Universidade do Minho
Escola de Engenharia

Inteligência Artificial para as Telecomunicações

ENGENHARIA DE TELECOMUNICAÇÕES E INFORMÁTICA

(Docentes: Paulo Jorge Freitas Oliveira Novais, Sérgio Manuel Carvalho Gonçalves)

Conceção e otimização de modelos de Aprendizagem Automática

Ano Letivo 2021/2022

Guimarães, 8 de janeiro de 2021

Hugo Miguel Miranda Reinolds - a83924@alunos.uminho.pt

Pedro Miguel Fonseca Sampaio - a79668@alunos.uminho.pt

Rui Filipe Ribeiro Freitas - a84121@alunos.uminho.pt

Tiago João Pereira Ferreira - a85392@alunos.uminho.pt

Índice

Índice de figuras

Índice de tabelas

Introdução

O objetivo principal deste trabalho é conceber e desenvolver um projeto de aprendizagem automática recorrendo aos modelos abordados ao longo do semestre na cadeira de Inteligência Artificial para as Telecomunicações. A ferramenta que utilizamos na realização deste projeto foi o *Knime*, software inicialmente desconhecido pelo grupo, mas que após alguma pesquisa conseguimo-nos adaptar bem.

Este projeto foi dividido em 3 fases principais onde na primeira foi necessária a pesquisa por um conjunto de dados que permitisse uma consulta de informação objetiva e para isso foi utilizado um *dataset* sobre a esperança de vida de vários países do mundo tendo em conta fatores como doenças, mortalidade infantil, grau de desenvolvimento do país em questão, entre outros. Na segunda fase procuramos preparar os dados de modo a conceber e otimizar modelos de aprendizagem automática em particular focando os modelos de aprendizagem supervisionada de regressão e de classificação. Numa última fase procuramos analisar os dados da melhor maneira possível recorrendo a gráficos e interpretando estes tendo em conta a utilidade dos resultados no contexto do problema da esperança de vida.

De modo a sermos capazes de cumprir com os objetivos propostos no enunciado foi necessário colocar em prática conhecimentos adquiridos ao longo das aulas teóricas e práticas da unidade curricular. Estes conhecimentos foram essenciais para perceber o que era pedido e assim conseguirmos chegar a uma solução eficiente do nosso ponto de vista.

Fundamentos

1. Aprendizagem automática (*Machine Learning*)

Ao longo dos anos a inteligência artificial tem sido bastante trabalhada com a finalidade de melhorar o dia-a-dia das pessoas, um dos pontos que foram desenvolvidos, sendo que ainda se encontra em desenvolvimento, é o *Machine Learning* que permite que aplicações de software façam precisas previsões de resultados.

Machine Learning, tem como característica essencial a capacidade de aprender de um modo autónomo. Utiliza assim sistemas de aprendizagem, que serão de seguida mencionadas.

1. Sistema de Aprendizagem com Supervisão
2. Sistema de Aprendizagem sem Supervisão
3. Aprendizagem por reforço

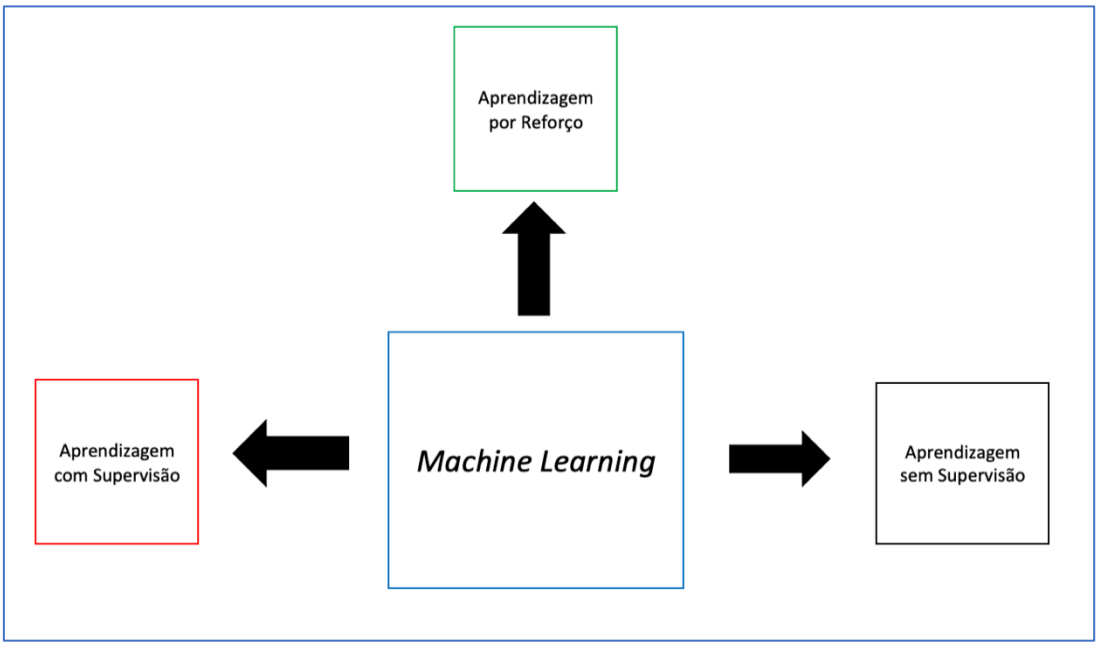

Figura 1 - Diagrama Machine Learning.

De seguida, iremos abordar cada método um a um, de forma que mais a frente ao explicarmos a componente prática exista melhor compreensão por parte de todos.

1.1. Sistema de aprendizagem com supervisão

Este paradigma de aprendizagem contém informações que são importantes mais tarde serem utilizados. Este método vai utilizar a informação adquirida, que podem ser resultados pretendidos, e os resultados produzidos pelo sistema, isto permite que após estes cálculos o sistema, possa estabelecer uma relação entre estes dois tipos de valores.

O sistema está dividido em duas componentes:

- Classificação -> quando os resultados são discretos;
- Regressão -> quando os resultados são contínuos

1.2. Sistema de aprendizagem sem supervisão

Este paradigma de aprendizagem, ao contrário do que foi mencionado anteriormente, não contém informação sobre os casos, apenas sabemos o problema, isto implica que, que é necessário encontrar técnicas e/ou métodos de aprendizagem para chegar a uma determinada solução.

Este sistema está divido em duas categorias que serão apresentadas de seguida:

- Segmentação -> baseia-se na organização dos dados;
- Associação -> Existe tratamento dos dados obtidos.

1.3. Aprendizagem por reforço

O paradigma de aprendizagem por reforço tem uma semelhança com a anterior, visto que também não possui informação, mas permite que após os resultados produzidos pelo sistema de aprendizagem são satisfatórios ou não.

É importante referir que dos tipos de sistemas apresentados, o grupo optou apenas por aplicar o primeiro “Sistema de Aprendizagem com Supervisão”, porque no contexto do problema proposto era o método mais fácil para se obter resultados mais fiáveis e com menor margem de erro.

Desenvolvimento

Neste capítulo são demonstrados os vários passos que nos levaram a chegar à solução por nós implementada. Esta solução foi a que achamos mais adequada tendo em conta o problema que escolhemos resolver que foi analisar o grau de desenvolvimento de um país e a esperança média de vida tendo em conta certos atributos particulares como doenças, mortalidade infantil, entre outros.

1. Especificação do projeto

Para a realização deste trabalho foi pedido que procurássemos e analisássemos um *dataset* que contivesse conhecimento relevante no contexto do nosso problema. Problema este que através de aprendizagem automática procura identificar um país como desenvolvido ou em desenvolvimento tendo em conta vários parâmetros como as doenças presentes num certo país e o grau de incidência na população, a taxa de mortalidade infantil e adulta, o nível de escolaridade, etc. Para realizar esta aprendizagem automática utilizamos alguns modelos de aprendizagem supervisionada, de regressão e de classificação explicados anteriormente. Após a realização do *workflow* é necessária a realização de uma análise crítica aos resultados e interpretar a sua utilidade no contexto dos problemas subjacentes.

2. Domínios e objetivos do dataset

O nosso *dataset* foi escolhido baseado numa pequena pesquisa realizada sobre os melhores assuntos para tratar quando queremos realizar uma aprendizagem supervisionada de regressão e de classificação. Este *dataset* foi elaborado por parte da WHO (World Health Organization) que é uma agência internacional especializada na área da saúde pelo que é uma fonte credível de dados. Quanto ao *dataset* em específico este dá-nos informações sobre o país, a esperança média de vida, várias informações sobre doenças, informações sobre a mortalidade tanto de jovens como adultos, população e outros aspetos que podem ser considerados importantes. Quanto aos problemas que decidimos procurar resolver foram 2 distintos, um para a aprendizagem supervisionada de regressão e outra para a de classificação. Para a aprendizagem de regressão procuramos obter uma estimativa de qual seria a esperança média de vida de um certo país tendo em conta os atributos disponíveis e para a aprendizagem de classificação procuramos obter informações sobre o grau de desenvolvimento de um dado país, ou seja, tendo em conta os atributos disponíveis este país era um país desenvolvido ou um país em desenvolvimento.

3. Preparação do dataset

Quanto ao tratamento realizado ao *dataset* este passou principalmente por realizar a leitura do ficheiro que continha o conjunto de dados através de um bloco "*CSV Reader*". Depois utilizamos um "*Rule-Base Row Filter*" de modo a retirar todas as linhas da base de dados onde estivessem presentes valores desconhecidos ou não aceites. Após isso realizamos a normalização dos dados com um bloco "*Normalizer*" e a respetiva repartição em 2 grupos, um de treino e um de teste. Esta repartição foi realizada com auxílio do bloco "*Partitioning*" no *Knime*. De seguida é demonstrada através de uma figura esta secção do nosso *workflow* que trata da preparação dos dados.

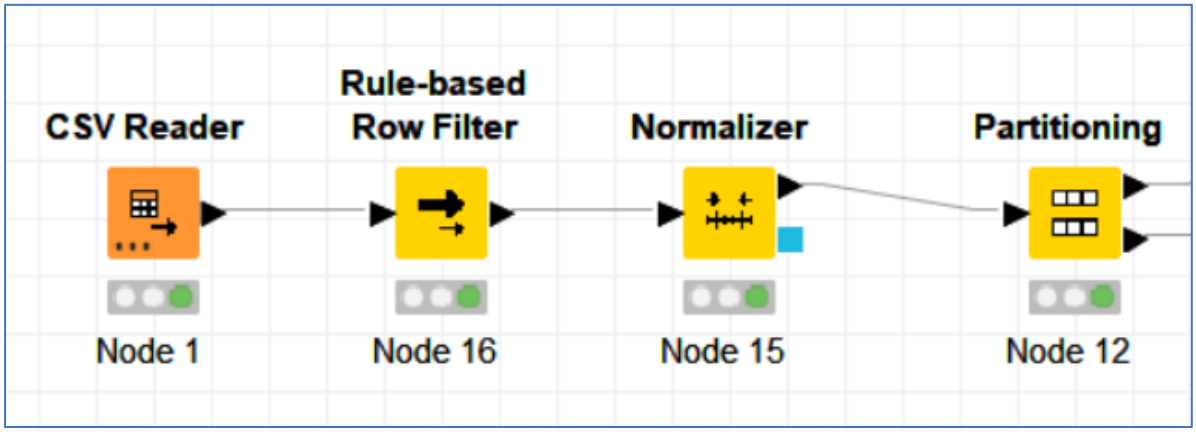

Figura 2 - Preparação do dataset.

4. Descrição dos modelos e workflows desenvolvidos

4.1. Análise de aprendizagem supervisionada de classificação

Relativamente ao primeiro *workflow* que o grupo realizou este é de aprendizagem supervisionada de classificação e realizamos 2 modelos, um baseado no algoritmo de árvore de decisão com utilização dos blocos "*Decision Tree Learner*" e "*Decision Tree Predictor*" no *Knime* e no algoritmo de regressão lógica que apesar do nome foi utilizado como classificação por permitir obter 2 valores. Neste caso utilizamos os blocos "*Logistic Regression Learner*" e "*Logistic Regression Predictor*" do *Knime*.

De seguida apresentamos uma figura que ilustra o primeiro *workflow* que realizamos, os últimos blocos do *workflow* são blocos de análise de dados que iremos falar no capítulo de testes mais à frente neste relatório.

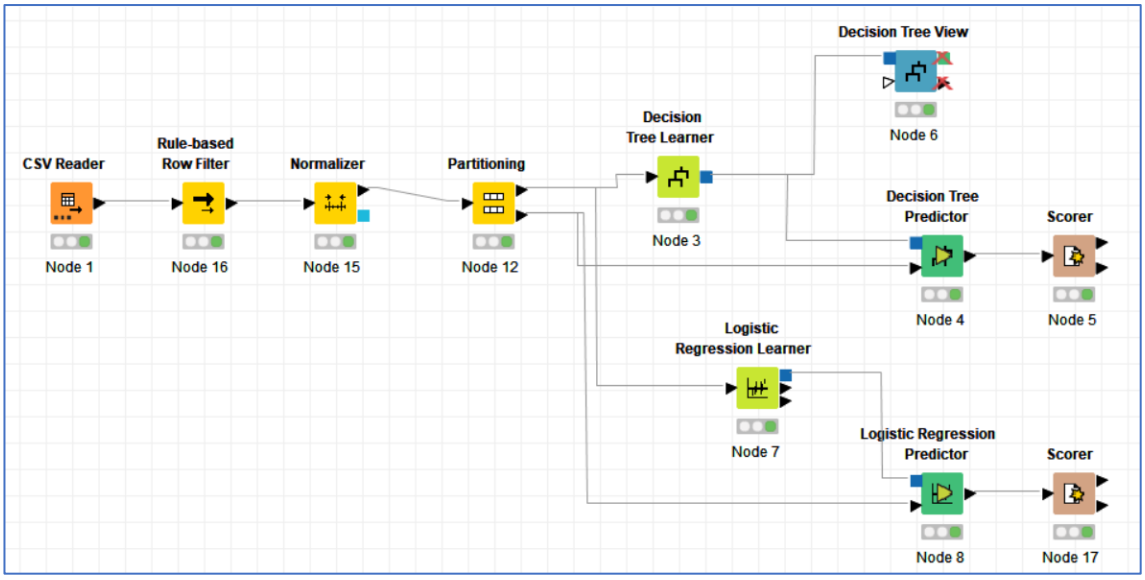

Figura 3 - Workflow de aprendizagem supervisionada de classificação.

Neste *workflow* não houve necessidade de grande configuração dos blocos de aprendizagem apenas tivemos de colocar qual a variável que queríamos observar e como queríamos dividir os dados na repartição. Nas próximas fotos apresentamos capturas de ecrã do bloco de "*Partitioning"* onde colocamos uma percentagem de 20% para os dados de teste e 80% para os dados de treino e do bloco "*Decision Tree Learner*" e "*Logistic Regression Learner*" onde colocamos a variável de análise Status que diz se um país é desenvolvido ou em desenvolvimento.

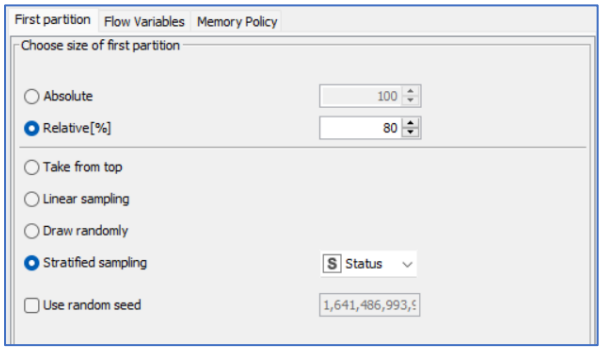

Figura 4 - Bloco Partitioning do Knime.

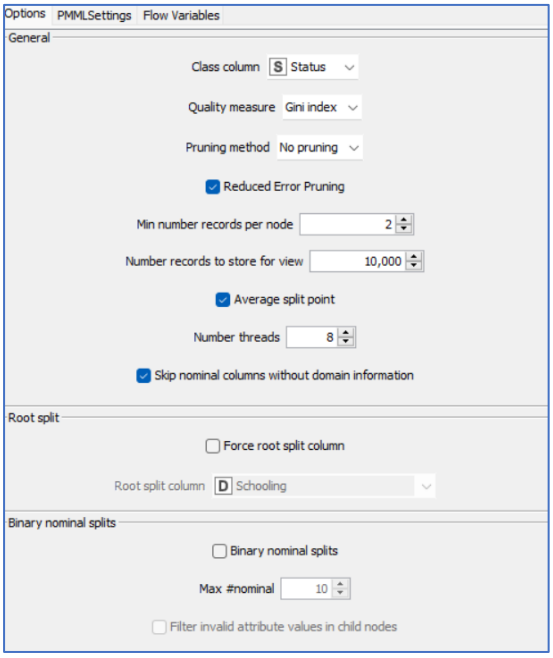

Figura 5 - Bloco Decision Tree Learner do Knime.

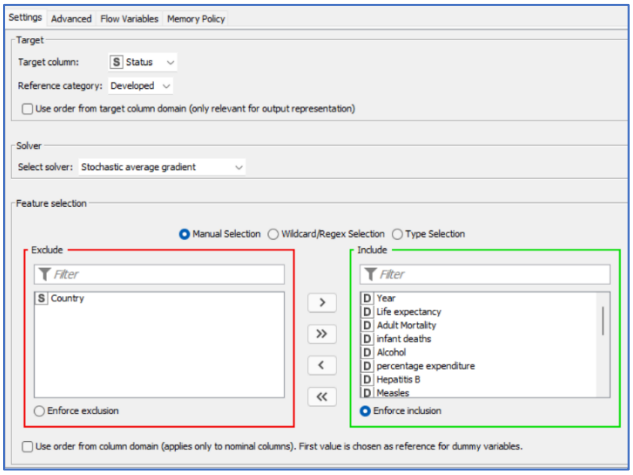

Figura 6 - Bloco Logistic Regression Learner do Knime.

4.2. *Análise de aprendizagem supervisionada de regressão*

Relativamente ao segundo *workflow* este foi realizado com o objetivo de estudar a aprendizagem supervisionada de regressão onde com o *dataset* por nós escolhido procuramos obter uma estimativa da esperança de vida de um país com base nos seus atributos. Isto foi realizado com auxílio do algoritmo de regressão linear onde no Knime foi utilizado o bloco "*Linear Regression Learner*" para a fase de aprendizagem e "*Regression Predictor*" para a realização da previsão da esperança de vida.

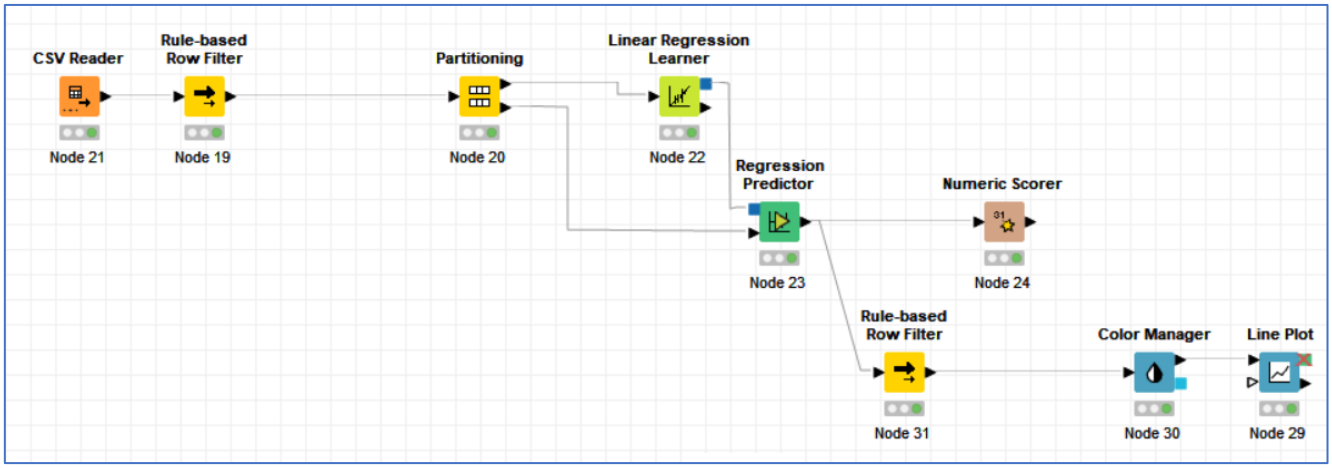

Figura 7 - Workflow de aprendizagem supervisionada de regressão.

Quanto à elaboração deste modelo a configuração foi feita apenas no bloco de aprendizagem da regressão linear onde tivemos de colocar qual era a variável que queríamos ensinar ao sistema. No nosso caso esta variável foi a esperança de vida como demonstrado na figura seguinte.

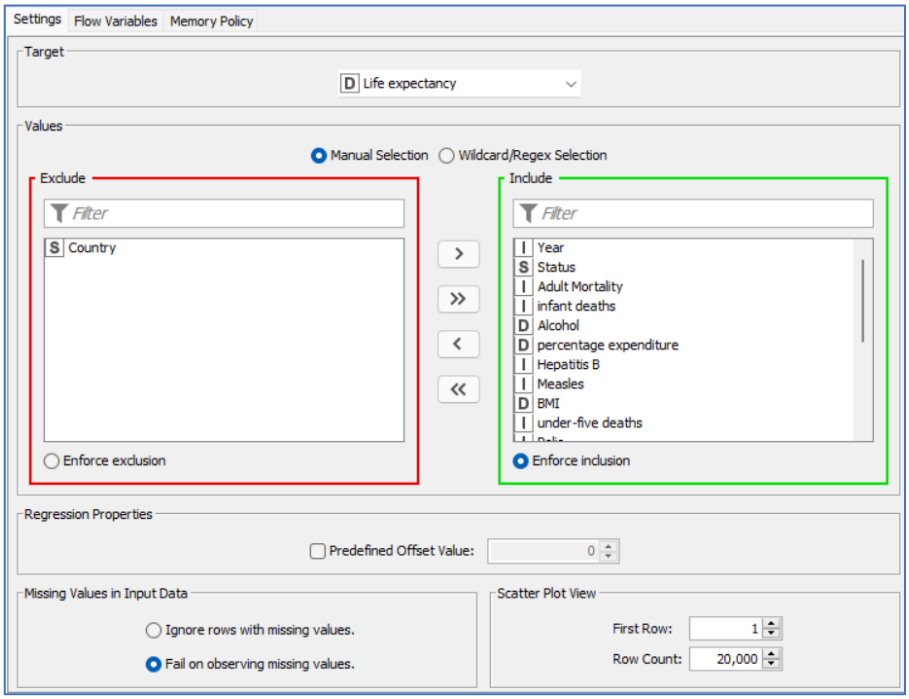

Figura 8 - Bloco Linear Regression Learner.

4.3. *Análise de ambas as técnicas de aprendizagem*

Neste terceiro *workflow* decidimos implementar um algoritmo que servisse para ambas as técnicas e o *Random Forest* pareceu-nos o mais apropriado. De seguida é demonstrado o *workflow* onde realizamos este algoritmo para a aprendizagem supervisionada de regressão e de classificação. Quanto à regressão utilizamos os blocos "*Random Forest Learner (Regression)*" e "*Random Forest Predictor (Regression)*" do Knime para a implementação. Já a classificação usamos os blocos "*Random Forest Learner*" e "*Random Forest Predictor*".

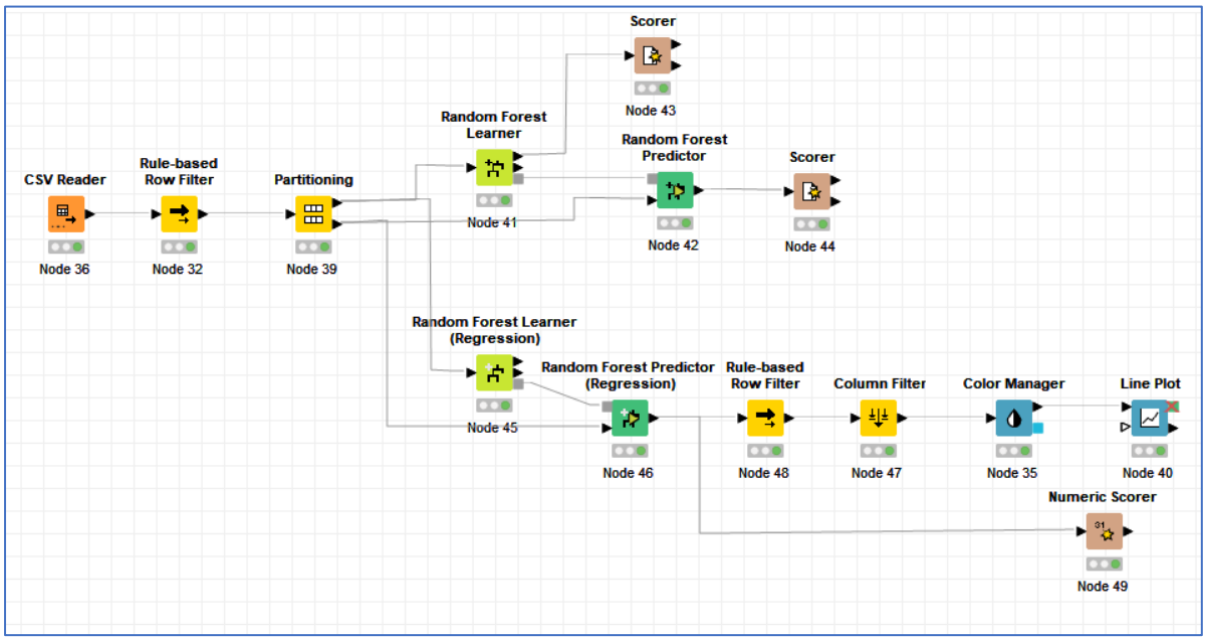

Figura 9 - Workflow com ambas as técnias de aprendizagem.

No que diz respeito à configuração dos blocos tivemos de realizar as mesmas mudanças faladas previamente, ou seja, alterando a variável que queríamos observar em que no caso da regressão era a variável da esperança de vida e no da classificação a variável do estado ou seja se um país era desenvolvido ou se encontrava em desenvolvimento. Para além disso no caso da regressão foi necessário filtrar por colunas de modo a realizar um gráfico que permitisse a visualização de resultados que iremos demonstrar no capítulo de testes do nosso relatório. Esta alteração no bloco "*Column Filter*" está presente de seguida.

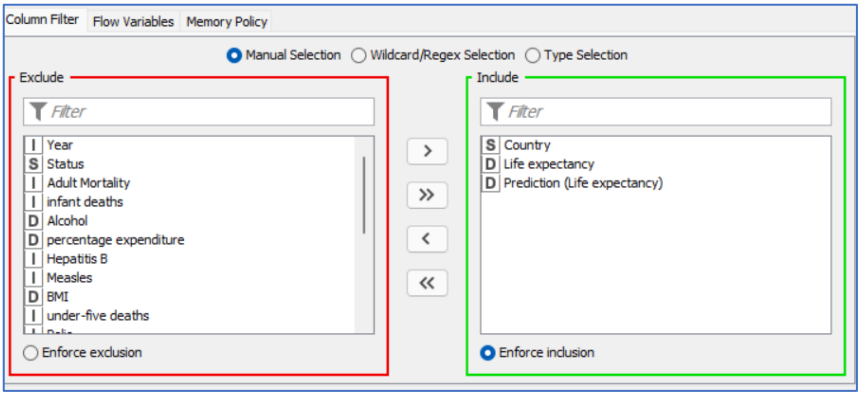

Figura 10 - Bloco Column Filter.

Testes e discussão de resultados

1. Aprendizagem supervisionada de classificação

Neste workflow o objetivo era prever se um país é desenvolvido ou se encontra em estado de desenvolvimento baseado nos fatores presentes no dataset. Foi atribuída uma percentagem de 20% para os dados de teste.

Na tabela seguinte está representada a confusion matrix da técnica Decision Tree Learner, onde relaciona os dados previstos com os dados reais. Como se pode ver o algoritmo tem uma boa percentagem de sucesso, mas previu que 6 países estavam desenvolvidos, enquanto estes encontravam-se em desenvolvimento. O algoritmo errou também nos países desenvolvidos, apontando que 9 países estavam desenvolvidos, mas na realidade não se encontravam.

Tabela 1 - Decision Tree Learner Confusion Matrix.

Decision Tree Learner		
Row ID	Developing	Developed
Developing	276	6
Developed	9	39

A tabela 2 corresponde a confision matrix do algoritmo Logistic Regression. Este teve um bom desempenho na previsão de países em desenvolvimento, errando apenas 6 países. Já na previsão de países desenvolvidos o caso foi diferente, este teve mais dificuldade e não obteve muito sucesso, errando 17 países.

Tabela 2 - Logistic Regression Confusion Matrix.

Logistic Regression		
Row ID	Developing	Developed
Developing	276	6
Developed	31	17

Depois de analisadas as duas tabelas podemos concluir que para este caso em específico, prever se um país se encontra em desenvolvimento ou desenvolvido, o algoritmo mais acertado a usar é o Decision Tree Learner, uma vez que foi o que teve uma menor percentagem de erros.

2. Aprendizagem supervisionada de regressão

Relativamente a este *workflow* o objetivo passou por prever qual seria a esperança média de vida de um país tendo em conta os vários fatores. Foi atribuída tal como anteriormente uma percentagem de 80% para a aprendizagem e 20% para a realização dos testes. Para a realização deste teste tivemos de realizar um filtro pelo ano de 2014 visto que o dataset que nós escolhemos continha um período de anos e para uma melhor visualização decidimos utilizar apenas 1. Isto foi obtido através de um bloco "*Rule-Based Row Filter*" do Knime.

A seguir é apresentada uma imagem ilustrativa do resultado por nós obtido com a comparação entre a esperança de vida presente no dataset e a esperança de vida prevista pelo modelo.

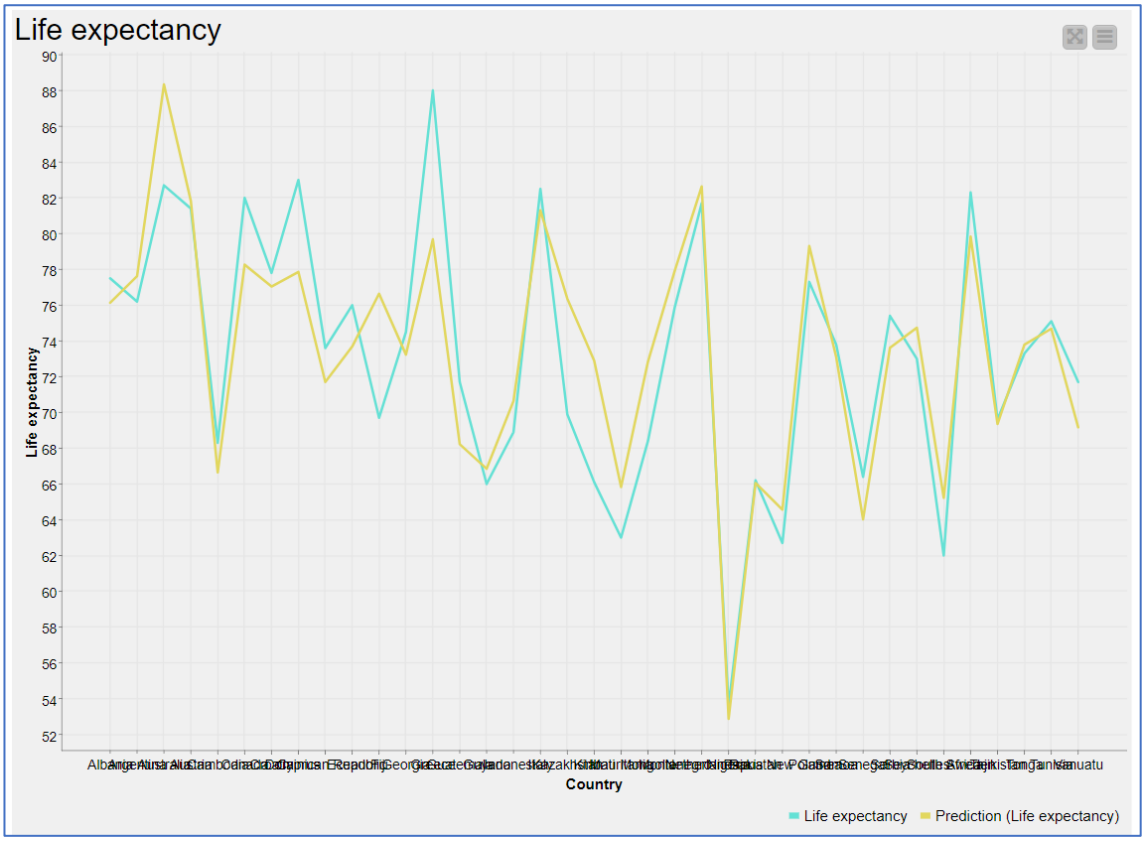

Figura 11 - Gráfico da previsão da esperança de vida vs esperança de vida.

Podemos então concluir que o modelo pode ser classificado como eficaz visto que os valores que obtivemos da previsão do que seria a esperança média de vida num país é relativamente similar à esperança de vida real.

3. Aprendizagem supervisionada de regressão e classificação

No terceiro e último *workflow* decidimos utilizar o mesmo algoritmo para ambas as técnicas de aprendizagem. Quanto à supervisionada de regressão procuramos novamente obter uma estimativa do que seria a esperança de vida num dado país e obtivemos o resultado apresentado de seguida.

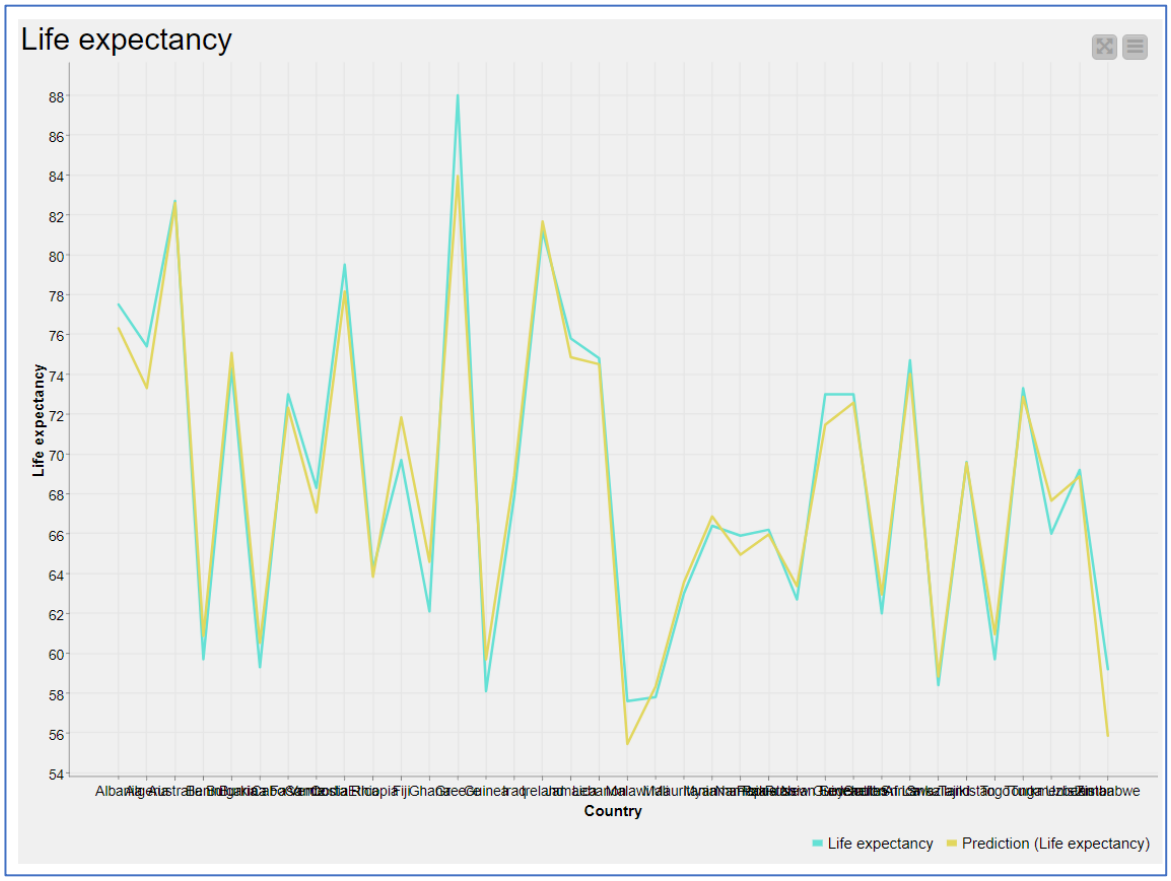

Figura 12 - Esperança de vida com algoritmo Random Forest.

Com a observação do gráfico anterior e comparando-o ao obtido quando foi utilizado o algoritmo de regressão linear podemos concluir que o algoritmo Random Forest é melhor e mais eficiente visto que as linhas da esperança de vida real e a prevista se encontram praticamente sobrepostas.

Relativamente à utilização do algoritmo Random Forest para a implementação da aprendizagem supervisionada de classificação foi utilizado o mesmo problema de procurar retirar se um país era desenvolvido ou em desenvolvimento tendo em conta os vários fatores do *dataset*. De seguida apresentamos os resultados obtidos através de uma tabela com a confusion matrix que relaciona os dados previstos com os reais.

Tabela 3 - Random Forest Confusion Matrix.

Random Forest Learner		
Row ID	Developing	Developed
Developing	419	3
Developed	3	70

Com a observação da tabela anterior podemos então concluir que este algoritmo é melhor que por exemplo o algoritmo Decision Tree visto que os valores que efetivamente deram "errados" são bastante menores.

Relativamente aos erros ocorridos fomos procurar ver quais eram os países que o algoritmo detetou erros e o país era a Grécia isto permitiu uma análise interessante visto que realmente na altura que foram retirados os dados do nosso dataset realmente o país encontrava-se em desenvolvimento, no entanto este encontra-se desenvolvido pelo que o erro no algoritmo foi justificável tendo em conta que o país estava na fronteira entre um país em desenvolvimento e um país desenvolvido. Isto permitiu provar a eficiência prática do modelo por nós implementado.

Conclusão

Após a conclusão do projeto podemos afirmar que nos sentimos satisfeitos com o resultado e com o nosso desempenho uma vez que cumprimos com os objetivos propostos. Relativamente às dificuldades que foram surgindo ao longo da realização do trabalho estas foram ultrapassadas com sucesso apesar do desconhecimento inicial de algumas tecnologias.

Na elaboração do projeto a parte que consumiu mais tempo foi o desenvolvimento dos *workflows* devido ao desconhecimento inicial da ferramenta *Knime* no entanto foi um trabalho conseguido com bastante eficácia graças ao comprometimento de todos os membros do grupo.

Acabamos assim este trabalho semestral felizes por termos implementado e testado tudo com a obtenção dos resultados que eram esperados e que achamos satisfatórios. Além disso adquirimos conhecimentos que nos irão certamente ajudar no nosso futuro.