Uma imagem com texto, Tipo de letra, captura de ecrã, logótipo

Descrição gerada automaticamente

Engenharia Conhecimento

Licenciatura em Engenharia Informática

2023/2024

**Modelos de Classificação e Regressão**

**Thyroid disease Data Set**

DEPARTAMENTO DE INFORMÁTICA

Grupo 6:

- Eduardo Proença 57551

- Tiago Oliveira 54979

- Bernardo Lopes 54386

## Introdução e Objetivos

Este projeto foi desenvolvido com o intuito de utilizar diferentes técnicas de aprendizagem automática, aprendidas ao longo da unidade curricular, para construir os melhores modelos de classificação e regressão possíveis.

Para tal, foi fornecido um conjunto de dados, uma versão do "thyroid0387" data set, que contém 31 características e 7338 instâncias. Os dados deste conjunto, não se encontram processados, sendo possível encontrar valores categóricos e contínuos, assim como valores em falta. Desta maneira, ao longo do desenvolvimento do projeto irão ser utilizadas técnicas como codificação, normalização e imputação de dados.

Quanto à construção dos modelos serão utilizadas técnicas como seleção de variáveis, *cross validation* e afinamento de hiperparâmetros.

Ao longo do projeto serão testados diferentes modelos como, DecisionTreeClassifier, LogisticRegression, Naive Bayes, KNN e SVM

**Objetivo 1 (O1)**

* Os alunos devem fornecer os melhores modelos de classificação possíveis usando quaisquer métodos abordados nas aulas
* A classificação deve ser de acordo com 8 classes da variável alvo “diagnoses”

**Objetivo (O2)**

* Podemos prever com exatidão a idade dos indivíduos, tendo em conta os outros atributos?
* Podemos prever com exatidão o sexo dos indivíduos, tendo em conta os outros atributos

**Objetivo (O3)**

* Quais são as características mais significativas dos melhores modelos obtidos acima (O1 e O2)

## Processamento dos dados

Devido ao facto de terem sido testados vários modelos e embora não seja a melhor prática, foi decido fazer o processamento dos dados antecipadamente, de maneira a reduzir a complexidade da tarefa e por sua vez o tamanho da mesma.

Desta forma, o ideal seria fazer um processamento de dados diferente para cada modelo, tendo em conta as suas características. Portanto, é de realçar que os resultados obtidos não são perfeitos, pelo que ainda existe margem para melhoria. No entanto, as técnicas utilizadas neste projeto, foram técnicas presentes na construção de diferentes tipos de modelos, e que na sua maioria permitem bons resultados.

**Perfil de dados**

Uma imagem com texto, captura de ecrã, número, Tipo de letra

Descrição gerada automaticamenteComeçou-se por fazer um *profiling* dos nossos dados (usando a classe *ProfileReport* do package *ydata\_profiling*), para obter uma informação mais detalhada sobre as nossas variáveis.

Figura 1: ProfileReport Overview

Devido ao facto dos valores em falta serem representados por “?” e algumas colunas estarem em formato de texto, para além de não ser possível ver valores em falta no *report*, este também nos mostra *text variables*, que na verdade são valores numéricos.

Desta forma, olhando para estes resultados e para a explicação fornecida no ficheiro “data.names”, foi possível identificar três tipos de variáveis: numéricas, booleanas e categóricas. Antes de proceder com a codificação dos dados, foi apagada a variável “[record identification]”, pois apenas representa um identificador único, o que não é relevante. Foram também substituídos os valores “?” por *NaN* para serem identificados os valores em falta.

**Codificação de dados**

**Divisão dos dados**

**Normalização**

**Imputação de dados**

## Seleção de variáveis

O processo de seleção de variáveis foi feito através do uso da classe *SequencialFeatureSelector* (SFS) juntamente com modelos regressão, como o *LinearRegression* e também classificação, como o *LogisticRegression*. Foram realizados diferentes testes, com vários números possíveis de variáveis, utilizando esta técnica. Em baixo encontra-se um exemplo dos resultados:

Uma imagem com diagrama, Desenho técnico, Esquema, texto

Descrição gerada automaticamente

Entre os diferentes números de variáveis testados, os que obtiveram melhor resultado, de acordo com o *F1 Score*, foram 13 e 25. Desta maneira, optou-se por utilizar 13 variáveis, pois para além de melhorar os nossos modelos, eliminando dados que possam ser irrelevantes, também ajuda a reduzir o tempo que é necessário para treinar os mesmos.

## Resultados dos modelos

## Afinação dos hiperparâmetros

## Conclusão