E RESPETIVAS COTAÇÕES	Existe uma grelha definida para a avaliação da componente técnica e o artigo científico cujo *template* é fornecido. O trabalho será avaliado em grupo e individualmente. O trabalho prático 2 tem um peso de 60% na nota final. O aluno deverá atingir nota mínima de 7,5 valores.

A nota final da UC é calculada seguindo a fórmula:

Classificação Final = x*TP1+y*TP2

onde:

- x=40% Nota mínima TP1 = 7,5 valores
- y=60% Nota mínima TP2 = 7,5 valores

É dada a possibilidade aos estudantes de poderem recuperar a nota dos Trabalhos práticos, desde que a respetiva classificação final seja <10, na época de recurso e na época especial.

Para melhoria de nota é dada a possibilidade aos estudantes de poderem melhorar a nota dos trabalhos práticos, desde que a classificação final seja >=10, na época de recurso ou na época especial.

2.8 Bibliografia

Desde 2015/2016 foi identificado como material de ensino, o seguinte:

- Textos e Slides, Ana Madureira
- Douglas C. Montgomery, Design and Analysis of Experiments, 8th edition, John Wiley &Sons, New York, 2013.
- Joaquim P. Marques de Sá, Applied Statistics Using SPSS, STATISTICA, MATLAB and R, 2nd Edition, John Wiley & Sons, 2007.

E como Ferramentas de ensino/aprendizagem:

- Moodle (http://moodle.isep.ipp.pt/)
- R-Studio (https://www.rstudio.com/products/rpackages/)

No ano letivo 2019/2020 foi identificado como material de ensino o seguinte:

- Textos e Slides, Ana Madureira e João Matos
- Douglas C. Montgomery, Design and Analysis of Experiments, 8th edition. John Wiley & Sons, New York, 2013.
- Christopher Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

Como Ferramentas de ensino/aprendizagem:

- Moodle (http://moodle.isep.ipp.pt/)
- R-Studio (https://www.rstudio.com/products/rpackages/)

E como Material de Ensino Complementar

- Sheldon M. Ross, Introduction to Probability and Statistics for Engineers and Scientists. 4th edition, Elsevier Academic Press, 2009.
- Tom Mitchell, Machine Learning. McGraw-Hill, 1997.

2.9 Análise dos Resultados Finais

Nesta secção, pretende-se realizar uma análise comparativa dos resultados académicos nos anos letivos de funcionamento de ANADI.

A Tabela 12 apresenta uma sistematização dos resultados do desempenho curricular dos alunos na UC de **Análise de Dados em Informática,** nos 5 anos em análise, e de funcionamento (no ano letivo corrente ainda não estão disponíveis os resultados). A informação foi extraída dos relatórios da UC nos diferentes anos letivos.

Tabela 12 - Estatísticas de Resultados Finais

Estatísticas de Resultados Finais	**2015/2016**	**2016/2017**	**2017/2018**	**2018/19**	**2019/2020**
Estudantes inscritos na UC - I	314	308	315	318	305
Estudantes sem frequência/reprovados por faltas - NF	16	20	21	29	0
Estudantes sem classificação – NC	14	13	19	24	37
Estudantes avaliados (com nota final diferente de NF e NC)	284	275	275	265	268
Estudantes reprovados	53	47	57	65	45
Estudantes aprovados	261	261	258	253	260
Taxas de Aprovação					
Base: Número de estudantes inscritos (I)	83,12%	84,74%	81,90%	79,56%	85,25%
Base: Número de estudantes que frequentaram (I-NF)	87,58%	90,63%	87,76%	87,54%	85,25%
Base: Número de estudantes avaliados (I-NF-FT)	91,90%	94,91%	93,82%	95,47%	97,01%

Considerando as estatísticas dos resultados finais dos diferentes anos letivos, foram calculadas as taxas de aprovação tendo por base diferentes perspetivas: o número de estudantes inscritos (I); o número de estudantes que frequentaram (I-NF), e o número de estudantes avaliados (I-NF-FT).

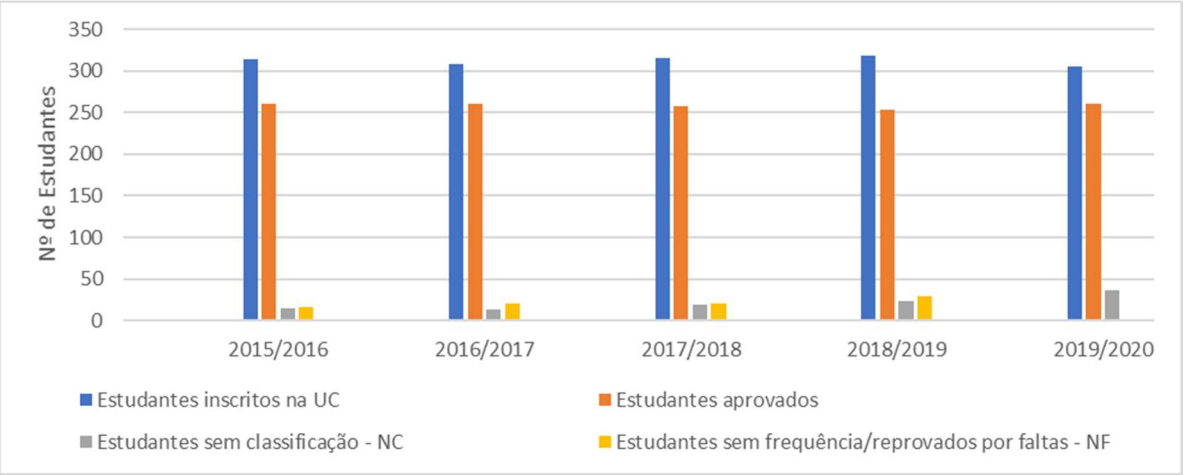

Figura 1 - Resumo do número de estudantes inscritos, aprovados, NC e NF

Da análise da Figura 1, é possível concluir não haver evidências estatísticas quanto a diferenças significativas no comportamento dos indicadores em análise, nos diferentes anos letivos, à exceção do ano letivo 2019/2020, em que por consequência das normas implementadas para fazer face à pandemia, não ter havido controlo de presenças, pelo que não houve, neste ano letivo, estudantes sem frequência (NF – Não Frequentou). Estes estudantes foram todos integrados na categoria de estudantes sem classificação (NC). Como se pode observar, os estudantes sem classificação (NC) e os estudantes sem frequência (NF), representam, em média, cerca de 12,3% dos alunos inscritos.

Refira-se ainda que considerando a situação de exceção associada ao período de pandemia vivido, outras variáveis poderão ter influenciado o processo de aprendizagem, não sendo possível relacionar a alteração curricular realizada na UC em 2019/2020, com os resultados e as taxas de aprovação.

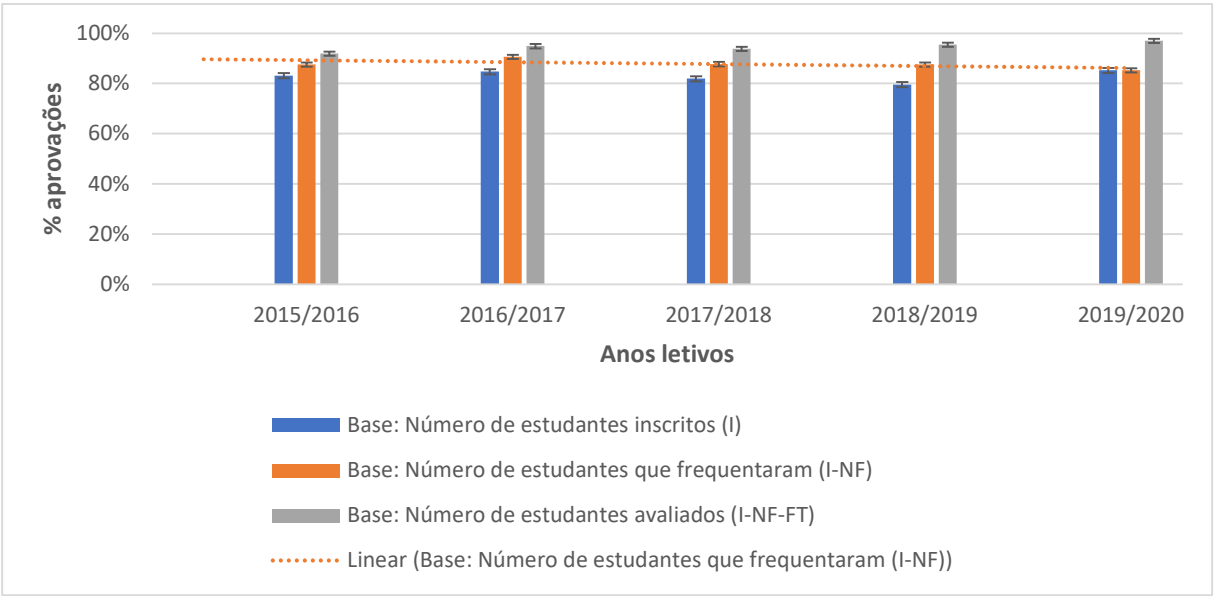

Figura 2 -Taxas de aprovação ANADI

A Figura 2, permite visualizar as taxas de aprovação nos anos letivos em análise, tendo por base o número de estudantes inscritos, o número de estudantes que frequentaram e o número de estudantes avaliados.

Tabela 13 - Sumário das Classificações Finais

Anos Letivos	Mínimo	1ºQuartil	Mediana	Média	3º Quartil	Máximo	Desvio Padrão	Skewness	Kurtosis
2015/2016	2.00	10.00	11.00	11.29	12.00	17.00	2.33	-0.9	3.41
2016/2017	2.00	12.00	14.00	13.35	15.00	19.00	2.61	-1.8	5.55
2017/2018	0.00	12.00	13.00	12.67	14.00	19.00	3.3	-2.09	5.69
2018/2019	3.00	11.00	12.00	11.77	13.00	17.00	2.17	-1.11	4.03
2019/2020	9.00	13.00	14.00	13.90	15.00	18.00	1.92	-0.56	-0.45

Das classificações finais dos alunos em cada ano letivo, foi possível extrair as métricas compiladas na Tabela 13, permitindo-nos analisar o comportamento das notas finais dos alunos. A média mais alta (13,9 valores) foi obtida no ano letivo 2019/2020, tendo-se alcançado igualmente uma menor dispersão das notas (com desvio padrão de 1,92). As medidas de forma, assimetria (*skewness*) e achatamento (*kurtosis*) permitem-nos caraterizar a forma de distribuição dos elementos da população (notas) em torno da média.

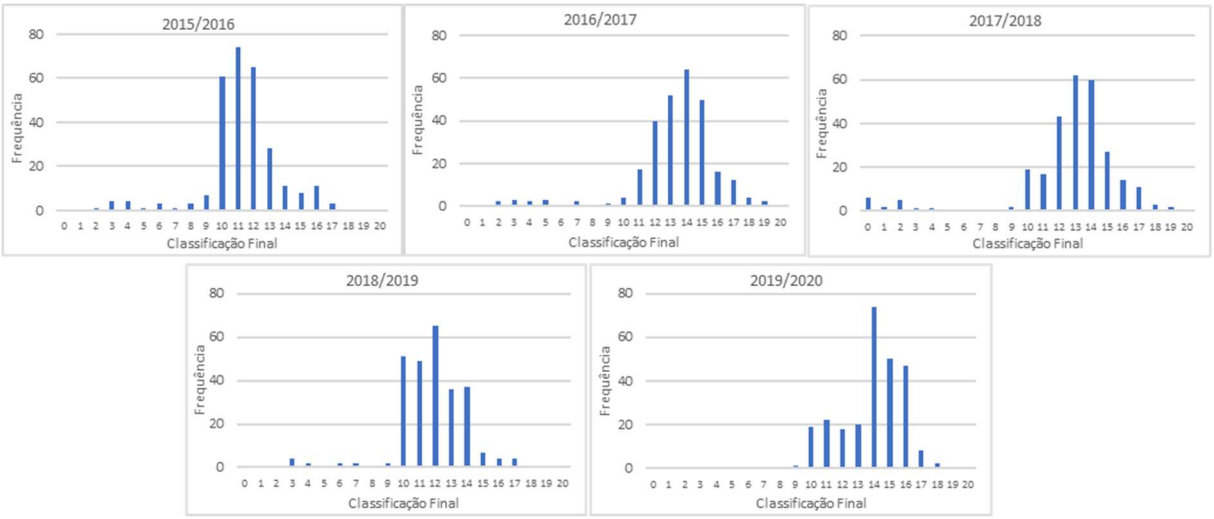

Figura 3 - Análise das classificações finais de ANADI

A análise dos valores destas métricas pode ser acompanhada da análise gráfica (Figura 3) das respetivas distribuições permitindo-nos ter uma perspetiva visual. Os dados obtidos permitem-nos concluir que a distribuição das notas nos diferentes anos letivos foi assimétrica à esquerda (valores negativos para a *skewness*). No que diz respeito ao achatamento, a análise dos dados da Tabela 13, permite-nos concluir que a distribuição das notas nos anos letivos 2015/2016 a 2018/2019 é leptocúrtica (valores positivos para a *kurtosis*). Para 2019/2020 evidencia ser ligeiramente platicúrtica (valor negativo para a *kurtosis*). Adicionalmente, analisando conjuntamente os valores da assimetria e do achatamento, poderemos comparar com a forma da distribuição teórica – Distribuição Normal – dado que os valores dos coeficientes referidos se aproximam de zero no intervalo]-0,5; +0,5[.

O gráfico da Figura 4 representa o comportamento e distribuição das classificações nos anos letivos em análise. É possível visualizar alguns *outliers*. O ano letivo 2015/2016 foi o que apresentou pior desempenho em termos de classificações a ANADI, com uma média de 11,29 valores. O ano letivo 2019/2020 parece indiciar ter sido o que apresentou melhor desempenho, com uma média global de 13,9 valores e um desvio padrão de 1,92, não havendo, no entanto, evidências estatísticas quanto à

existência de diferença significativa entre 2016/2017 (média de 13,35 valores e desvio padrão de 2,61) e 2019/2020.

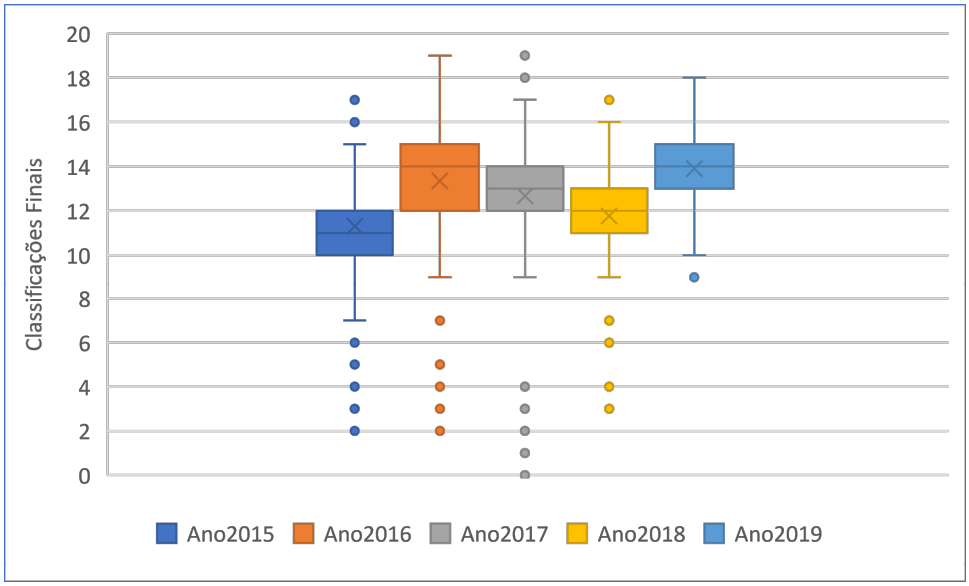

Figura 4 - Análise das classificações finais de ANADI

Os resultados refletem o sucesso escolar dos alunos que efetivamente frequentaram a UC com interesse e empenho e que se enquadraram nos critérios de êxito definidos na Ficha de Unidade Curricular de ANADI. Os resultados poderiam ainda melhorar, nomeadamente a média global, se os alunos fossem mais assíduos, participassem nas aulas e tirassem partido das atividades pedagógicas desenvolvidas no âmbito dos diferentes tipos de aulas. Refira-se ainda, no sentido da melhoria contínua, o interesse e necessidade das aulas TP serem alteradas para aulas PL, dado o enquadramento da UC no âmbito da aplicação de técnicas estatísticas a problemas no domínio da Análise de Dados em Informática. Neste sentido, as aulas requerem o uso de packages de análise estatística, não sendo possível a sua adequada dinamização em salas sem PCs. Para além disso, neste tipo de aulas, o acompanhamento do docente deve ser mais próximo, pelo que em turmas de 40-50 alunos se torna muito difícil. Nesta quarta edição da UC, considera-se terem sido alcançados os objetivos definidos, tendo em conta o esforço de todos os intervenientes, desde os alunos que se prontificaram a trazerem os seus portáteis, à gestão que se prontificou, dentro do possível, a escalonar as aulas TP em laboratórios, em vez das tradicionais salas de TP, e aos docentes que, apesar das condições difíceis, desenvolveram as medidas pedagógicas necessárias ao bom funcionamento da UC.

Os resultados dos inquéritos pedagógicos não têm apresentado significância estatística pelo que se optou por não serem apresentados e analisados.

Contudo, tem sido recolhido o testemunho dos alunos durante a defesa dos TP's, quanto à pertinência e ao interesse do TP e do tema em particular no seu perfil académico e em alguns casos, profissionais, uma vez que temos um grupo muito significativo de alunos com estatuto de trabalhador-estudante. O retorno recebido indica haver uma opinião global muito positiva por parte dos alunos relativamente ao funcionamento da UC e das competências adquiridas, particularmente o de Análise de Usabilidade e a sua utilização no suporte à escrita do capítulo Teste e Validação do relatório de Projeto/Estágio.

Este ano letivo está programada a realização de inquéritos aos alunos para cada módulo no sentido da melhoria contínua.

3. Ensino de Análise de Dados em Informática

Neste capítulo, será realizada a análise de ofertas educativas na área de Análise de Dados em Informática no ISEP/P.PORTO, mas também noutras instituições de ensino superior, a nível nacional e internacional. Sendo ANADI uma UC do 1º ciclo, contextualiza-se a UC e identificam-se as suas dependências e contribuições. Finalmente, propõe-se a reformulação da UC **Análise de Dados em Informática** da LEI.

3.1 No ISEP

A Declaração de Bolonha, com início oficial em junho de 1999, define um conjunto de etapas e de passos a implementar pelos sistemas de ensino superior europeu, no sentido da construção de um espaço de ensino superior globalmente harmonizado e reconhecido. A ideia de base é de que, salvo algumas especificidades nacionais, deve ser possível a um qualquer estudante do ensino superior, iniciar a sua formação académica, continuar os seus estudos, concluir a sua formação superior e obter um diploma europeu reconhecido em qualquer instituição do ensino superior de qualquer estado-membro [7]. Neste enquadramento, os sistemas de ensino superior deveriam ser dotados de uma organização estrutural de base idêntica, oferecer cursos e especializações semelhantes e comparáveis em termos de conteúdos e de duração, e conferir diplomas de valor equivalente tanto académico como profissional.

O processo de Bolonha estabeleceu o Espaço Europeu do Ensino Superior para facilitar a mobilidade dos estudantes e dos graduados, fazendo com que o ensino superior seja mais inclusivo e acessível e tornar o ensino superior na Europa mais atrativo e competitivo a nível mundial. No âmbito do Espaço Europeu do Ensino Superior, todos os países participantes se comprometeram no sentido da harmonização das estruturas do ensino superior a [7]:

- introduzir um sistema de ensino superior de três ciclos, que consiste em estudos de licenciatura, mestrado e doutoramento;
- assegurar o reconhecimento mútuo das qualificações e dos períodos de aprendizagem no estrangeiro concluídos noutras instituições do ensino superior;
- aplicar um sistema de garantia da qualidade, de modo a reforçar a qualidade e a relevância da aprendizagem e do ensino.

3.1.1 Licenciatura em Engenharia Informática

A missão do Ciclo de Estudos assenta na formação de graduados em Engenharia Informática capazes de uma rápida e harmoniosa integração com o meio envolvente (numa perspetiva europeia), sensibilizados para o empreendedorismo como forma preferencial de intervenção profissional/social e dotados de valências que lhes permitam reagir apropriadamente aos desafios da nova sociedade do século XXI, suportada por tecnologias de conhecimento omnipresentes.

Durante a formação incluem-se ainda a habituação a rigorosos métodos de trabalho, a aplicação de boas práticas e uma cultura permanente de avaliação (avaliar e ser avaliado). Para além do "saber fazer" e do "saber conceber", os licenciados deverão estar imbuídos da cultura de "saber aprender", numa perspetiva de formação e adaptação permanentes ao longo da vida, e da cultura de "saber ser", de forma a terem uma postura de cidadania ativa e construtiva.

Na prossecução desta missão será adotada uma praxis pedagógica que incluirá diversos paradigmas formativos e uma exposição a aplicações e casos de estudo que interliguem teoria e prática académicas com situações reais, escolhidas com base em critérios de relevância e utilidade.

Em termos de perfil, o licenciado em Engenharia Informática deverá ser possuidor de competências de conceção, desenvolvimento e administração de infraestruturas computacionais confiáveis e de sistemas de informação orientados para a sociedade do conhecimento. O diplomado neste CE será capaz de participar adequadamente em atividades relacionadas com a consultadoria, o empreendedorismo ou a investigação aplicada.

Como resultado da prática associada ao CE, o diplomado em Engenharia Informática estará ainda imbuído de códigos de conduta/boas práticas e será conhecedor de métodos de trabalho potenciadores de elevada eficácia e eficiência.

A unidade curricular ANADI surgiu como resultado da reforma do plano de estudos (Tabela 2) da Licenciatura em Engenharia Informática (LEI) concretizada no ano letivo de 2015/2016 [3]. A candidata é regente da UC desde então, tendo lhe sida proposta a dinamização da UC "**Análise de Dados em Informática**" incorporando algumas sugestões de auditores externos, fruto dos processos de avaliação e acreditação dos ciclos de estudo a funcionar no DEI. A sua génese foi pensada como resultando da integração de competências de duas áreas científicas: Ciências e Tecnologias da Especialidade/Engenharia Informática e Ciências de Base no sentido de compensar a lacuna do CE na área da Estatística e principalmente na Análise de Dados.

No sentido de se adequar por um lado às necessidades dos empregadores, e por outro, acompanhar os desenvolvimentos científicos e tecnológicos, a UC foi sofrendo pequenas alterações e adequações ao longo das suas edições.

Serão descritas ao longo deste documento as adequações realizadas quer em termos de conteúdos programáticos quer em termos de metodologias de avaliação, de modo a melhor refletir nas competências dos alunos a evolução científica e tecnológica.

Estando a Unidade Curricular (UC) de ANADI inserida numa Licenciatura do Ensino Superior Politécnico convém atender a algumas características deste tipo de estabelecimento de ensino. O ponto 1 do artigo 7º da Lei 62/2007, de 10 de setembro de 2007, que corresponde ao Regime Jurídico das Instituições de Ensino Superior, define a especificidade do Ensino Superior Politécnico da seguinte forma:

"Os institutos politécnicos e demais instituições de ensino politécnico são instituições de alto nível orientadas para a criação, transmissão e difusão da cultura e do saber de natureza profissional, através da articulação do estudo, do ensino, da investigação orientada e do desenvolvimento experimental "

A UC de Análise de Dados em Informática integra o 2º semestre do 3º ano da Licenciatura em Engenharia Informática. Esta UC visa dar conhecimentos estruturantes no planeamento do estudo estatístico no âmbito da Análise Exploratória de Dados (AED) para problemas da área da Informática, particularmente relacionados com o processo de análise de dados no suporte à tomada de decisão de planeamento e gestão.

A informática está na base de todas as atividades da sociedade atual, desde aplicações aos sistemas inteligentes, passando pelas bases de dados empresariais. A Engenharia Informática é transversal a todas as áreas ou setores de atividade e é a base para a construção do futuro, permitindo desenvolver soluções informáticas avançadas nas mais diversas áreas de atividade, tais como: comércio eletrónico,

saúde, ambiente, seguros, banca, transportes, desporto, jogos, economia e redes sociais, sendo nesta conjuntura uma área com grande procura de profissionais com formação nesta área.

Neste momento, o DEI tem como ofertas letivas uma licenciatura – em Engenharia Informática – que inclui a conceção e desenvolvimento de software, sistemas de informação, administração de redes e segurança e tratamento de dados, e três mestrados: o Mestrado em Engenharia Informática, o Mestrado em Engenharia de Inteligência Artificial e o Mestrado em Engenharia de Sistemas Computacionais Críticos.

A LEI inclui como principais objetivos a conceção e desenvolvimento de software, sistemas de informação, a administração de redes e segurança e o tratamento de dados. A sua estrutura curricular [3] é baseada nas melhores práticas internacionais (ACM, IEEE e CDIO Initiative), tendo sido distinguida com a certificação de qualidade EUR-ACE da Ordem dos Engenheiros. O ciclo de estudos da LEI apresenta como principais objetivos, em consonância com a missão do ISEP/P.PORTO:

- A formação integral de profissionais de engenharia informática que sejam capazes de se constituir como agentes de progresso, da inovação e da intervenção cultural e social;
- O desenvolvimento da investigação aplicada e da disponibilização social do conhecimento adquirido;
- O estímulo das capacidades que permitam aos graduados da LEI compreender e intervir globalmente no desenho do futuro da humanidade.

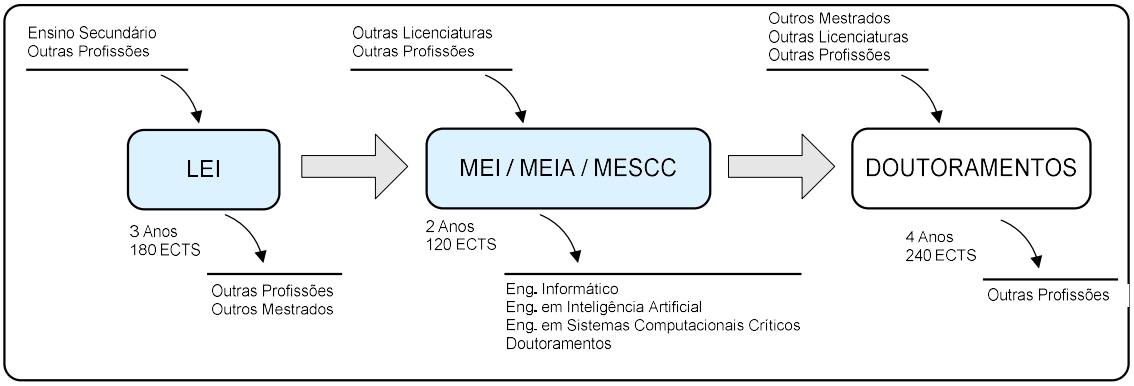

Figura 5 - Oferta letiva do DEI/ISEP: LEI, MEI, MEIA e MESCC

A Figura 5 ilustra o modelo adotado segundo as 3 vertentes principais: a duração de cada ciclo de estudos, o esforço realizado (em ECTS) e tipos de acesso (entrada e saída) de cada ciclo. Refira-se que até ao momento as Instituições do Ensino Superior Politécnico (IESP) não conferem ainda o grau de doutor (3º ciclo). Nos politécnicos nacionais e no ISEP/IPP em particular, existem atualmente condições para a formação ao nível do 3º ciclo, havendo estudantes de doutoramento a realizar os seus trabalhos nos centros de I&D residentes ou externos, com a orientação efetiva de docentes e investigadores do ISEP, mas que têm de estar inscritos em universidades, nacionais ou estrangeiras, nas quais irão defender as suas teses. Os trabalhos de doutoramento são um elemento importante no desenvolvimento da I&D+i realizada e esta é cada vez mais importante para a promoção do conhecimento necessário ao desenvolvimento económico, social e das empresas da região.

Salienta-se o conjunto de recomendações apresentadas pela OCDE em fevereiro de 2018, com o propósito de reforçar o desempenho e o impacto das atividades e das instituições de Investigação e Desenvolvimento (I&D) e de ensino superior em Portugal, numa perspetiva internacional e num contexto multidisciplinar. De forma a atingir estes objetivos, foram introduzidas algumas alterações ao regime jurídico dos graus e diplomas de ensino superior, através Decreto-Lei n.º 65/2018 de 16 de agosto, que vem reforçar as exigências sobre a capacidade das instituições de ensino superior para

desenvolver atividades de I&D, segundo o subsistema em causa, passando estas exigências a ser consideradas para efeitos de acreditação em todos os ciclos de estudos. Neste sentido, foi realçada a garantia que a acreditação de ciclos de estudos conducentes ao grau de doutor dependeria da existência de ambientes próprios de investigação de elevada qualidade, designadamente considerando os resultados da avaliação das unidades de I&D, regularmente realizada pela FCT, e a integração alargada dos docentes desse ciclo de estudos em unidades com classificação mínima de Muito Bom na área científica correspondente [8].

Várias são as iniciativas que têm vindo a ser desenvolvidas a nível nacional, no sentido de retirar a limitação legal que impede os politécnicos de outorgar o grau de doutor, ficando a acreditação em cada caso dependente dos requisitos atuais, já contemplados no Regime Jurídico dos graus e diplomas do ensino superior, na sua redação atual (Decreto-Lei n.º 65/2018, de 16 de agosto). A concretização destes objetivos implica, no entanto, a necessidade da alteração da Lei de Bases do Sistema Educativo – Lei n.º 46/86, de 14 de outubro - alterada pelas Leis n.ºs 115/97, de 19 de setembro, 49/2005, de 30 de agosto, e 85/2009, de 27 de agosto – e do Regime Jurídico das Instituições de Ensino Superior – Lei n.º 62/2007, de 10 de setembro.

São estas as razões pelas quais, no âmbito deste documento, nos centramos apenas ao nível dos CE do 1º e 2º ciclo, embora deixando em aberto a interligação com o 3º ciclo, que se espera num futuro próximo.

3.1.2 Mestrados no Departamento em Engenharia Informática

Em consonância com a missão do ISEP, os mestres deverão ser capazes de antecipar tendências das tecnologias e da sociedade, descobrir e dotar-se atempadamente do conhecimento necessário, atuar rapidamente com vista ao aproveitamento das oportunidades emergentes e contribuir para mudar construtivamente a sociedade (“saber liderar”, “saber empreender” e “saber inovar”). Como referido anteriormente, em termos de continuidade de estudos à LEI, o DEI faculta 3 mestrados[9]:

- **Mestrado em Engenharia Informática (MEI) [4]**
  - Com 4 áreas de especialização: Sistemas Computacionais, Sistemas de Informação e Conhecimento, Sistemas Gráficos e Multimédia, e Engenharia de Software. Encontra-se a aguardar acreditação pela A3ES a área de Engenharia de Dados;
  - Possui a certificação EUR-ACE e ABET - entidade global de acreditação de programas universitários em ciências naturais e aplicadas, informática, engenharia e tecnologias de engenharia.
- **Mestrado em Engenharia de Inteligência Artificial (MEIA) [5]**
  - Com um foco abrangente nos vários domínios da IA, incentiva à resolução de problemas complexos, ao raciocínio crítico e à criatividade;
  - O plano de estudos tenta ir ao encontro das necessidades das empresas que apontam a IA como o domínio de maior aposta no futuro.
- **Mestrado em Engenharia de Sistemas Computacionais Críticos (MESCC) [6]**
  - Confere competências para analisar de forma coerente e metódica um Sistema Computacional Crítico (SCC) como um todo, identificando possíveis problemas e oportunidades de melhoria;
  - O plano de estudos inclui unidades curriculares que abrangem as áreas de análise, projeto, implementação e operação de SCC, tanto na agregação de novo valor aos produtos desenvolvidos pelas empresas, como na abordagem de desafios sociais relevantes.

A UC de ANADI surge neste contexto como uma UC introdutória aos tópicos relacionados em Ciências dos Dados e Técnicas de Aprendizagem Automática, particularmente importantes ao MEI nas áreas de especialização - Sistemas de Informação e Conhecimento e Engenharia de Dados – e ao MEIA.

3.1.3 Relação com outras Unidades Curriculares da LEI

A unidade curricular de **Análise de Dados em Informática** é lecionada no 2º semestre do 3º ano da LEI. Esta UC visa dar conhecimentos estruturantes no planeamento do estudo estatístico no âmbito da Análise Estatística de Dados (AED) para problemas da área da Informática, particularmente relacionados com o processo de análise de dados no suporte à tomada de decisão de planeamento e gestão. Requer conhecimento específico prévio sobre análise de desempenho de algoritmos, usabilidade e fiabilidade de sistemas, engenharia de software e conceitos básicos de estatística.

O âmbito da UC **Análise de Dados em Informática** é relativamente abrangente, envolvendo conhecimentos multidisciplinares, oriundos de vários domínios do conhecimento da área da Engenharia Informática e da Matemática (estatística) e da Inteligência Artificial.

A avaliação da UC é baseada em trabalhos práticos com apresentação e defesa, promovendo a transparência dos critérios de avaliação, definidos no enunciado dos trabalhos. As UC's na área de Programação (Algoritmia e Programação, Estruturas de Informação e Paradigmas da Programação) fornecem os fundamentos de algoritmia, programação e estruturação necessários para a compreensão de metodologias e algoritmos e a sua implementação prática nos trabalhos desenvolvidos na UC. A UC de Algoritmia Avançada dota os estudantes de conhecimentos importantes Métodos de Resolução Automática de Problemas, Pesquisa e Otimização, assim como, uma abordagem inicial ao tema da Inteligência Artificial. Estes conhecimentos são fundamentais para a componente de desenvolvimento dos Trabalhos Práticos (Projeto), em que se espera a implementação computacional do estudo estatístico, usando preferencialmente a ferramenta computacional RSudio [10].

A UC na área da Estatística (Matemática Computacional - MATCP) fornece os fundamentos matemáticos necessários ao conhecimento e compreensão da terminologia e conceitos básicos da teoria da estatística necessários à compreensão na análise estatística descritiva, mas principalmente inferencial.

As unidades curriculares lecionadas na LEI que se relacionam mais diretamente com ANADI encontram-se esquematizadas na Figura 6. As UCs específicas referem-se às áreas que ANADI tem usado como caso de estudo (Análise de fiabilidade, Análise de Usabilidade e Análise de Desempenho).

A UC de ANADI tem apenas 3 horas de contacto semanais, em 15 semanas, no 2º semestre do 3ºano da LEI do ISEP, pelo que se favorece a perspetiva aplicacional das técnicas de análise estatística no suporte à análise de dados, com especial enfoque na correta interpretação dos outputs dos métodos implementados num script do R. No âmbito de ANADI, um dos objetivos mais relevantes é dotar os alunos de competências na componente de interpretação, análise e discussão dos resultados.

Pretende-se que, com as competências adquiridas em ANADI, os alunos sejam capazes de desenvolver e construir de forma mais robusta, o capítulo de Testes e Validação do relatório de fim de curso no âmbito de Projeto/Estágio (PESTI) da LEI. Realça-se a importância que outras UC's possuem no processo de desenvolvimento do projeto/Estágio (PESTI), nomeadamente Engenharia de Software, Engenharia de Aplicações e Arquitetura de Sistemas.

Sendo ANADI uma UC com foco mais orientado para a utilização de ferramentas estatísticas e não especialmente para a componente teórica e matemática, da análise estatística no desempenho de modelos e algoritmos de Aprendizagem Automática, considera-se que a UC apresenta interesse para a continuidade de estudos ao nível dos Mestrados em Engenharia Informática nas áreas de especialização - Sistemas de Informação e Conhecimento e Engenharia de Dados - e ao Mestrado em Engenharia de Inteligência Artificial do ISEP/P.PORTO mas também para outros mestrados na área da Engenharia Informática ou Ciência dos Dados.

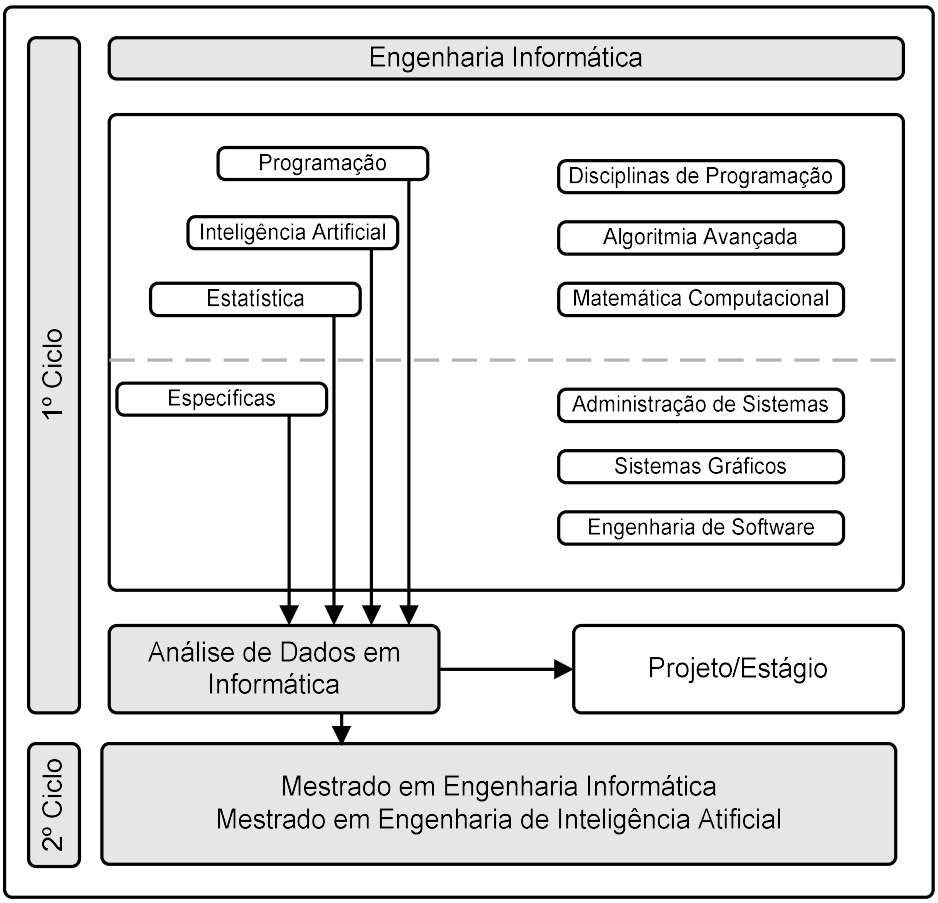

Figura 6 - Enquadramento da UC de ANADI com outras UCs da LEI e do MEI/MEIA

A UC de ANADI surge neste contexto como uma UC introdutória aos tópicos relacionados em Ciência dos Dados e Técnicas de Aprendizagem Automática, particularmente importantes no suporte às UCs do MEI e do MEIA do ISEP. Identificou-se também relevância da UC ao nível do mestrado em Engenharia Biomédica do ISEP, sendo esta uma área muito rica em casos de estudo para tratamento de dados, sejam eles estruturados, mas particularmente não estruturados (imagens, áudio, vídeo).

3.2 Noutras Instituições

Nesta secção, pretende-se analisar ofertas educativas na área de Análise de Dados, no contexto da Engenharia Informática, a nível nacional e internacional.

3.2.1 Instituições de Ensino Superior Nacionais

Relativamente a Instituições de Ensino Superior a nível nacional, são várias as que oferecem a unidade curricular de "Análise de Dados" (ou similar), como parte integrante do CE ao nível da Licenciatura ou Mestrado (1º ou 2º ciclo). Na Tabela 14, foi realizada uma sistematização de algumas UCs que abordam os conteúdos programáticos, mesmo que parcialmente, em CE na área da Engenharia Informática ou similar, tendo por base informação disponibilizada nas páginas web das respetivas instituições.

Tabela 14 - Unidades Curriculares para o ensino da **Análise de Dados** (instituições nacionais)

Instituição	Ciclo de Estudos	Modelo Ano|Semestre	Unidade Curricular
Universidade Nova de Lisboa	Mestrado Integrado em Engenharia Informática [11]	4º|1º	Prospeção e Análise de Dados
Faculdade de Engenharia da Universidade do Porto	Mestrado Integrado em Engenharia Informática e Computação [12]	1º|2º 2º|2º	Métodos Estatísticos Inteligência Artificial
Universidade do Minho	Mestrado Integrado em Engenharia Informática [13]: -Perfil de Especialização: Engenharia do Conhecimento -Perfil de Especialização: Ciência de Dados -Perfil de Especialização: Machine Learning: Fundamentos e Aplicações	2º|1º 4º|1º 4º|1º 4º|1º	Estatística Aplicada Análise de Dados Aprendizagem Automática Métricas em Machine Learning
Instituto Superior Técnico – Universidade de Lisboa	Mestrado em Engenharia Informática e de Computadores - Alameda e Taguspark [14] Mestrado de Bolonha Engenharia e Ciência de Dados [15]	1º|1º 1º|1º 1º|1º	Ciência de Dados Aprendizagem Automática Estatística Computacional
Universidade de Coimbra	Mestrado em Engenharia e Ciência de Dados [16] Mestrado em Engenharia Informática - Sistemas Inteligentes [17]	1º|1º 1º|1º	Tópicos de Ciência dos Dados Metodologias Experimentais em Informática
Faculdade de Ciências da UP	Mestrado em Ciência de Dados (Data Science) [18] Pósgraduação Estatística Computacional e Análise de Dados [19]	1º|1º 1º|2º	Introdução à Ciência de Dados Estatística e Análise de Dados
Universidade de Aveiro	Licenciatura em Engenharia Informática [20] Mestrado em Engenharia Informática [21]	3º|1º 1º|1º	Tópicos de Aprendizagem Automática Exploração de Dados
Universidade de Trás-os-Montes e Alto Douro [22]	Licenciatura em Engenharia Informática [23] Licenciatura em Matemática Aplicada e Ciência de Dados [24]	2º|1º 3º|1º 1º|2º 2º|2º 3º|2º	Métodos Estatísticos Inteligência Artificial Estatística aplicada Laboratório avançado de matemática e ciência de dados Aprendizagem automática
Instituto Superior de Engenharia de Coimbra (ISEC-IPC)	Licenciatura em Engenharia Informática - Ramo de Sistemas de Informação [25] Mestrado em Engenharia Informática - Tecnologias da Informação e do Conhecimento [26]	1º|2º 2º|1º 3º|1º 1º|2º 1º|2º	Métodos Estatísticos Introdução à Inteligência Artificial Inteligência Computacional Machine Learning Análise de Dados
Instituto Superior de Engenharia de Lisboa (ISEL-IPC)	Licenciatura em Engenharia Informática e de Computadores [27] Mestrado em Engenharia Informática e de Computadores [28]	2º|1º 1º|1º 2º|1º	Probabilidade e Estatística Aprendizagem e Mineração de Dados Inteligência Artificial e Sistemas Cognitivos

Da pesquisa realizada foi também possível verificar que os diferentes CE introduzem, de uma forma genérica, os fundamentos da análise de dados, em particular e parcialmente, na perspetiva de Métodos Estatísticos, Probabilidades e Estatística, e Inteligência Artificial, abordando as bases teóricas e práticas da Estatística ao nível do 1º ciclo ou vertente de iniciação à Inteligência Artificial. Em ciclos de estudo do 2º ciclo, a Análise de Dados, quer do ponto de vista aplicacional de ferramentas

estatísticas, quer com recurso a técnicas de Aprendizagem Automática, é mais frequentemente encontrada. Refira-se ainda que alguns dos CE analisados abordam esta problemática numa perspetiva de *Business Intelligence*.

3.2.2 Instituições de Ensino Superior Internacionais

Nesta secção, pretende-se analisar ofertas educativas na área de Análise de Dados ou afim, no contexto da Engenharia Informática a nível internacional. Na Tabela 15, foi realizada uma sistematização, não exaustiva, das UC que abordam os conteúdos programáticos, mesmo que parcialmente, em CE na área da Engenharia Informática ou similar, tendo por base informação disponibilizada nas páginas web das respetivas instituições.

Tabela 15 - Unidades Curriculares para o ensino da **Data Analytics ou afim** (instituições internacionais)

Instituição	Ciclo de Estudos	Modelo Ano|Semestre	Unidade Curricular
Texas A&M University, USA	Master of Science in analytics [29]		Applied Analytics (Machine Learning)
Oxford University UK	Computer science [30]		Probabilistic model checking Probability and computing
Universidad Autónoma de Madrid [31]	Grado en Ingeniería Informática Grado en Ingeniería Informática y Matemáticas		Probabilidad y Estadística Probabilidad y Estadística
Universidad Complutense de Madrid [32]	Grado en Ingeniería Informática Grado en Ingeniería del Software Grado en Ingeniería de Computadores		Probabilidad y Estadística Estadística aplicada Métodos Estadísticos
Université Sorbonne Paris [33]	Licence Informatique	3º|2º	Probabilités statistiques et application à l'analyse de données
Jacobs University, Bremen, Germany [34]	M.Sc. in Data Engineering		Big Data and Data Engineering
IU International University of Applied Sciences, Germany [35]	Master Data Science (online)	1º|1º	Advanced Statistics Data Science
Barcelona Graduate School of Economics [36]	Master Program in Data Science Methodology		Statistics and Machine Learning

Os conteúdos programáticos de ANADI são comumente encontrados em ciclos de estudo, particularmente ao nível do 2º ciclo, em mestrados na área da Ciência dos Dados [37].

3.3 Reformulação da Unidade Curricular

O termo "Computing" não engloba apenas uma única área de estudo. Durante a década de 1990, mudanças nesta área, nas tecnologias de comunicação e seus efeitos sociais levaram a modificações nesta família de disciplinas [38]:

- A Engenharia Informática emerge da Engenharia Eletrotécnica;
- As Ciências da Computação evoluíram para uma área científica mais madura;
- Os Sistemas de Informação expandem-se à medida que os computadores se tornaram a base de processos organizacionais e ambientes de trabalho;
- As Tecnologias de Informação emergem como uma nova disciplina que fomenta a construção e manutenção de infraestruturas informáticas;
- A Engenharia de Software emerge como uma disciplina baseada em Informática/Engenharia Informática.

Os avanços no desenvolvimento de currículos académicos, a nível mundial, expandiram o âmbito das disciplinas tradicionais na área de *Computing*: Engenharia Informática, Informática, Sistemas de Informação, Tecnologias da Informação e Engenharia de Software. A necessidade contínua e crescente de segurança da informação, e dos dados como recurso e impulsionadores do processo de tomada de decisões têm sido as áreas centrais de desenvolvimento na última década. Recentes esforços curriculares conduziram a desenvolvimentos significativos na área de cibersegurança, na ciência dos dados e noutras áreas emergentes de estudo. Embora estes esforços sejam geralmente reconhecidos no contexto das fronteiras do ensino da Informática ou da Engenharia Informática.

O *Computing Curricula 2020* (CC2020) é uma iniciativa lançada conjuntamente por várias sociedades profissionais, em 2021, com o objetivo de sintetizar o estado atual das diretrizes curriculares para ciclos de estudo que concedem diplomas ao nível da licenciatura nesta área científica, bem como propor uma visão para futuras diretrizes curriculares [38]. Esta iniciativa visa não só refletir o estado da arte no ensino e prática da Engenharia Informática ou afins, mas também fornecer informação sobre o futuro da área do ensino da Engenharia Informática para a próxima década. Este documento envolveu um grupo de trabalho com representantes de instituições do ensino superior, da indústria e de estruturas governamentais. A *Association for Computing Machinery* (ACM) e o *IEEE Computer Society* (IEEE-CS) lideraram o grupo de trabalho envolvido na criação do CC2020.

Após reflexão sobre as últimas recomendações curriculares da ACM e do IEEE-CS com a publicação do CC2020 [38], é possível identificar uma transição do processo de aprendizagem baseada no conhecimento (*knowledge-based learning*) para a aprendizagem baseada em competências (*competency-based learning*). Neste contexto, a competência requer a demonstração do comportamento humano com conhecimento e competências. Em termos gerais, pode pensar-se em competências como as qualidades que um indivíduo deve possuir para ser eficaz numa profissão, função, tarefa ou dever.

Nos meios académicos é consensual considerar-se que o sucesso no desenvolvimento de carreira requer três aspetos [38]:

- **Conhecimentos (Knowledge)** — "*know-what*" — uma proficiência em conceitos e conteúdos fundamentais e na aplicação de aprendizagem a novas situações;
- **Competências (Skills)** — "*know-how*" — a capacidade de realizar tarefas com determinados resultados;
- **Disposições (Dispositions)** — "*know-why*" — tendências intelectuais, sociais ou morais.

Tendo por base trabalhos prévios, o Relatório CC2020 [38] especifica a competência como sendo composta pelas dimensões K-S-D observadas no desempenho de uma tarefa, T.

Competências = [Knowledge + Skills + Dispositions] na tarefa T

Uma competência enumera conhecimentos, *skills* e disposições que são observáveis no cumprimento de uma tarefa realizada no âmbito de um contexto de trabalho. A figura 7 ilustra a estrutura conceptual da competência.

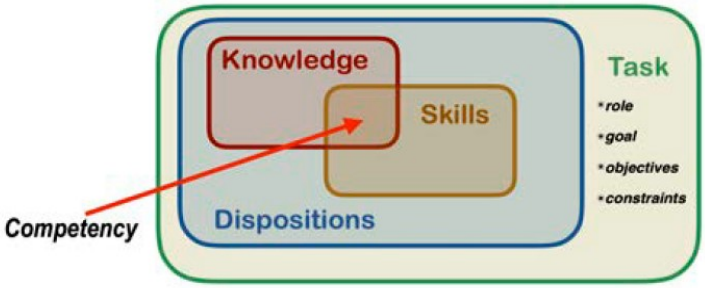

Figura 7 - Estrutura Conceptual do Modelo de Competências CC2020 [38]

Trabalhos recentes têm vindo a enfatizar uma visão baseada em competências no ensino da Engenharia Informática. Esta tendência tem sido um importante pivô na definição de padrões curriculares que codificam as expectativas para ir além da simples comunicação de conhecimento. No contexto mais amplo da indústria, profissional e na sociedade, uma descrição curricular centrada nas competências centra-se na capacidade de um indivíduo para realizar e aplicar a sua formação em informática num serviço prático e profissional à sociedade.

Um modelo de competências coerente para a definição de um plano de estudos na área da Engenharia Informática deve promover e descrever claramente os benefícios práticos dos ciclos de estudo aos seus *stakeholders*: estudantes, docentes, empregadores, entidades acreditadoras, legislador, a sociedade. Descrever a competência informática num contexto prático muda o foco da descrição do corpo de conhecimento em relação a uma área disciplinar e orienta-a para a concretização e desempenho pragmático dos alunos. As descrições do que os licenciados podem fazer em situações práticas, substituem descrições de aprendizagem e memorização de conteúdos. A competência descreve mais eficazmente as expectativas de resultados (*outcomes*), desafia as instituições de ensino superior a prepararem profissionais em Informática mais competentes, e permite que a sociedade reconheça os propósitos e os benefícios do ensino da informática no âmbito do quadro de competências definido.

Várias são as áreas tecnológicas e científicas emergentes para as quais os ciclos de estudo, nesta área científica, se devem preparar para dar respostas em termos de formação dos seus graduados: desde a Computação em Nuvem, Cidades Inteligentes, Sustentabilidade, Computação Paralela, Internet das Coisas (IoT), Deep learning (DL), Moeda Digital, Blockchain, Realidade virtual/aumentada, cibersegurança, aspetos de segurança, privacidade, e ética da Inteligência Artificial, entre outros.

Considerando a reformulação da UC Matemática Computacional (MATCP) da LEI, no ano letivo 2020/2021, incluindo os tópicos teóricos e práticos subjacentes a Análise Estatística Descritiva e Inferencial, torna-se oportuno reformular os conteúdos de ANADI no próximo ano letivo, 2021/2022, reduzindo esta componente e reforçando a componente de Aprendizagem Automática, particularmente para as técnicas de aprendizagem não supervisionada e *Deep Learning*, sendo até à edição atual apenas abordados na perspetiva teórica.

Pretende-se que o estudante, após concluir esta UC, reconheça a necessidade de utilizar técnicas de AED, que compreenda alguns dos algoritmos de aprendizagem automática e use ferramentas de desenvolvimento de AA para criação, treino e validação de modelos de análise de dados (R, Phyton). No final desta UC, o estudante deverá ser capaz de:

CO1. Discutir as técnicas básicas de AED e AA para a conceção de experiências no domínio de *Data Science*;
C02. Analisar e organizar dados de uma diversidade de fontes;
CO3. Discutir as diferentes técnicas de estatística descritiva e inferencial para implementar a AED;
CO4. Identificar, selecionar e usar ferramentas de AED e AA adequadas no suporte ao processo de *Data Science.*

CO5.Formular problemas reais no contexto de terminologia de AA e identificação da abordagem mais adequada para a resolução do problema;
CO6.Especificar o processo de construção e otimização de modelos de AA relevantes dos dados;
CO7.Especificar o processo de análise de desempenho dos modelos;
CO8.Trabalhar em grupo e produzir relatórios técnicos e artigos científicos, e concretizar a comunicação oral em Português/Inglês;
O9. Caracterizar o estado da arte das técnicas estatísticas/ferramentas para AED e AA e sua potencial aplicação em engenharia e ciências aplicadas.

3.3.1 Conteúdos programáticos

O programa da UC foi revisto em 2019/2020 para se ajustar quer à evolução tecnológica, quer às necessidades e desafios do mercado de trabalho. Os conteúdos programáticos da UC "Análise de Dados em Informática" para a edição 2021/2022 são os seguintes:

1. Artificial Intelligence/Data Science/BigData/Data Analytics (4h - 50%T+50%TP)
2. Statistical Learning versus Machine Learning (1h – 100%T)
3. Técnicas de Aprendizagem Automática (40h-30%T+80%TP)
   - Conceitos básicos
   - Noções e aplicabilidade de técnicas de AA
   - Problemas: Regressão e Classificação
   - Aprendizagem supervisionada (Árvores de Decisão, Redes Neuronais, kNN, SVM)
   - Aprendizagem não-supervisionada (*Clustering*, *Kmeans*)
   - Aprendizagem por reforço
   - *Deep Learning*

3.3.2 Metodologia de Avaliação

A avaliação da UC será suportada num Trabalho Prático, com 2 iterações, com peso de 100% na nota final da UC. A avaliação consiste na realização de 1 Trabalho Prático em grupo, de realização extra-aulas (submetido no Moodle), com defesa/apresentação obrigatória, em grupo e individual.

Instrumentos de avaliação:

- TP1 - Trabalho Prático 1 (Teamwork Project assignment - iteration 1), 3 semanas, realização extra-aulas, 15 horas de trabalho estimadas por aluno, defesa obrigatória, peso de 40% na nota final.
- TP2 - Trabalho Prático 1 (Teamwork Project assignment - iteration 2), 4 semanas, realização extra-aulas, 20 horas de trabalho estimadas por aluno, defesa obrigatória, peso de 60% na nota final.

É dada a possibilidade aos estudantes de poderem recuperar a nota dos Trabalhos Práticos, desde que a nota final seja <10 (reprovado na época normal), na época de RECURSO e/ou na época ESPECIAL, mantendo-se a fórmula de cálculo.

O Trabalho prático pode ser melhorado, mantendo-se a mesma fórmula de cálculo da classificação final:

Classificação Final = x*TP1+y*TP2

Em que:

x=40% Min TP1 = 7,5 valores

y=60% Min TP2 = 7,5 valores

3.3.3 Metodologia de Ensino

As aulas T têm como principal foco a sistematização de conceitos teóricos relacionados com técnicas de estatística na AED de dados oriundos de problemas da Engenharia Informática. As aulas TP terão uma forte componente prática, com uso intensivo da linguagem de programação R para apoiar a dinâmica e objetivos da aula, definidos em Fichas TP com a proposta de problemas (similares às apresentadas no Anexo 1). Preferencialmente, as aulas devem ser em laboratórios com PC's e portáteis.

Considerando que o principal objetivo da UC é dotar os estudantes de competências para analisar e organizar os dados a partir de uma diversidade de fontes, o uso de técnicas de estatística descritiva e inferencial para implementar a AED e o desenho/especificação de testes/experiências para avaliar a análise de desempenho de abordagens baseadas em Aprendizagem Automática. O recurso a Phyton Considera-se que a forma mais eficaz para avaliar as competências adquiridas é através de trabalhos práticos com dinamização do trabalho em equipa, a escrita científica e a apresentação oral dos resultados.

3.3.4 Bibliografia

- Nailong Zhang, A Tour of Data Science: Learn R and Python in Parallel, Chapman & Hall/CRC Data Data Science Series), 2020.
- Douglas C. Montgomery, Design and Analysis of Experiments, 8th edition, John Wiley & Sons, New York, 2013.
- Luis Torgo, A Linguagem R. Programação para Análise de Dados, Escolar Editora, 2009.
- Tutoriais, recursos online, e artigos científicos selecionados sobre tópicos específicos da UC
- Jake Vanderplas, Python Data Science Handbook, OReilly Media & Inc & USA, 2016.
- Chris Albon, Machine Learning with Python Cookbook: Practical solutions from preprocessing to deep learning, O'Reilly Media, Inc, USA, 2018.

Esta reformulação vem reforçar as competências, em análise de dados, dos nossos estudantes à saída do 1º ciclo, tornando-os mais competitivos para o mercado de trabalho, mas também melhor preparados para a continuidade de estudos ao nível do 2º ciclo, quer nos mestrados residentes no DEI/ISEP, quer para outros mestrados externos à instituição.

4. Conclusões

Na unidade curricular de **Análise de Dados em Informática** pretende-se que os alunos adquiram conhecimentos fundamentais do estado da arte na área de Análise de Dados e Aprendizagem Automática, e que sejam capazes de reconhecer as questões em aberto e os desafios atualmente existentes associados à área de *BigData* e *Data Science* (Ciência dos Dados) em direta consonância com os objetivos traçados no *Computing Curricula 2020 da ACM/IEEE-CS* [38]. Espera-se que estes possam alavancar futuras atividades de investigação, mas também que lhes permitam a aplicação dos conhecimentos adquiridos em situações concretas, na resolução de problemas reais. Em particular, espera-se a utilização de ferramentas estatísticas na análise de dados e do desempenho de modelos de técnicas de Aprendizagem Automática na resolução de problemas de classificação e regressão.

A unidade curricular de ANADI orienta o seu foco especialmente para Análise de Dados adquiridos em contextos reais, dada a pertinência, os desafios e as oportunidades que este tipo de problemas proporciona. No entanto, originam também novos desafios de análise e processamento (informação adquirida em ambientes não controlados), o que justifica a necessidade da procura de novas abordagens e a investigação de novas metodologias e algoritmos.

Começou-se por descrever o funcionamento, objetivos de ANADI desde a sua 1ª edição em 2015/2016, conteúdos programáticos, metodologias de ensino e avaliação e bibliografia adotada. Foi realizado um estudo estatístico dos resultados académicos nas diferentes edições da UC.

No capítulo 3, foi feito um esforço no levantamento e respetiva sistematização de ofertas educativas na área de Análise de Dados em Informática no ISEP/P.PORTO, mas também noutras instituições de ensino superior, a nível nacional e internacional. Neste sentido, sendo uma UC do 1º ciclo, contextualiza-se a UC e identificam-se as suas dependências com outras UCs lecionadas no CE de Engenharia Informática da LEI e nos mestrados do Departamento em Engenharia Informática. Finalmente, propõe-se a reformulação da UC **Análise de Dados em Informática** da LEI.

Reconhece-se que a unidade curricular proposta apresenta bastantes desafios, por se tratar de uma unidade curricular claramente ambiciosa, tanto ao nível do esforço que implica ao nível da docência como para os alunos, principalmente pela exigência da fundamentação teórica, abrangendo um conjunto de conceitos muito vasto. No entanto, tendo em atenção o interesse e atualidade da área em que se insere, às bases gerais que proporciona, aos exemplos e casos de estudo concretos que apresenta, acredita-se que o resultado final será certamente compensador.

Referências

[1]. D.C. Howell, Fundamental Statistics for the Behavioral Sciences, 4ª Edição, Belmont, CA: Duxbury Press, 1999.

[2]. J. Marôco, Análise Estatística com o SPSS Statistics, 5ªedição, Report Number, 2011.

[3]. Alteração do plano de estudos da Licenciatura em Engenharia Informática, lecionada no Instituto Superior de Engenharia do Porto, despacho nº 2074/2015, IIª série do Diário da República nº 40, de 26 de fevereiro de 2015.

[4]. Publicação do plano de estudos do Mestrado em Engenharia Informática, lecionado no Instituto Superior de Engenharia, despacho nº 10143/2015 do Diário da República nº176, Série II de 09 de setembro de 2015.

[5]. Estrutura curricular e plano de estudos do mestrado em Engenharia de Inteligência Artificial, lecionado no Instituto Superior de Engenharia do Porto, despacho nº 6374/2020 do Diário da República nº 115, Série II de 16 de junho de 2020.

[6]. Estrutura curricular e do plano de estudos do mestrado em Engenharia de Sistemas Computacionais Críticos, lecionado no Instituto Superior de Engenharia, despacho nº 6942/2020 do Diário da República n.º129, Série II de 6 de julho de 2020.

[7]. The Bologna Process and the European Higher Education Area|Educação e formação (europa.eu), https://ec.europa.eu/education/policies/higher-education/bologna-process-and-european-higher-education-area_pt, retirado em 8 de abril 2021.

[8]. Alteração ao regime jurídico dos graus e diplomas do ensino superior, Decreto-Lei n.º 65/2018 do Diário da República n.º 157, 1.ª Série, de 16 de agosto 2018.

[9]. Página do portal ISEP: https://www.isep.ipp.pt/

[10]. Página do RStudio: https://www.rstudio.com/

[11]. Plano de Estudos, Mestrado Integrado em Engenharia Informática, Universidade Nova de Lisboa, https://guia.unl.pt/pt/2020/fct/program/935#structure

[12]. Plano de Estudos, Mestrado Integrado em Engenharia Informática e Computação, FEUP, https://sigarra.up.pt/feup/pt/cur_geral.cur_planos_estudos_view?pv_plano_id=2496&pv_ano_lectivo=2012&pv_tipo_cur_sigla=MI&pv_origem=CUR

[13]. Plano de Estudos, Mestrado Integrado em Engenharia Informática, Universidade do Minho, Departamento de Informática), https://miei.di.uminho.pt/plano_estudos.html

[14]. Plano de Estudos, Mestrado Bolonha em Engenharia Informática e de Computadores – Alameda, IST - https://fenix.tecnico.ulisboa.pt/cursos/meic-a/paginas-de-disciplinas

[15]. Plano de Estudos, Mestrado Bolonha em Engenharia e Ciência de Dados, IST, https://fenix.tecnico.ulisboa.pt/cursos/mecd

[16]. Plano de Estudos, Mestrado em Engenharia e Ciência de Dados, Universidade de Coimbra, https://apps.uc.pt/courses/PT/programme/8521/2021-2022?id_branch=20361

[17]. Plano de Estudos, Mestrado em Engenharia Informática, Universidade de Coimbra, https://apps.uc.pt/courses/PT/unit/88863/20326/2020-2021

[18]. Plano de Estudos, Mestrado em Ciência de Dados (Data Science), Faculdade de Ciências da UP, https://sigarra.up.pt/fcup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia_id=454421

[19]. Plano de Estudos, Pós graduação Estatística Computacional e Análise de Dados, Faculdade de Ciências da UP, https://sigarra.up.pt/fcup/pt/cur_geral.cur_planos_estudos_view?pv_plano_id=27481&pv_ano_lectivo=2020&pv_tipo_cur_sigla=

[20]. Plano de Estudos, Licenciatura em Engenharia Informática, Universidade de Aveiro, https://www.ua.pt/pt/c/383/p

[21]. Plano de Estudos, Mestrado em Engenharia Informática, Universidade de Aveiro, Universidade de Aveiro (ua.pt)

[22]. Plano de Estudos, Lic. em Engenharia Informática, UTAD, https://www.utad.pt/estudar/cursos/engenharia-informatica/

[23]. Licenciatura em Engenharia Informática, UTAD, https://side.utad.pt/cursos/einformatica/disciplinas/paginas/

[24]. Licenciatura em Matemática Aplicada e Ciência de Dados, UTAD, https://side.utad.pt/cursos/lmatcd/disciplinas/paginas/

[25]. Licenciatura em Engenharia Informática, ISEC, https://www.ipc.pt/ipc/oferta-formativa/licenciatura-em-engenharia-informatica-2/

[26]. Mestrado em Engenharia Informática - Tecnologias da Informação e do Conhecimento, ISEC, https://www.isec.pt/PT/estudar/mestrados/MestInfSist/#lnkPlanoCurricular

[27]. Licenciatura em Engenharia Informática e de Computadores, ISEL, https://www.isel.pt/media/uploads/LEICPlano%20Curricular.pdf

[28]. Mestrado em Engenharia Informática e de Computadores, ISEL, https://www.isel.pt/cursos/mestrados/engenharia-informatica-e-de-computadores/plano-curricular

[29]. Texas A&M University, USA, https://mays.tamu.edu/ms-analytics/curriculum-overview/

[30]. Oxford University, UK, https://www.ox.ac.uk/admissions/undergraduate/courses-listing/computer-science

[31]. Universidad Autónoma de Madrid, http://www.uam.es/UAM/ConoceGradosUAM/1446772206517.htm?language=es&nodepath=Conoce%20los%20grados%20de%20la%20UAM&pid=1446772206517

[32]. Complutense de Madrid, https://informatica.ucm.es/listado-de-asignaturas-2020-2021#GII

[33]. Licence Informatique Université Sorbornne, Paris, http://odf.univ-paris13.fr/fr/offre-de-formation/feuilleter-le-catalogue-1/sciences-technologies-sante-STS/licence-lmd-XA/licence-informatique-program-gl4inf-116-2-2.html

[34]. M.Sc. in Data Engineering, Jacobs University, Bremen, Germany , https://www.masterstudies.com/M.Sc.-in-Data-Engineering/Germany/Jacobs-University/

[35]. Master Data Science (online), IU International University of Applied Sciences, Germany, https://www.onlinestudies.com/Master-Data-Science-(MSc)/Germany/IUOnline/

[36]. Master Program in Data Science Methodology, Barcelona Graduate School of Economics, https://www.masterstudies.com/Master-Program-in-Data-Science-Methodology/Spain/Barcelona-GSE/

[37]. Masters Programs in Data Science in Europe 2021, https://www.masterstudies.com/Masters-Degree/Data-Science/Europe/

[38]. A Computing Curricula Series Report 2020 (CC2020), Paradigms for Global Computing Education encompassing undergraduate programs in Computer Engineering, Computer

Science, Cybersecurity, Information Systems, Information Technology, Software Engineering with data science, Association for Computing Machinery (ACM) &IEEE Computer Society (IEEE-CS), 2020 December 31, ISBN: 978-1-4503-9059-0, DOI:10.1145/3456302, https://dl.acm.org/citation.cfm?id=3456302 (retirado em 1 de junho de 2021).

Anexo A – Enunciados das Fichas TP's 2019/2020

- Ficha Teórico-Prática 1 – Estatística Descritiva
- Ficha Teórico-Prática 2 – Testes de Hipóteses Paramétricos
- Ficha Teórico-Prática 3 – Testes de Hipóteses Não-Paramétricos
- Ficha Teórico-Prática 4 – Regressão Linear
- Ficha Teórico-Prática 5 – Regressão: Arvores de Regressão
- Ficha Teórico-Prática 6 – Classificação: Arvores de Decisão
- Ficha Teórico-Prática 7 – Classificação: Redes Neuronais
- Ficha Teórico-Prática 8 – Classificação: K-vizinhos mais próximos

Licenciatura em Engenharia Informática – DEI/ISEP
Análise de Dados em Informática 2019/2020

Ficha Teórico-Prática 1

Estatística Descritiva

Objetivos:

- Familiarização com a ferramenta R no suporte à Análise Exploratória de Dados;
- Breve revisão de Estatística Descritiva;
- Análise e discussão dos resultados.

1. Numa aula prática laboratorial de Algoritmia e Programação, o docente decidiu realizar um estudo do desempenho dos alunos, no sentido de avaliar qual o tipo de erro mais realizado. Para tal, sugeriu aos alunos a codificação de um dado algoritmo em C++. De seguida, pediu-lhes que compilassem o programa e analisassem o nº de erros léxicos, sintáticos e semânticos cometidos.

Aluno	Erros Léxicos	Erros Sintáticos	Erros Semânticos
1	2	5	1
2	3	2	0
3	0	1	0
4	0	0	0
5	3	2	1
6	2	4	1
7	1	5	0
8	2	6	0
9	1	3	1
10	2	6	0
11	2	4	1
12	3	7	1
13	4	12	1

   a) Construa um gráfico com os diagramas de extremos e quartis (box plot) que nos permita analisar o comportamento dos alunos pelo nº de erros cometidos de cada tipo. Qual o tipo de erro mais cometido? Analise o gráfico, referindo a concentração e a dispersão dos dados.

   b) Construa as tabelas de frequências para cada tipo de erro. Da análise da tabela, indique o valor mediano de cada tipo de erro. Qual é o número de erros mais comuns em cada tipo de erro?

   c) Determine a média, o desvio padrão, o mínimo e o máximo para cada tipo de erro. Com base nestas medidaS, o que pode afirmar sobre os dados?

   d) Construa um gráfico que permita visualizar a forma da distribuição de frequências da amostra. O que pode observar no gráfico?

2. Uma empresa do ramo químico produz diferentes tipos de produtos, mas somente pode fabricar um de cada vez. O gestor de produção tem que decidir quanto ao problema de sequenciamento

de 5 tarefas numa única máquina. Para cada tarefa j (j=1, ..., 5), seja pj o tempo de processamento, dj a data de entrega e wj a penalização no caso da tarefa j se atrasar. O objetivo consiste em encontrar uma sequência que minimize a soma dos atrasos pesados 1 || ΣwjTj . O número de soluções admissíveis para este problema corresponde a 5! =120.

Instâncias	Óptimo	ACO	PSO	ABC
Wa40_1	913	1003	913	913
Wa40_2	1225	1781	1225	1225
Wa40_3	537	1557	537	537
Wa40_4	2094	3020	2094	2094
Wa40_5	990	1180	990	990
Wa40_6	6955	17008	7614	6955
Wa40_7	6324	10116	6324	6324
Wa40_8	6865	10853	7048	6865
Wa40_9	16225	23467	16289	16225
Wa40_10	9737	16800	9741	9737
Wa40_11	17465	38563	20334	17465
Wa40_12	19312	33729	19961	19312
Wa40_13	29256	49262	30812	29256
Wa40_14	14377	28531	14497	14377
Wa40_15	26914	46192	27718	26914
Wa40_16	72317	112463	76364	72317
Wa40_17	78623	102721	83190	78623
Wa40_18	74310	109266	77302	74310
Wa40_19	77122	110673	83191	77122
Wa40_20	63229	91617	67679	63229
Wa40_21	77774	102521	83164	77774
Wa40_22	100484	130534	106972	100484
Wa40_23	135618	177417	136910	135618
Wa40_24	119947	159194	126935	119947
Wa40_25	128747	163745	130907	128747
…	…	…	…	…

Foram consideradas para o estudo computacional instâncias do problema de sequenciamento de Máquina Única (WT - “Weighted Tardiness”) com 40 tarefas, 50 e 100 tarefas. Foram retirados os resultados das diferentes MetaHeurísticas (MH), ACO, PSO e ABC, em apenas uma corrida, pretendendo-se analisar o seu desempenho. A tabela descreve um excerto dos dados obtidos para as instâncias com 40 tarefas.

a) Determine os erros relativos para cada instância das várias MH e construa tabelas de frequências para cada uma delas.

b) Compare graficamente o desempenho das MH na resolução das instâncias em estudo. É possível identificar uma técnica que seja a mais eficaz?

c) Calcule um sumário de estatísticas para os erros relativos de cada MH. Compare os resultados obtidos com os resultados das alíneas anteriores.

Exercício Complementar

3. Considere os ficheiro "teeab.csv" contendo os registos dos tempos de entrega de 120 encomendas, em unidades de tempo (u.t.) dois operadores A e B.

   a) Construa a tabela de frequências para os tempos de entrega para cada um dos operadores A e B.

   b) Represente os dados usando gráficos adequados.

   c) Calcule a média, a mediana e os restantes quartis do tempo de entrega por encomenda para os operadores A e B.

   d) Calcule o mínimo, o máximo, o intervalo interquartil e o desvio padrão do tempo de entrega por encomenda para os operadores A e B.

   e) Caracterize, quanto à simetria e achatamento, os dados referentes a ambos os fornecedores.

   f) Baseando a sua resposta nos resultados das alíneas anteriores, caraterize e compare a distribuição dos dados referentes a ambos os operadores quanto à localização, dispersão e forma. Comente.

Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática 2019/2020

Ficha Teórico-Prática 2

Testes de Hipóteses

Objetivos:

- Testar o valor hipotético de um parâmetro
- Familiarização com a ferramenta R no suporte aos testes de hipóteses paramétricos;
- Análise e discussão de resultados.

1. Escolheram-se aleatoriamente 15 computadores portáteis de uma determinada marca, e obtiveram-se as seguintes medidas para as suas espessuras (em mm):

30	30	30	30	31	32	32	32	32	33	33
34	34	34	35							

Considerando que a espessura do computador é uma variável aleatória normal, teste a hipótese H0: $\mu = 32,5$ contra H1:$\mu \neq 32,5$ (admita que $\alpha = 0,05$).

2. Uma loja online indicou no seu website que a entrega é realizada, em média, em 5 dias. Um cliente habitual efetuou uma reclamação, afirmando que o tempo médio de entrega foi superior ao valor indicado pela loja. Para averiguar se o cliente tem razão, foram analisadas aleatoriamente 36 compras efetuadas no respetivo website, tendo-se registado os tempos de entrega, em dias, abaixo indicados:

5	4	4	5	5	5	6	5	4	4	3	4	4	5	5	7	6	5
6	4	6	5	5	6	6	6	4	4	5	5	5	3	6	3	6	5

Considerando um nível de significância de 1%:

a) Formule a hipótese nula e a respetiva hipótese alternativa para o problema. Trata-se de um teste unilateral ou bilateral?

b) Indique o valor observado da estatística teste.

c) O que se pode concluir relativamente à reclamação do cliente?

3. Numa empresa, o departamento de controlo da qualidade quer efetuar alguns testes sobre o peso de um determinado modelo de portátil, que, segundo as especificações de fabrico, é de 2,5 kg. Com o objetivo de verificar se o peso efetivo de cada portátil ultrapassa o indicado nas suas especificações foram selecionados, aleatoriamente, 16 portáteis do referido modelo. Os pesos, em gramas são os indicados na tabela seguinte.

2550	2550	2450	2560	2520	2530	2530	2500
2490	2510	2520	2520	2530	2510	2550	2550

Considerando que o peso dos computadores segue uma distribuição normal:

a) Formule a hipótese nula e a respetiva hipótese alternativa para o problema. Trata-se de um teste unilateral ou bilateral?

b) Para um nível de significância de 10%, indique o valor observado da estatística teste. É possível concluir que o peso de cada portátil indicado nas especificações de fabrico é, de facto, o peso correto?

4. Registaram-se as velocidades máximas (em dpi) de 24 ratos de computadores de uma determinada marca, divididos em 2 grupos: com e sem fio. Os resultados foram:

Com fio	2300	2000	1800	2000	2400	2200	2000	1800	1900	2100	2200	2400
Sem fio	2400	2200	1800	1900	1800	1900	2100	2050	2200	2000	1900	2000

a) Indique se as amostras são independentes ou emparelhadas.

b) Para um nível de significância de 1%, assumindo que a velocidade é uma variável aleatória normal, teste a hipótese de igualdade das velocidades médias, nos ratos com e sem fio.

5. Considera-se que os indivíduos canhotos têm mais força na mão esquerda do que na mão direita. Para testar esta hipótese foram registadas as forças (em N), da mão direita e da mão esquerda, de 6 pessoas canhotas:

Força (em N)	Pessoa					
	1	2	3	4	5	6
mão esquerda	140	90	125	130	95	121
mão direita	138	89	126	128	92	122

a) Indique se as amostras são independentes ou emparelhadas.

b) Assumindo a normalidade dos dados, diga se estes apoiam a hipótese, com um nível de significância de 5%.

6. O diretor de um hotel resolveu investir numas obras durante o ano de 2014, com o objetivo de modernizar o empreendimento, melhorar a qualidade dos seus equipamentos, e assim atrair um maior número de clientes. Para verificar se esta remodelação teve um efeito positivo, foram registadas as taxas mensais de ocupação, em %, ? em 2013 (antes das obras) e em 2015 (depois das obras). Os dados obtidos são apresentados na tabela seguinte.

Mês / **Taxas médias de ocupação, em %**	**Jan**	**Fev**	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**	**Ago**	**Set**	**Out**	**Nov**	**Dez**
Antes das obras (2013)	20	35	40	55	60	75	95	100	90	80	45	25
Depois das obras (2015)	25	30	45	75	80	100	100	100	100	85	65	30

Admitindo que os dados da amostra provêm de uma distribuição normal:

a) Indique se as amostras são independentes ou emparelhadas.

b) Formule a hipótese nula e a respetiva hipótese alternativa para o problema. Trata-se de um teste unilateral ou bilateral?

c) Usando um nível de significância de 0,05, indique o p-valor e interprete.

d) Que podemos concluir acerca do investimento realizado?

7. Um engenheiro informático pretende averiguar se existe uma diferença significativa entre a duração das baterias de computadores com as mesmas características, mas de marcas distintas. Para tal, 12 computadores da marca A e 12 computadores da marca B foram selecionados aleatoriamente e registou-se para cada um a duração da bateria, em horas, como podemos observar na tabela abaixo.

Marca A	6,3	5,2	6,0	6,1	6,5	5,6	5,8	6,0	5,9	5,8	5,9	6,2
Marca B	4,8	6,7	7,1	5,0	6,2	6,1	6,0	5,9	7,0	4,5	5,3	6,2

Com um grau de confiança de 90% e assumindo que a duração da bateria é uma variável aleatória normal, o que podemos concluir?

8. No website de uma universidade, é publicada a seguinte notícia: "mais de 85% dos estudantes usam o Windows como sistema operativo". Para averiguar a veracidade desta notícia, 200 estudantes foram escolhidos aleatoriamente e registou-se o sistema operativo usado. Os dados obtidos encontram-se no ficheiro Ex 8_dados_TP6e7_TH.xlsx.

a) Formule a hipótese nula e a respetiva hipótese alternativa para o problema. Trata-se de um teste unilateral ou bilateral?

b) Para um nível de significância de 5%, que podemos concluir acerca da notícia publicada?

9. Pretende-se realizar um estudo sobre o uso das redes sociais pelos jovens portugueses entre os 20 e os 25 anos. Para tal, foram escolhidos aleatoriamente 150 jovens e para cada um registou-se: o sexo, o tempo diário dispensado nas redes sociais em horas e a rede social usada (Facebook, Twitter, Instagram, Linkedin). Os dados recolhidos encontram-se no ficheiro Ex 9_dados_TP6e7_TH.xlsx.

Considerando um grau de confiança de 95%:

a) Construa um intervalo de confiança para o tempo médio diário passado nas redes sociais pelos jovens do sexo masculino.

b) Verifique se os tempos diários médios passados nas redes sociais diferem significativamente entre os homens e as mulheres.

c) Podemos concluir que mais de metade dos jovens utiliza o Facebook?

10. Uma empresa do ramo químico produz diferentes tipos de produtos, mas somente pode fabricar um de cada vez. O gestor de produção tem que decidir quanto ao problema de sequenciamento de 5 tarefas numa única máquina. Para cada tarefa j (j=1, ..., 5), seja pj o tempo de processamento, dj a data de entrega e wj a penalização no caso da tarefa j se atrasar. O objetivo consiste em encontrar uma sequência que minimize a soma dos atrasos pesados 1 || ΣwjTj . O número de soluções admissíveis para este problema corresponde a 5! =120.

Instâncias	Óptimo	ACO	PSO	ABC
Wa40_1	913	1003	913	913
Wa40_2	1225	1781	1225	1225
Wa40_3	537	1557	537	537
Wa40_4	2094	3020	2094	2094
Wa40_5	990	1180	990	990
Wa40_6	6955	17008	7614	6955
Wa40_7	6324	10116	6324	6324
Wa40_8	6865	10853	7048	6865
Wa40_9	16225	23467	16289	16225
Wa40_10	9737	16800	9741	9737
Wa40_11	17465	38563	20334	17465
Wa40_12	19312	33729	19961	19312
Wa40_13	29256	49262	30812	29256
Wa40_14	14377	28531	14497	14377
Wa40_15	26914	46192	27718	26914

Foram consideradas para o estudo computacional instâncias do problema de sequenciamento de Máquina Única (WT - "Weighted Tardiness") com 40, 50 e 100 tarefas. Foram retirados os resultados das diferentes MetaHeurísticas (MH), ACO, PSO e ABC, em apenas uma corrida, pretendendo-se analisar o seu desempenho. A tabela descreve um excerto dos dados obtidos para as instâncias com 40 tarefas (Ficheiro de Dados SingleMachine-Técnicas de Otimização.csv da ficha TP1).

a) Construa um intervalo de confiança para o desempenho médio da Meta-Heurística PSO.
b) Verifique se os desempenhos das MH, PSO e ABC diferem significativamente entre si.

11. Os sistemas eletrónicos, tais como as redes de computadores, os sistemas de gestão e os sistemas de segurança são ferramentas essenciais para a garantia da continuidade e eficiência dos processos de negócio e fazer com que as organizações funcionem de forma eficaz. Um técnico de informática pretende analisar o desempenho de uma fonte de alimentação ininterrupta, também conhecida pelo acrónimo UPS (sigla em inglês de Uninterruptible Power Supply). Uma UPS é um sistema de alimentação secundário de energia elétrica que fornece energia de emergência e permite a estabilização da alimentação elétrica (reduzindo os efeitos dos picos de tensão) aos equipamentos eletrónicos a este ligado. Para tal foi medida a tensão (volts) à saída da fonte de alimentação em vários momentos ao longo do dia, tendo-se registado os valores constantes da tabela a baixo. Considera-se que a fonte está em perfeitas condições se a tensão for à saída próxima de 220 volts e a variabilidade for de $\pm$10 volts.

214	220	228	245	226	229	233	218	230	239	212	220	225	250	240	219	244	230
237	241	234	214	225	242	235	236	233	218	256	225	223	231	236	241	228	244

a) Construa um intervalo de confiança a 95% para a tensão média à saída da UPS.
b) Teste a hipótese de a tensão média ser 220 volts (determine o p-valor associado). Compare com o resultado obtido na alínea anterior.
c) A fonte encontra-se em perfeitas condições de funcionamento?
d) Assumindo uma distribuição normal para a tensão, determine um intervalo de confiança a 95% para o desvio-padrão. O que conclui?

12. Na tabela seguinte encontram-se os quocientes entre o custo final e o custo inicialmente previsto dos projetos de I&D realizados em 4 grandes empresas.

Empresa	**Custo final/Custo previsto**					
A	1.0	0.8	1.9	1.1	2.7	
B	1.7	2.5	3.0	2.2	3.7	1.9
C	1.0	1.3	3.2	1.4	1.3	2.0
D	3.8	2.8	1.9	3.0	2.5	

Admitindo a normalidade dos dados, pretende-se investigar se o fator "Empresa" tem efeito sobre o agravamento dos projetos. Considere um nível de significância ($\alpha = 0.05$).

Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática 2019/2020

Ficha Teórico-Prática 3

Testes de Hipóteses não Paramétricos

Objetivos:

- Familiarização com a ferramenta R no suporte aos testes de hipóteses não paramétricos;
- Análise e discussão de resultados.

1. Um grupo de 38 alunos fez um curso intensivo de informática e foi submetido a duas formas de aprendizagem. No final do curso, cada aluno escolheu o método de ensino da sua preferência, como mostra a tabela seguinte:

	Método 1	Método 2	Total
Frequência	25	13	38
Proporção	0,658	0,342	1

Espera-se que não exista diferença entre os métodos de ensino. Teste a hipótese nula com um nível de significância de 5%.

2. Numa empresa de informática, um novo produto (produto A) foi desenvolvido e o departamento de marketing gostaria de determinar se este terá tanto sucesso como o produto favorito existente (produto B). Para tal, 150 participantes foram selecionados para testar os dois produtos. Cada participante experimentou ambos os produtos numa ordem aleatória e indicou aquele que preferiu. Os resultados deste teste encontram-se no ficheiro *Ex 2_dados_TP9e10_THNP.csv*)

participante	produto
1	B
2	B
3	A
4	A
5	A
6	B

Com um nível de significância de 10%, podemos rejeitar a hipótese de que os dois produtos têm igual popularidade?

3. 50 estudantes de uma universidade, escolhidos aleatoriamente, deram a sua opinião sobre a sua marca preferida de computador, entre "Asus" e "Samsung". Os resultados mostraram que 30 preferem a marca "Samsung". Há evidências suficientes para concluir que "Samsung" é a marca preferida, comparado a com a "Asus"? Considere $\alpha = 0,05$.

4. A tabela em baixo mostra a distribuição (aproximada) do número de acesso diárioS durante uma semana a uma nova aplicação de telemóvel (frequências esperadas). Os gestores da aplicação suspeitam que existe uma diferença de acessos nas semanas de férias. Para tal, foram registados os dados de uma semana de férias em particular (frequências observadas).

	Domingo	Segunda	Terça	Quarta	Quinta	Sexta	Sábado
Frequências observadas	35 000	24 000	27 000	32 000	25 000	36 000	31 000
Frequências esperadas	35 000	24 500	27 300	31 700	24 900	36 000	31 100

Com um nível de significância de 5%, teste a hipótese de que em semanas de férias existe uma distribuição do número de acessos à aplicação diferente do habitual.

5. Um dado, com 6 faces numeradas de 1 a 6, foi lançado 60 vezes e registou-se sucessivamente o número da face voltada para cima. Os resultados podem ser observados no ficheiro Ex 5_dados_TP9e10_THNP.csv:

Lançamento n.º	face voltada para cima
1	1
2	5
3	2
4	4
5	6

...

Com um nível de significância de 5%, podemos afirmar que o dado está viciado? Use o teste de ajustamento do Qui-quadrado.

6. Foram registados os tempos de execução, em segundos, de um determinado algoritmo em 20 computadores selecionados aleatoriamente:

10,9	10,2	14,9	9,4	9,9
11,8	8,9	8,8	11,1	11,7
9,2	13,3	9,8	7,5	9
6,4	9,5	12,4	12	9,1

a) Construa um QQ-plot para a amostra e análise da possibilidade de esta seguir uma distribuição normal.
b) Utilize o teste Kolmogorov-Smirnov para verificar se os dados são provenientes de uma população com distribuição N(10;22). Use α=0,05.
c) Use agora o teste de Shapiro-Wilk para avaliar a normalidade dos dados. Compare e discuta com a alínea anterior.
d) Teste a hipótese de o tempo de execução médio ser 10 segundos.

7. Podemos observar na tabela seguinte o número de polegadas de 60 televisões que estão à venda numa determinada loja (os dados encontram-se no ficheiro Ex 7_dados_TP9e10_THNP.csv):

46	53	58	60	60	49	59	48	46	78
37	58	46	46	47	48	42	50	63	48
62	49	47	36	40	39	61	43	53	42
59	60	52	34	40	36	67	44	40	45
40	56	51	51	35	47	53	49	50	48
39	60	48	48	49	50	44	52	65	50

Estude a hipótese destas medições serem normalmente distribuídas, recorrendo ao teste de normalidade de Lilliefors (Use α=0,05).

8. Escolheram-se aleatoriamente 15 computadores portáteis de uma determinada marca, e obtiveram-se as seguintes medidas para as suas espessuras (em mm):

30	30	30	30	31	32	32	32	32	33	33
34	34	34	35							

Teste a hipótese H0: μ = 32,5 contra H1:μ ≠ 32,5 (admita que α = 0,05).

9. Numa empresa, o departamento de controlo da qualidade quer efetuar alguns testes sobre o peso de um determinado modelo de portátil, cujo peso, segundo as especificações de fabrico, é de 2,5 kg. Com o objetivo de verificar se o peso efetivo de cada portátil ultrapassa o indicado nas suas especificações foram selecionados, aleatoriamente, 16 portáteis do referido modelo. Os pesos, em gramas, são os indicados na tabela seguinte:

2550	2550	2450	2560	2520	2530	2530	2500
2490	2510	2520	2520	2530	2510	2550	2550

Com um nível de significância de 5%, é possível concluir que o peso de cada portátil indicado nas especificações de fabrico é de facto o peso correto?

10. Uma máquina que embala pacotes de arroz foi recentemente calibrada por forma a que o peso de um pacote de arroz fosse normalmente distribuído com a média de 1 quilograma e desvio padrão 5,1 gramas. Recolheu-se uma amostra aleatória de 10 pacotes de arroz embalados pela máquina e obtiveram-se os seguintes resultados:

1007	990	997	1010,1	1001,5	999	1002,5	1007,1	1010	1010,5

Perante a amostra obtida será possível afirmar que as normas estão a ser respeitadas? (Use α=0.05).

11. Um engenheiro informático desconfia que o tempo de execução de um determinado algoritmo ultrapassa os 25 segundos. Para tal, oito computadores foram selecionados aleatoriamente e registou-se os tempos, em segundos, para executar o algoritmo:

25,36 24,64 25,17 25,37 24,56 24,56 24,80 25,21 25,38 24,55

Verifique se a desconfiança do engenheiro é válida (use α=0,05).

12. Um engenheiro informático instalou um novo dispositivo eletrónico em 6 computadores selecionados aleatoriamente e registou os tempos de arranque, em segundos, antes e depois da instalação do dispositivo. Os resultados podem ser observados na tabela abaixo:

	Computador					
	1	**2**	**3**	**4**	**5**	**6**
Tempo de arranque antes da instalação do dispositivo (s)	14	9,0	12,5	13	9,5	12,1
Tempo de arranque depois da instalação do dispositivo (s)	13,8	8,9	12,6	12,8	9,2	14,2

Considerando um nível de significância de 5% verifique se a instalação do novo dispositivo permite reduzir o tempo de arranque do computador, usando, se possível:

a) o Teste do Sinal;
b) o Teste de Wilcoxon;
c) um teste paramétrico adequado.

13. Um veterinário desconfia que a ração seca para cães de marca B leva a um aumento de peso, comparado com a ração seca de marca A. Foram então selecionados aleatoriamente 8 cães de raças distintas, alimentados durante 6 meses com a ração A e os outros seis meses com a ração B. Os seus pesos, em quilogramas, foram registados no final de cada fase, como podemos observar na tabela seguinte:

Cão	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Peso medido após 6 meses de alimentação com a ração de **marca A** (kg)	31,2	26,5	24,1	10,2	25,3	12,1	30,3	39,2
Peso medido após 6 meses de alimentação com a ração de **marca B** (kg)	35,8	21,3	15,8	11,1	28,5	10,3	31,6	25,4

O que podemos concluir acerca da desconfiança do veterinário? Considere um nível de significância $\alpha = 0{,}05$. Utilize, se possível:

a) o Teste do Sinal;
b) o Teste de Wilcoxon;
c) um teste paramétrico adequado.

14. Numa avaliação de desempenho de computadores, doze utilizadores de nível avançado na área da informática foram selecionados aleatoriamente para avaliar três computadores com características iguais mas de marcas distintas (A, B e C). O objetivo do estudo é verificar se a marca do computador influencia a avaliação do utilizador. Na tabela abaixo, podemos observar as classificações de cada computador, segundo cada utilizador, numa escala de 1 a 10 (os dados encontram-se no ficheiro *Ex 3_dados_TP9e10_THNP_Parte2.csv*):

Marca	**Utilizador**	**Resposta**
A	1	7
A	2	6
A	3	6
A	4	7
A	5	7
A	6	8

Com um nível de significância de 5%, verifique se a marca do computador influencia a avaliação do utilizador.

15. Num teste de usabilidade, foram comparadas duas páginas web, webA e webB. Um conjunto de 12 participantes foi divido em dois grupos e a cada um dos grupos foi pedido que realizasse um teste de usabilidade a uma das páginas. Na apreciação global, medida numa escala 1-10, obtiveram-se os seguintes resultados:

webA	3	4	2	6	2	5
webB	9	7	5	10	6	8

Com um nível de significância de 5%, utilize um teste não paramétrico adequado para analisar a hipótese de que a página *webB* se encontra mais adequada às exigências dos utilizadores.

16. O ficheiro Processadores.txt contém um conjunto de medições (devidamente controladas) de velocidade (GHz) de 2 processadores com as mesmas características, mas de marcas diferentes. Use um procedimento adequado para investigar se existem diferenças entre as velocidades dos dois processadores (use um nível de significância de 5%).

17. A empresa criadora de uma determinada página web, decidiu realizar um teste de usabilidade à página recentemente desenvolvida para um serviço de *streaming*. Para isso selecionou um grupo de participantes de várias idades agrupadas da seguinte forma: 15 – 30 anos, 31 – 45 anos e 46 – 60 anos. Após a execução das tarefas, os participantes responderam a um questionário, em que, em particular, se pretendia analisar o grau de facilidade em alterar os dados de perfil de utilizador. A tabela seguinte apresenta os resultados recolhidos com base numa escala de Likert com 5 níveis.

15 – 30 anos	4	5	3	2	5	4	4	5	
31 – 45 anos	3	4	3	3	5	3	4	4	
46 – 60 anos	2	3	4	2	3	5	4	3	2

Para um nível de significância de 5%, poderá concluir-se que existem diferenças na perceção de usabilidade da funcionalidade referida em função da idade dos participantes?

18. Observe a tabela seguinte:

TEMPO DE RESPOSTA DE UM MONITOR (ms)		
Marca e modelo	Anunciado	Medido
A	8	4.8
B	12	5
C	16	17,5
D	8	6,4
E	8	6,7
F	8	4,3
G	27	20,3
H	12	8
I	8	3,5
J	16	6,3

Utilizando o método de Kendall, verifique se existe alguma relação entre as duas variáveis (use um nível de significância de 5%).

19. Pretende-se verificar se o número de dispositivos informáticos usados (computador, tablet e telemóvel) depende do nível de escolaridade. Para tal, 250 pessoas foram escolhidas aleatoriamente e os seguintes resultados foram obtidos:

	Número de dispositivos informáticos			
	1	**2**	**3**	**+ de 3**
Básico	8	13	9	10
Secundário	25	30	12	8
Universitário	15	27	50	43

Use o teste do Qui-Quadrado para verificar a dependência entre as duas variáveis consideradas (use um nível de significância de 5%).

20. Numa determinada instituição de ensino, as classificações de Estatística são categorizadas em 6 níveis: Excelente, Muito Bom, Bom, Suficiente, Insuficiente e Mau. Já no caso do Cálculo, as classificações podem ser A, B, C, D, E e F, por ordem decrescente de valor. No ficheiro Notas encontram-se os registos das classificações obtidas por 25 alunos. Pretende-se estudar se a nota obtida a Cálculo está associada positivamente com a classificação obtida em Estatística. Use um nível de significância de 5% e a correlação de Spearman para analisar o problema.

21. Um engenheiro informático, responsável pela gestão de um conjunto de servidores, pretende analisar se existe uma correlação entre as falhas de conetividade e o número de acessos diários. Para tal, gerou um script para registar diariamente e durante o período de 1 mês, o número de falhas e o número de acessos a um determinado servidor. Os valores obtidos encontram-se no ficheiro Servidor.
    - **a)** Construa um diagrama de dispersão dos dados e verifique a existência de uma associação entre as duas variáveis.
    - **b)** Calcule um coeficiente de correlação apropriado e conclua, a um nível de significância de 5%, se existe uma correlação positiva entre as duas variáveis.

Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática 2019/2020

Ficha Teórico-Prática 4

Regressão Linear

Objetivos:

- Familiarização com a ferramenta R no suporte a problemas relativos Regressão Linear Simples;
- Análise e discussão de resultados.

1. Considere os seguintes valores observados das variáveis X e Y:

xi	yi
21	185,79
24	214,47
32	288,03
47	424,84
50	454,58
59	539,03
68	621,55
74	675,06
62	562,03
50	452,93
41	369,95
30	273,98

a) Usando um diagrama de dispersão, verifique se existe uma relação linear entre as duas variáveis.

b) Estime os parâmetros da reta de regressão e *y*(40).

c) Calcule o coeficiente de correlação linear de Pearson e comente os resultados.

d) Determine se os pressupostos relativos aos resíduos se verificam.

2. Os seguintes dados mostram o volume de produção de trigo, em milhares de toneladas, numa dada região entre 1986 e 1994.

Ano	Volume de produção
1986	285
1987	270
1988	294
1989	279
1990	260
1991	262
1992	258
1993	272
1994	255

a) Construa o diagrama de dispersão e acrescente a reta de regressão correspondente. Comente os resultados.

b) Calcule os coeficientes de determinação e de correlação. Comente os resultados.

c) Usando um teste de Durbin-Watson, verifique a independência dos resíduos.

3. Na seguinte tabela apresentam-se os montantes dos seguros de vida e os rendimentos anuais, em milhares de unidades monetárias (u.m.), de 12 agregados familiares de certo país.

Rendimento anual (milhares de u.m.)	Capital seguro (milhares de u.m.)
14	31
19	40
23	49
12	20
9	21
15	34
22	54
25	52
15	28
10	21
12	24
16	34

a) Construa o diagrama de dispersão para estes dados e adicione a reta de regressão. Comente os resultados.
b) Estime o montante do seguro de vida para um agregado familiar com rendimento anual de 20000 u.m.
c) Verifique se os resíduos gozam de homocedasticidade e se são independentes.
d) Teste a normalidade dos resíduos.

4. Um engenheiro mecânico pretende analisar o acabamento da superfície das peças de metal produzidas num torno e suspeita que este está dependente da velocidade (em rotações por minuto) do torno e do tipo de ferramenta de corte usada. O ficheiro "**ExemploMontegomery-12-11.csv**" contém os dados da amostra recolhida.
a) Apresente um modelo de regressão linear adequado e interprete os seus coeficientes.
b) Estime os parâmetros da reta de regressão.
c) Calcule o coeficiente de determinação ajustado e comente os resultados.
d) Determine se os pressupostos relativos aos resíduos se verificam.
e) Verifique se existe multicolinearidade.

Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática 2019/2020

Ficha Teórico-Prática 5

Modelos de regressão lineares. Árvores de regressão.

Objetivos:

- Modelos de regressão linear simples e múltipla, usando R;
- Modelos de árvores de regressão, usando R;
- Avaliação dos modelos.

Pretende-se avaliar o impacto que o orçamento, em publicidade, em três canais (youtube, facebook e jornal) têm sobre as vendas de uma empresa. Os dados disponíveis são o orçamento em publicidade em milhares de dólares e o montante das vendas. A publicidade em cada um dos canais foi repetida 200 vezes com diferentes orçamentos e as vendas observadas foram recolhidas. O objetivo é prever as vendas futuras da empresa usando modelos de regressão lineares e árvores de regressão.

1. Comece por carregar o *dataset* " marketing" do package = "datarium".
2. Analise os dados.
3. Separe o conjunto de dados inicial em dois subconjuntos treino e teste, segundo o critério holdout (70% treino/30% teste).
4. Obtenha um modelo de regressão linear simples usando apenas um dos canais de publicidade.
   a) Apresente a função linear resultante.
   b) Visualize a reta correspondente ao modelo de regressão linear simples e o respetivo diagrama de dispersão.
   c) Calcule o erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE) do modelo sobre os 30% de casos de teste.
5. Repita as alíneas anteriores, com um modelo de regressão linear múltipla usando os três canais.
6. Simplifique o modelo.
7. Obtenha a árvore de regressão usando a função *rpart* para prever as vendas futuras da empresa em função dos orçamentos em publicidade nos três canais.
8. Visualize a árvore de regressão.
9. Calcule o erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE) da árvore de regressão sobre o conjunto de teste.

Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática 2019/2020

Ficha Teórico-Prática 6

Classificação: Árvores de Decisão.

Objetivos:

- Modelos de árvores de decisão, usando R;
- Avaliação dos modelos.

O conjunto de dados "BreastCancer" a analisar, contém atributos que foram obtidos a partir de imagens digitalizadas de pequenas amostras de massa mamária de pacientes e descrevem as características dos núcleos celulares presentes nessas imagens. Os atributos deste conjunto de dados são os seguintes:

Id	Identifier
Cl.thickness	Clump Thickness
Cell.size	Uniformity of cell size
Cell.shape	Uniformity of cell shape
Marg.adhesion	Marginal adhesion
Epith.c.size	Single Epithelial cell size
Bare.nuclei	Bare Nuclei
Bl.cromatin	Bland Chromatin
Normal.nucleoli	Normal Nucleoli
Mitoses	Mitoses
Class	Class (benign/malignant)

Pretende-se determinar a qual das duas classes (benigna ou maligna) o tumor pertence.

1. Comece por carregar o dataset "BreastCancer" da biblioteca "mlbench" para o ambiente do R. Verifique a sua dimensão e obtenha um sumário dos dados.
2. Usando os gráficos apropriados, analise os vários atributos do conjunto de dados.
3. Separe o conjunto de dados inicial em dois subconjuntos treino e teste, segundo o critério holdout, (70% treino/30% teste), aplique a função "rpart" sobre os dados de treino para gerar um modelo de classificação e visualize a árvore de decisão.
4. Apresente a matriz de confusão e a taxa de acerto do modelo gerado.

5. Repita o processo anterior de aprendizagem/teste 10 vezes (com amostras diferentes em cada repetição) colecionando, em cada iteração, a percentagem de acerto obtida pela respetiva árvore. Apresente o valor médio da percentagem de acerto nas 10 repetições e o respetivo desvio padrão.
6. Elabore uma função para apresentar a matriz de confusão e as medidas de avaliação: taxa de acerto (accuracy), recall, precision e F1 de um modelo.
7. Repita novamente o processo de aprendizagem com a função "rpart" usando agora o método de treino k-fold cross validation e a função anterior para obter as medidas de avaliação de cada modelo.
8. Obtenha o valor médio e o respetivo desvio padrão das medidas obtidas anteriormente.

Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática 2019/2020

Ficha Teórico-Prática 7

Classificação: Redes Neuronais

Objetivos:

- Redes Neuronais;
- Análise e discussão dos resultados.

O betão é um dos materiais mais importantes em engenharia civil. A resistência à compressão do betão é uma função altamente não linear da sua idade e dos seus componentes. O objetivo é desenvolver modelos que prevejam a resistência (*strength*) do betão em função dos seguintes atributos previsores:

cement	cimento (kg / m3)
slag	cinza volante (kg / m3)
ash	escória de alto-forno (kg / m3)
water	água (kg / m3)
superplastic	superplastificante (kg / m3)
coarseagg	agregado grosso (kg / m3)
fineagg	agregado fino (kg / m3)
age	idade do teste (dias)

1. Comece por carregar o ficheiro (”concrete.csv”) para o ambiente do R, verifique a sua dimensão e obtenha um sumário dos dados.
2. Usando os gráficos apropriados explore os vários atributos do conjunto de dados.
3. Verifique se os dados precisam ser normalizados. Em caso positivo elabore uma função para realizar a normalização **min-max** que mapeia os valores das variáveis no intervalo [0-1]:

$$y' = \frac{y - min_y}{max_y - min_y}$$

4. Separe o conjunto de dados em dois subconjuntos treino e teste, segundo o critério holdout (70% treino/30% teste).
5. Treine uma rede neuronal usando a função **neuralnet** para prever a resistência (strength) do betão. Comece por considerar como configuração da rede apenas um nó na camada interna. Visualize a rede e avalie as previsões da rede usando o conjunto de teste.
6. Repita a alínea anterior considerando as seguintes configurações da rede:
   a) 1 nível interno, com 3 nós
   b) 2 níveis internos, com 6 e 2 nós, respetivamente
7. Usando o método de treino **k-fold cross validation** obtenha modelos de previsão de strength com:
   a) uma rede neuronal com a configuração anterior com melhor desempenho
   b) um modelo de regressão linear múltipla

   c) um modelo árvore de regressão obtenha a média e o desvio padrão do RMSE de cada modelo

8. Verifique se a diferença de desempenho entre os dois melhores modelos obtidos anteriormente é estatisticamente significativa.

Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática 2019/2020

Ficha Teórico-Prática 8

Classificação: K-Vizinhos-mais -próximos

Objetivos:

- K-Vizinhos-mais -próximos;
- Análise e discussão dos resultados.

O Abalone é um molusco com uma concha peculiar em forma de orelha forrada a madrepérola. O seu valor económico está positivamente correlacionado com a sua idade, sendo por isso importante determinar a idade com precisão. No entanto, os produtores estimam a idade deste molusco cortando a concha e através de um microscópio contam o número de anéis na concha. Este processo além de demorado é pouco preciso e aumenta o custo do molusco. O objetivo é prever a idade do abalone através de modelos usando as medições físicas do molusco. Para isso considere o seguinte *dataset*:

Sex	nominal	M, F, and I (infant)	
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer		+1.5 gives the age in years

1. Comece por carregar o ficheiro (”abalone.data”) para o ambiente do R, verifique a sua dimensão e obtenha um sumário dos dados.
2. Usando os gráficos apropriados, explore os vários atributos do conjunto de dados e realize as seguintes transformações aos dados:
   a) conversão do atributo categórico **Sex** para numérico
   b) normalização dos dados
3. Separe o conjunto de dados em dois subconjuntos treino e teste, segundo o critério **holdout**, (70%treino/30% teste).
4. Aplique o algoritmo K-vizinhos-mais-próximos para prever o atributo **Rings** usando os valores ímpares de K no intervalo (1..50). Recolha para cada valor de K o RMSE da previsão. Verifique qual o valor de k que minimiza o RMSE. Não se esqueça de voltar a desnormalizar as previsões para obter um RMSE comparável com o valor Rings.
5. Derive um novo atributo **Age** a partir do atributo **Rings** e discretize este novo atributo em duas classes: **Young** e **Adult**.

6. Aplique o algoritmo K-vizinhos-mais-próximos para prever o atributo **Age** usando os valores ímpares de K no intervalo (1..50). Recolha para cada valor de **K** a taxa de acerto (accuracy) da previsão. Verifique qual o valor de **k** que maximiza a taxa de acerto.
7. Usando o método de treino **k-fold cross validation** obtenha modelos de previsão do atributo **Agecom**:
   a) o algoritmo K-vizinhos-mais-próximos e o valor de **k** obtido na alínea anterior
   b) um modelo árvore de regressão e obtenha a média e o desvio padrão taxa de acerto dos modelos
8. Verifique se a diferença de desempenho entre os modelos obtidos anteriormente é estatisticamente significativa.

Anexo B – Enunciados de Trabalhos Práticos

B 1. Enunciados de Trabalhos – Edição 2015-2016

- **B 1.1. Análise de Usabilidade**
- **B 1.2. Análise de Fiabilidade**
- **B 1.3. Análise de Desempenho**

Trabalho Prático

Análise de Dados em Informática

Engenharia Informática - 3º ano 2º semestre
Ano Lectivo 2015/2016

---

1. Objectivos

Objectivo Geral:

- Análise de Usabilidade

Objectivos Específicos:

- Definir plano da sessão de avaliação
- Definir e construir inquérito de satisfação
- Análise e discussão dos Resultados
- Escrita do relatório

2. Calendarização

Lançamento das propostas de trabalhos: até 28 de Fevereiro de 2016
Constituição dos grupos: até 5 de Março de 2016
Entrega do trabalho: até 20 de Março de 2016 (23:55)
Apresentação e discussão: em data a marcar pelo professor de TP

A identificação dos grupos deve ser efectuada junto do professor das aulas Teórico Práticas (TP).

3. Normas

- O trabalho deve ser realizado em grupo (max. 3 alunos), extra aulas. A apresentação e discussão poderão ser realizadas individualmente.

- A **data final de ENTREGA** do trabalho é 20 de Março de 2016. No entanto os grupos terão de cumprir as seguintes fases intermédias:
  - **Até 05.03.16**: Identificação do grupo e descrição informal do trabalho, ao professor de teórico-práticas, por *email*.
  - **Até 20 de Março de 2016(23:55)**: Entrega do trabalho no moodle.

  Independentemente destes prazos, os grupos deverão ser capazes de, quando o professor o solicitar, reportar o estado de desenvolvimento do trabalho.

- A entrega do trabalho consta de um relatório e respectivos anexos. Deverá submeter todos os documentos num ficheiro compactado. O nome do ficheiro deverá seguir a seguinte notação:

  ANADI_XXX_Nºaluno1_Nºaluno2_ Nºaluno3.zip, em que **XXX** representa a turma PL.

  Exemplo: **ANADI_3AD_7777777_8888888_9999999.zip.**
- Trabalhos cujo nome não respeite a notação indicada, **serão penalizados em 10%.**

- A **Entrega do trabalho deverá ser submetida no moodle até à data de entrega definida**. **Não serão aceites trabalhos fora do prazo**.

- A apresentação e defesa trabalho decorrerá em dia e hora a marcar por cada professor das teórico-práticas. No dia da apresentação, **TODOS** os elementos do grupo deverão estar presentes. Os elementos ausentes não terão classificação.

- A avaliação do trabalho será realizada por uma equipa de docentes.

- Cada grupo é responsável por gerir o seu processo de desenvolvimento. Dificuldades e problemas deverão ser comunicados atempadamente ao professor das aulas práticas laboratoriais.

3.1 *Relatório*

No relatório final deverão ser documentadas todas as fases da Análise da Usabilidade, organização do estudo estatístico, análise e discussão dos resultados e conclusões.

3.2 Avaliação

Na avaliação do trabalho serão considerados:

- Guia da sessão da avaliação de usabilidade,
- Inquérito
- A análise exploratória de dados
- A apresentação e discussão
- Participação individual de cada um dos elementos

Definição do problema	**20%**
Guia da sessão da Avaliação de usabilidade	**20%**
Inquérito	**20%**
Análise Exploratória de Dados	**40%**

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua participação. A equipa de avaliação de trabalhos práticos irá validar, no momento da defesa do trabalho, a participação de cada um dos elementos do grupo na concretização dos objectivos do trabalho e do grupo.

4. Descrição do Trabalho

Na realização deste trabalho pretende-se que os alunos desenvolvam o processo de Análise de Usabilidade, com o objetivo de testar e avaliar a usabilidade de sistemas, portais e sites.
Um sistema interativo é constituído por dois elementos principais: a parte computacional, também designada por funcional, e pela parte comunicacional (Hix e Hartson, 1993). A primeira é invisível aos utilizadores e é responsável pelo processamento das ações, pelo espoletar das funções adequadas, e por fornecer ao utilizador o necessário feedback. A segunda, parte comunicacional, é responsável pela comunicação e é frequentemente entendida como sendo o interface com o utilizador, pois é a parte visível. O conceito de usabilidade pretende promover todos os aspetos que facilitem a utilização de um sistema computacional. Seguindo os passos definidos por Hix e Hartson (1993) para avaliar a usabilidade, deve ser desenvolvida uma experiência de utilização do sistema a avaliar e organizar sessões de avaliação com participantes onde procedem à recolha de dados. Posteriormente devem realizar a análise dos dados recolhidos e formular as conclusões.
Pretende-se no âmbito do 1º Trabalho Prático testar e avaliar a usabilidade de sistemas, portais e sites. Sugerem-se os seguintes:

- Site das Finanças: efatura
- Portal do ISEP
- Moodle do ISEP
- Página web do DEI
- Ebay
- Facebook
- Entre outros

5 – Referências Bibliográficas:

NIELSEN, J. - Usability Engineering. AP Professional (1993).
HIX, D. e HARTSON, H. R. - Developing User Interfaces: Ensuring Usability Through Product & Process. John Wiley & Sons, inc (1993).
PIAIRO, EDUARDO, Desenvolvimento de Interface Inteligente para Suporte à Gestão e Controlo de Produção, Dissertação de Mestrado, FEUP, 2012.

Trabalho Prático

Análise de Dados em Informática

Análise de Fiabilidade

Engenharia Informática - 3º ano 2º semestre
Ano Lectivo 2015/2016

---

1. Objectivos

Objectivo Geral:

- Análise de Fiabilidade dos servidores do DEI
- Análise de Fiabilidade da rede wireless EDUROAM no DEI

Objectivos Específicos:

- Definir plano da sessão de avaliação de fiabilidade
- Planeamento e Organização do processo de recolha de dados
- Análise e discussão dos Resultados com recurso ao R
- Escrita do relatório

2. Calendarização

Lançamento das propostas de trabalhos: até 18 de Abril de 2016
Entrega do trabalho: até 16 de Maio de 2016 (23:55)
Defesa e discussão: em data a marcar pelo professor de TP

A identificação dos grupos deve ser efectuada junto do professor das aulas Teórico Práticas (TP).

3. Normas

- O trabalho deve ser realizado em grupo (max. 3 alunos), extra aulas. A defesa e discussão poderão ser realizadas individualmente.
- O grupo deve ser o mesmo em todos os trabalhos práticos.
- Deverá ser usado o R como ferramenta de suporte ao estudo estatístico
- A **data final de ENTREGA** do trabalho é 16 de Maio de 2016, no moodle.
  Independentemente destes prazos, os grupos deverão ser capazes de, quando o professor o solicitar, reportar o estado de desenvolvimento do trabalho.
- A entrega do trabalho consta de um relatório e respectivos anexos. Deverá submeter todos os documentos num ficheiro compactado. O nome do ficheiro deverá seguir a seguinte notação:

  ANADI_YYY_XXX_Nºaluno1_Nºaluno2_ Nºaluno3.zip, onde **YYY** representa a sigla do docente das TP, e **XXX** representa a turma TP.

  Exemplo: **ANADI_EOS_3AD_7777777_8888888_9999999.zip.**

- Trabalhos cujo nome não respeite a notação indicada, **serão penalizados em 10%.**
- A **Entrega do trabalho deverá ser submetida no moodle até à data de entrega definida**. **Não serão aceites trabalhos fora do prazo**.
- A defesa e discussão dos trabalhos decorrerá em dia e hora a marcar por cada professor das teórico-práticas. No dia da apresentação, **TODOS** os elementos do grupo deverão estar presentes. Os elementos ausentes não terão classificação.
- A avaliação do trabalho será realizada por uma equipa de docentes.
- Cada grupo é responsável por gerir o seu processo de desenvolvimento. Dificuldades e problemas deverão ser comunicados atempadamente ao professor das aulas teórico-práticas.

3.1. Relatório

No relatório final deverão ser documentadas todas as fases da Análise de Fiabilidade realizadas, contextualização do tema, recolha dos dados, organização do estudo estatístico, análise e discussão dos resultados, conclusões e anexos.

3.2. Avaliação

Na avaliação do trabalho serão considerados os seguintes aspetos:

- Contextualização

- Organização e Recolha dos dados,
- A análise exploratória de dados com recurso ao R
- A defesa e discussão
- Participação individual de cada um dos elementos

Contextualização	**20%**
Fiabilidade da rede do DEIspace	**40%**
Fiabilidade da rede sem fios EDUROAM no DEI	**30%**
Conclusões	**10%**

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua participação. A equipa de avaliação de trabalhos práticos irá validar, no momento da defesa do trabalho, a participação de cada um dos elementos do grupo na concretização dos objectivos do trabalho e do grupo.

4. Descrição do Trabalho

Na realização deste trabalho pretende-se que os alunos desenvolvam o processo de Análise de Fiabilidade, com o objetivo de avaliar a fiabilidade da rede do DEIspace e da rede sem fios EDUROAM no DEI.

1. Análise de Fiabilidade da rede do DEIspace

Em anexo é fornecida uma lista de falhas (indisponibilidade) de vários servidores do DEIspace, verificados de 5 em 5 minutos, entre o dia 2 de Dezembro de 2014 às 20:00 e o dia 24 de Fevereiro de às 10:00.
O registo de falhas contém a data/hora (pela qual está ordenado) em que a falha foi detetada, o nome do servidor e a duração da falha em minutos. A ausência de registos de falha significa que o servidor esteve disponível.
Com base nos dados disponibilizados:

a) Determinar a disponibilidade:
- total (durante todo o período de verificação) de cada um dos servidores.
- no ano de 2015 do servidor "ipp";
- no mês de Janeiro 2015 do servidor "srv3";
- no dia 14 de Maio 2015 do servidor "sw1700a";

Comente os resultados obtidos.

b) Determine a função de fiabilidade do servidor "sw1700a" no dia 14 de Maio 2015. Apresente e analise o seu gráfico.

c) Se λ_i representa a taxa de falhas no dia i do mês de Dezembro de 2014, do servidor "sw1800a", em que i∈{1,2,…,31}, verifique se a taxa média real de falhas está abaixo dos 0.01 falha/minuto. Opte por um grau de confiança que considerar mais adequado e justifique a sua escolha.

d) Valide se existe uma diferença significativa entre as durações médias de falhas dos servidores "sw1800a" e "srv3" durante o ano de 2015. Use o mesmo grau de confiança que usou na questão 4 e justifique

e) Usando o teste ANOVA, verifique se existe uma diferença significativa entre as durações médias de falhas de todos os servidores durante o período total de verificação. Use o mesmo grau de confiança que usou na questão 4 e justifique.

f) Retire conclusões com base nos resultados das alíneas anteriores

2. Análise de fiabilidade da rede sem fios EDUROAM no DEI

Planeie e organize o processo de recolha e análise de fiabilidade da rede sem fios EDUROAM no DEI. Existem algumas queixas dos utilizadores relativamente à disponibilidade da rede sem fios EDUROAM no ISEP, pretende-se por isso levar a cabo uma recolha de dados automatizada sobre a disponibilidade deste serviço para posterior análise estatística.
Muitas das queixas dos utilizadores referem-se ao processo de acesso e autenticação, por isso esse processo deve ser contemplado na recolha de dados.

a) Desenvolver um sistema de recolha de dados baseado nos seguintes princípios:

- Deverá ser usado um posto de trabalho (portátil) onde será implementado o sistema de recolha que poderá ser implementado através de um pequeno script.
- Deve estar previamente configurado o acesso à rede EDUROAM com as credenciais apropriadas. Para evitar interferências, outras configurações de rede sem fios devem ser retiradas.
- De 10 em 10 minutos a interface de rede sem fios é desativada, seguindo-se o seguinte procedimento.
- Dois minutos após a desativação a interface é novamente ativada.
- Três minutos depois da reativação é testado o acesso à internet e o resultado é guardado num ficheiro de texto juntamente com uma etiqueta data/hora.

b) Realização prática da recolha de dados:

- Cada grupo de trabalho terá de disponibilizar equipamento próprio (portátil) para realização da recolha.
- Uma vez que a rede sem fios EDUROAM suporta canais da banda de 2,4 GHz (802.11g) e também da banda de 5 GHz (802.11a e 802.11n) é importante que seja associado aos dados recolhidos se o equipamento usado suporta ou não a banda de 5 GHz.
- No equipamento referido deverá ser implementado e devidamente testado o referido sistema de recolha de dados.
- Cada grupo deverá recolher informação durante um período contínuo não inferior a 24 horas (1 dia - fins de semana incluídos).
- O DEI disponibiliza o alojamento para os equipamentos dos alunos durante o processo de recolha (Sala B405). Durante o período de recolha, salvo necessidade absoluta não será facultado acesso ao equipamento.
- Será definido um calendário para cada grupo efetuar a recolha de dados, os dias para troca de equipamentos (inicio e fim de períodos de recolha de dados) são todos os dias úteis: SEG; TER; QUA; QUI e SEX.

c) Descreva todas as fases necessárias à respectiva concretização, desde o processo de planeamento e organização para a recolha dos dados até à análise estatística. Organize com base nos resultados obtidos o estudo estatístico. Retire conclusões com base no estudo estatístico realizado. Justifique todas as decisões que tomar.

Os dados recolhidos por cada grupo, deverão ser submetidos no moodle. Depois de um tratamento prévio deverão ser disponibilizados aos restantes grupos para tratamento estatístico. Instruções adicionais serão distribuídas a seu tempo.

5. Referências Bibliográficas

Stapelberg, R. F. (2009). Handbook of Reliability, Availability, Maintainability and Safety in Engineering Design. Springer.

Trabalho Prático

Análise de Dados em Informática

Análise de Desempenho

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2015/2016

---

1. Objetivos

Objetivo Geral:

- Análise de Desempenho de técnicas de otimização na resolução de problemas de escalonamento

Objetivos Específicos:

- Definir plano da Análise de Desempenho
- Análise e discussão dos Resultados com recurso ao R
- Escrita de artigo científico

2. Calendarização

Entrega do trabalho: até 12 de Junho de 2016 (23:55)
Defesa e discussão: em data a marcar pelo professor de TP

3. Normas

- O trabalho deve ser realizado em grupo (max. 3 alunos), extra aulas. A defesa e discussão poderão ser realizadas individualmente.
- O grupo deve ser o mesmo em todos os trabalhos práticos.
- Deverá ser usado o R como ferramenta de suporte ao estudo estatístico
- A **data final de ENTREGA** do trabalho é 12 de Junho de 2016, no moodle.
  Independentemente destes prazos, os grupos deverão ser capazes de, quando o professor o solicitar, reportar o estado de desenvolvimento do trabalho.
- A entrega do trabalho consta de um artigo científico. Deverá submeter todos os documentos num ficheiro compactado. O nome do ficheiro deverá seguir a seguinte notação:

  ANADI_YYY_XXX_Nºaluno1_Nºaluno2_ Nºaluno3.zip, onde **YYY** representa a sigla do docente das TP, e **XXX** representa a turma TP.

  Exemplo: **ANADI_EOS_3AD_7777777_8888888_9999999.zip.**

- Trabalhos cujo nome não respeite a notação indicada, **serão penalizados em 10%.**
- A **Entrega do trabalho deverá ser submetida no moodle até à data de entrega definida**. **Não serão aceites trabalhos fora do prazo**.
- A apresentação, numa aula, **em formato de comunicação (10 minutos)** e discussão dos trabalhos decorrerá em dia e hora a marcar por cada professor das teórico-práticas. No dia da apresentação, **TODOS** os elementos do grupo deverão estar presentes. Os elementos ausentes não terão classificação.
- A avaliação do trabalho será realizada por uma equipa de docentes.
- Cada grupo é responsável por gerir o seu processo de desenvolvimento. Dificuldades e problemas deverão ser comunicados atempadamente ao professor das aulas teórico-práticas.

3.1. *Artigo Científico*

No 3.1.Artigo Científico deverão ser documentadas todas as fases da Análise de Desempenho realizadas, contextualização do tema, organização do estudo estatístico, análise e discussão dos resultados e conclusões.

3.2. Avaliação

Na avaliação do trabalho serão considerados os seguintes aspetos:

- Revisão do estado da arte
- A análise exploratória de dados com recurso ao R

- A apresentação numa aula e discussão
- Participação individual de cada um dos elementos

Contextualização (Abstract, Introdução, estado da arte)	**20%**
Análise de desempenho de técnicas de otimização	**40%**
Análise da influência dos parâmetros no desempenho das técnicas de otimização	**20%**
Conclusões	**10%**
Apresentação e Discussão	**10%**

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua participação. A equipa de avaliação de trabalhos práticos irá validar, no momento da defesa do trabalho, a participação de cada um dos elementos do grupo na concretização dos objetivos do trabalho e do grupo.

4. Descrição do Trabalho

O objetivo deste trabalho é praticar as competências na utilização de Análise Exploratória de Dados (Estatística Descritiva, Análise de Inferência e correlação) na Análise de Desempenho. Deve ser produzido um artigo científico (português ou inglês), conforme *template* indicado, com o estado da arte sobre análise de desempenho de algoritmos, os dados em análise, a análise e discussão dos resultados e conclusões.
O desempenho de técnicas de otimização deve ser suportado num conjunto exaustivo de testes computacionais, seguindo os seguintes aspetos [1-3]:

- definição do plano de testes - objetivos dos testes, seleção das instâncias e variáveis de entrada
- definição dos critérios de avaliação - qualidade das soluções, esforço computacional e robustez
- e a análise de resultados - visualização gráfica dos resultados, interpretação dos resultados, análise de significância.

4.1. Análise de desempenho de técnicas de otimização

O problema de escalonamento Job-Shop pode ser definido da seguinte forma: existe um conjunto de tarefas a processar num conjunto de máquinas; cada tarefa tem uma ordem de processamento nas máquinas específicas, isto é, a tarefa é composta por uma lista ordenada de operações, as quais são caracterizadas pela máquina onde são realizadas e pelo respetivo tempo de processamento. Às tarefas e às máquinas estão associadas algumas restrições, nomeadamente:

- não existem restrições de precedência entre operações de diferentes tarefas;
- o processamento de uma operação não pode ser interrompido;
- cada máquina só pode processar uma tarefa de cada vez;
- cada tarefa só pode ser processada numa máquina de cada vez.

São fornecidos os resultados de 3 técnicas de otimização (TO1, TO2, TO3) na resolução de um conjunto de instâncias do problema de escalonamento minimização do tempo de conclusão de todas as tarefas(C_{max}) e os tempos computacionais obtidos (u.t). Com base nos dados disponibilizados:

a) Analise o desempenho das técnicas de otimização

b) Verifique se existe uma diferença significativa no desempenho das técnicas de otimização. Identifique a mais eficaz e a mais eficiente.

4.2. Análise da influência dos parâmetros no desempenho das técnicas de otimização

Com vista a avaliar o impacto ou uma possível relação entre algumas regras usadas tradicionalmente na construção de uma solução inicial e a obtenção de uma solução final de melhor qualidade, isto é, mais próxima da solução ótima global do problema, foram efetuados alguns testes computacionais utilizando o algoritmo de Pesquisa Tabu(PT) em instâncias do problema do tipo WT ("Weighted Tardiness"), com 40, 50 e 100 tarefas.
São fornecidos os resultados da Pesquisa Tabu na resolução de um conjunto de instâncias do problema de escalonamento de minimização dos atrasos pesados (WT) e os tempos computacionais (u.t.) obtidos, com diferentes parâmetros:

a) Para cada instância foram consideradas 10 corridas do algoritmo de PT, partindo de soluções iniciais, geradas pelos mecanismos MGSI1 e MGSI2. Com base nos dados disponibilizados (melhor, pior e média das soluções obtidas nas 10 corridas), analise a influência do mecanismo de geração da solução inicial no desempenho da técnica de otimização PT.

b) Para cada instância foram consideradas 10 corridas do algoritmo de PT, com vizinhanças geradas pelos mecanismos MGV1 e MGV2. Com base nos dados disponibilizados (melhor, pior e média das soluções obtidas nas 10 corridas), analise a influência do mecanismo de geração de vizinhanças no desempenho da técnica de otimização PT.

5. Referências Bibliográficas

[1]. Barr, B.L. Golden, J.P. Kelly, M.G.C. Resende, and W.R. Stewart, **Guidelines for Designing and Reporting on Computational Experiments with Heuristic Methods**, 2001.

[2]. David S. Johnson, **A Theoretician's Guide to the Experimental Analysis of Algorithms**, AT&T Labs - Research, http://www.research.att.com/dsj/, 2001.

[3]. El-GhazaliTalbi, **Metaheuristics: From Design to Implementation**, Wiley, 2009.

B 2. Enunciados de Trabalhos – Edição 2019-2020

B 2.1 - Análise de Usabilidade

B 2.2 - Análise de Desempenho

Trabalho Prático

Análise de Dados em Informática

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2019/2020

---

1. Objetivos

Objetivo Geral:

- Análise de Usabilidade

Objetivos específicos:

- Análise de Usabilidade
  - Definir plano da sessão de avaliação
  - Definir, construir e recolher dados de inquérito de satisfação
  - Análise e discussão dos Resultados com recurso ao R
- Escrita do relatório

1. Calendarização

Lançamento das propostas de trabalhos: até 6 de março de 2020

Entrega do trabalho: até **19 de abril de 2020** (23:55)
Defesa e discussão: em data a marcar pelo professor de TP

3. Normas

- O grupo deve ser o mesmo em todos os trabalhos práticos. A defesa e discussão poderão ser realizadas individualmente.
- Deverá ser usado o R como ferramenta de suporte ao estudo estatístico.
- A **data final de ENTREGA** do trabalho é **19 de abril de 2020**, no moodle. Independentemente destes prazos, os grupos deverão ser capazes de, quando o professor o solicitar, reportar o estado de desenvolvimento do trabalho.
- A entrega do trabalho consta de um **relatório e respetivos anexos**, **inquérito, guia da sessão de avaliação de usabilidade, dados recolhidos, ficheiros de dado**s e **script do R em ficheiro**. Deverá submeter todos os documentos num ficheiro compactado. O nome do ficheiro deverá seguir a seguinte notação:

  ANADI_YYY_XXX_Nºaluno1_Nºaluno2_Nºaluno3.zip, onde **YYY** representa a sigla do docente das TP, e **XXX** representa a turma TP.

  Exemplo: **ANADI_AIM_3DA_7777777_8888888_9999999.zip.**

- Trabalhos cujo nome não respeite a notação indicada, **serão penalizados em 10%.**
- A **Entrega do trabalho deverá ser submetida no moodle até à data de entrega definida. Não serão aceites trabalhos fora do prazo**.
- A defesa e discussão dos trabalhos decorrerá em dia e hora a marcar por cada professor das teórico-práticas. No dia da apresentação, **TODOS** os elementos do grupo deverão estar presentes. Os elementos ausentes não terão classificação.
- A avaliação do trabalho será realizada por uma equipa de docentes.
- Cada grupo é responsável por gerir o seu processo de desenvolvimento. Dificuldades e problemas deverão ser comunicados atempadamente ao professor das aulas teórico-práticas.

3.1 Relatório

No relatório final deverão ser documentadas todas as fases da Análise de Usabilidade realizadas, contextualização do tema, recolha dos dados, organização do estudo estatístico, análise e discussão dos resultados, conclusões e anexos.

3.2. Avaliação

Na avaliação do trabalho serão considerados os seguintes aspetos:

- Contextualização e enquadramento teórico em cada uma das áreas temáticas
- Organização e recolha dos dados
- A análise exploratória de dados com recurso ao R
- A defesa e discussão
- Participação individual de cada um dos elementos

Análise de Usabilidade	
• Enquadramento teórico	10%
• Plano da sessão da Avaliação de Usabilidade	10%
• Inquérito	10%
• Análise Exploratória de Dados	25%
• Inferência Estatística	25%
Conclusões	**10%**
Relatório	**10%**

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua participação. A equipa de avaliação de trabalhos práticos irá validar, no momento da defesa do trabalho (via videoconferência), a participação de cada um dos elementos do grupo na concretização dos objetivos do trabalho e do grupo. **Os elementos ausentes não terão classificação.**

4. Descrição do Trabalho

Na realização deste trabalho pretende-se que os alunos desenvolvam o processo de Análise de Usabilidade, com o objetivo de testar e avaliar a usabilidade de sistemas, portais, sites e apps.

Um sistema interativo é constituído por dois elementos principais: a parte computacional, também designada por funcional, e pela parte comunicacional (Hix e Hartson, 1993). A primeira é invisível aos utilizadores e é responsável pelo processamento das ações, pelo espoletar das funções adequadas, e por fornecer ao utilizador o necessário feedback. A segunda, a parte comunicacional, é responsável pela comunicação e está a cargo da chamada interface com o utilizador (user interface ou UI), pois é a parte visível. Relativamente ao conceito de usabilidade, este é definido como a eficiência e adequabilidade na concretização de determinados objetivos por determinados utilizadores (Karray et al., 2008). A definição de usabilidade de um sistema computacional alterou-se ao longo tempo devido à melhor compreensão da interação entre humano e computador e dos fenómenos que rodeiam esse processo de comunicação. Uma das mais conhecidas definições de usabilidade foi apresentada por Nielsen (1993): "*a usabilidade apresenta múltiplos componentes e é tradicionalmente associada a cinco atributos: facilidade de aprendizagem, eficiência, facilidade de memorização, reduzida taxa de erros, e satisfação de utilização*". A usabilidade pretende promover todos os aspetos que facilitem a utilização de um sistema computacional. Seguindo os passos definidos por Hix e Hartson (1993) para avaliar a

à recolha de dados. Posteriormente devem realizar a análise dos dados recolhidos e formular as conclusões.

O desenvolvimento de aplicações tem por base um trabalho de estudo sobre qual o conceito de design a adotar, a usabilidade (UI – interface do utilizador) e a experiência do utilizador (UX).

Pretende-se com este trabalho realizar a análise de usabilidade do modelo de interação, de business apps de modo a avaliar a satisfação dos utilizadores. Sugere-se a análise de usabilidade dos sites de comércio eletrónico:

- Continente Online
- Auchan Online
-

Nota: Com os constrangimentos que todos estamos a viver, estes sites têm estado muito congestionados para necessidades prementes, pelo que sugeria libertá-los. Assim sendo, podem identificar 2 sites de e-commerce que possam analisar a *interface* e o modelo de interação.

Realça-se que o objetivo não é verificar qual deles é o mais utilizado ou tem os produtos mais baratos. Pretende-se analisar e avaliar a usabilidade de cada site de comércio eletrónico, do ponto de vista do modelo de interação:

- Definir o plano da sessão da Avaliação de Usabilidade.
- Construir o Inquérito para recolher os dados referentes ao grau de satisfação dos utilizadores dos 2 sites de comércio eletrónico.
- Realize a Análise Exploratória dos Dados recorrendo a técnicas de Estatística Descritiva. Identifique os tipos de dados utilizados, represente os dados em gráficos adequados, calcule medidas descritivas de localização, dispersão e forma. Comente. Identificar o que melhor se adequa em termos de usabilidade do modelo de interação.
- Para cada site, calcule uma medida de associação entre o grau de satisfação e o tempo de utilização do referido site. Comente os resultados obtidos verificando, ao nível de 5%, se o grau de associação é significativo.
- Identifique o site de comércio eletrónico (Continente Online-Auchan Online, ou outro par de sites de e-commerce) que melhor contribui para a satisfação dos utilizadores. Caso existam diferenças significativas, identifique qual o site preferido pelos utilizadores. Responda, considerando uma significância de 5%.
  Nota: Defina no inquérito as questões adequadas (para avaliar os 2 sites do ponto de vista do modelo de interação) para poder responder a esta questão.

5. Referências Bibliográficas

KARRAY, F., ALEMZADEH, M., SALEH, J. A., ARAB, M. N., Human-Computer Interaction: Overview on State Art, International Journal on Smart Sensing and Intelligence, 2008.

Nielsen, J. Usability Engineering. AP Professional, 1993.

HIX, D. and HARTSON, H. R. - Developing User Interfaces: Ensuring Usability Through Product & Process. John Wiley & Sons, inc ,1993.

PIAIRO, E., Desenvolvimento de Interface Inteligente para Suporte à Gestão e Controlo de Produção, Dissertação de Mestrado, FEUP, 2012.

HEUMANN, C., M. SCHOMAKER and SHALABH, Introduction to statistics and data analysis, Springer International Publishing, 2016.

Trabalho Prático

Análise de Dados em Informática

Análise de Desempenho

Engenharia Informática - 3º ano 2º semestre
Ano Letivo 2019/2020

---

1. Objetivos

Objetivo Geral:

- Análise de Desempenho de técnicas de aprendizagem automática

Objetivos Específicos:

- Definir a metodologia de trabalho
- Análise e discussão dos Resultados com recurso ao R
- Escrita de artigo científico

2. Calendarização

Entrega do trabalho: até **13 de junho de 2020 pelas 23:55**
Defesa e discussão: em data a marcar pelo professor de TP

3. Normas

- O grupo deve ser o mesmo em todos os trabalhos práticos.

- Deverá ser usada a ferramenta R.

- A **data final de ENTREGA** do trabalho é **13 de junho de 2020 pelas 23:55**, no moodle.
  Independentemente destes prazos, os grupos deverão ser capazes de, quando o professor o solicitar, reportar o estado de desenvolvimento do trabalho.

- A entrega do trabalho consta de um artigo científico (**máx. 8 páginas**) conforme *template* disponibilizado no moodle, apresentação *powerpoint* com resumo do trabalho realizado, entre outros. Deverá submeter todos os documentos num ficheiro compactado. O zip file deve conter:
  - artigo científico em pdf
  - dados utilizados em formato csv
  - script completo (e comentado) do código criado em R para resolver o problema
  - apresentação PowerPoint com resumo do artigo para 10 minutos (ppt)

- O nome do ficheiro zip deverá seguir a seguinte notação:

  ANADI_YYY_XXX_Nºaluno1_Nºaluno2_Nºaluno3.zip, onde **YYY** representa a sigla do docente das TP, e **XXX** representa a turma TP.

  Exemplo: **ANADI_AMD_3AD_7777777_8888888_9999999.zip.**

- Trabalhos cujo nome não respeite a notação indicada **serão penalizados em 10%.**

- A **entrega do trabalho deverá ser submetida no moodle até à data de entrega definida**. **Não serão aceites trabalhos fora do prazo**.

- A apresentação, **em formato de comunicação (10 minutos),** e discussão dos trabalhos decorrerá em dia e hora a marcar por cada professor das teórico-práticas. No dia da apresentação, **TODOS** os elementos do grupo deverão estar presentes. Os elementos ausentes não terão classificação.

- A avaliação do trabalho será realizada pelo docente das aulas TP.

- Cada grupo é responsável por gerir o seu processo de desenvolvimento. Dificuldades e problemas deverão ser comunicados atempadamente ao professor das aulas teórico-práticas.

3.1. *Artigo Científico*

No Artigo Científico (máx. 8 páginas) deverão ser documentadas todas as fases da metodologia de trabalho seguida, contextualização do tema, exploração, preparação dos dados, análise e discussão dos resultados, conclusões e referências bibliográficas.

3.2. Avaliação

Na avaliação do trabalho serão considerados os seguintes aspetos:

- Revisão do estado da arte (algoritmos de aprendizagem automática e análise de desempenho);
- Desenvolvimento de modelos de *Machine Learning*;
- A qualidade do processo de análise de dados seguido, a organização do código, a avaliação dos modelos criados e as conclusões alcançadas;
- Organização, qualidade da escrita, apresentação e clareza do artigo científico;
- A apresentação numa aula e discussão;
- Participação individual de cada um dos elementos.

Contextualização (Abstract, Introdução, estado da arte)	**10%**
Análise de desempenho de técnicas de aprendizagem	**70%**
Conclusões	**10%**
Apresentação e Discussão	**10%**

Nota: A nota de cada um dos elementos do grupo será definida de acordo com a sua participação. No momento da defesa do trabalho será validada a participação de cada um dos elementos do grupo na concretização dos objetivos do trabalho e do grupo.

4. Descrição do Trabalho

O objetivo principal deste trabalho consiste na aplicação de algoritmos de aprendizagem automática na exploração de dados e respetiva comparação dos mesmos usando os testes estatísticos mais adequados. Deve ser produzido um artigo científico (português ou inglês), conforme *template* indicado, com o estado da arte sobre os diferentes algoritmos, os modelos desenvolvidos, os resultados obtidos, a análise e discussão dos resultados e as conclusões.
Pretende-se que façam a análise salarial da população masculina de uma dada região, com base nos atributos abaixo descritos, através de modelos de classificação/regressão usando os algoritmos estudados: regressão linear, árvores de decisão, k-vizinhos-mais-próximos e redes neuronais.

O conjunto de dados a analisar neste trabalho diz respeito a salários e outras informações para 3000 trabalhadores do sexo masculino. Pretende-se que explorem as relações entre salário e os restantes atributos deste conjunto de dados:

Atributo	Descrição
Ano	Data de registo da informação
Idade	Idade do funcionário
Sexo	Sexo do funcionário
Estado	Estado civil
Raça	Etnia
Grau	Nível de estudos do funcionário
Emprego	Tipo de emprego
Saude	Nível de saúde do funcionário
Seguro	Funcionário com seguro de saúde (sim/não)
Salario	Salário dos funcionários

1. Comece por carregar o ficheiro ("**dados_emprego.csv**") para o ambiente do R, verifique a sua dimensão e obtenha um sumário dos dados.

2. Faça um estudo da regressão linear entre a variável dependente (**Salario**) e a variável independente (**Idade**):
   a. Calcule a correlação entre as variáveis **Salário** e **Idade**;
   b. Encontre a reta de regressão linear entre a variável dependente (**Salário**) e a variável independente (**Idade**);
   c. Verifique as condições sobre os resíduos (normalidade, independência e homocedasticidade).

3. Encontre o modelo de regressão linear generalizado onde a variável dependente é o "**Salario**" e as variáveis independentes são a "**Idade**", o "**Grau**" e o "**Emprego**". Com base no modelo encontrado, estime:
   a. Uma pessoa com 32 anos, com o emprego de "Industria", para todos os diferentes graus de escolaridade;
   b. Uma pessoa com 32 anos com o "12º ano escolaridade", para todos os diferentes tipos de emprego.

4. Derive um novo atributo **Nivel**, discretizando o atributo **Salario** em duas classes: Alto e Baixo, usando como valor de corte a mediana.

5. Faça uma Análise Exploratória de Dados, usando os gráficos apropriados, de modo a analisar os vários atributos (numéricos e categóricos) do conjunto de dados.

6. Usando o método ***k-fold cross validation*** estude a capacidade preditiva de alguns métodos de previsão relativamente ao novo atributo **Nivel**:
   a) Um modelo árvore de decisão;
   b) Uma rede neuronal. Deve avaliar o desempenho para diferentes configurações;
   c) Compare as soluções obtidas para as medidas de avaliação ***accuracy***, **precision**, ***recall*** e ***F1;***
   d) Verifique se existe diferença significativa no desempenho dos diversos algoritmos (use um nível de significância de 5%). Identifique a técnica de aprendizagem automática que apresenta melhor desempenho.

7. Usando o método ***k-fold cross validation*** obtenha a previsão do atributo **Salario** com um modelo:
   a) K-vizinhos-mais-próximos e o valor de k mais adequado
   b) Árvore de regressão
   c) Obtenha a média e o desvio padrão do RMSE dos modelos
   d) Verifique se existe diferença significativa no desempenho dos dois melhores modelos obtidos anteriormente (use um nível de significância de 5%). Identifique o modelo que apresenta o melhor desempenho.

5. Referências Bibliográficas

[1]. Christopher Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
[2]. Tom Mitchell, Machine Learning. McGraw-Hill, 1997.