Aprendizagem Supervisionada para previsão da popularidade de notícias em plataformas de mídia social em linguagem Python (Tema C1/Grupo 29)

Bruno Vale Fernandes
Mestrado Integrado em
Engenharia Informática e Computação
Faculdade de Engenharia da
Universidade do Porto
Email: up200707284@fe.up.pt

Fábio Oliveira
Mestrado Integrado em
Engenharia Informática e Computação
Faculdade de Engenharia da
Universidade do Porto
Email: up201604796@fe.up.pt

Ricardo Boia
Mestrado Integrado em
Engenharia Informática e Computação
Faculdade de Engenharia da
Universidade do Porto
Email: up201505244@fe.up.pt

***Resumo*—Este artigo tem como objetivo apresentar o trabalho desenvolvido, que consistiu na elaboração de modelos para a previsão da popularidade de notícias em plataformas de redes sociais recorrendo a técnicas de aprendizagem supervisionada. Faz parte do âmbito do artigo expor a implementação do projeto, assim como os modelos de aprendizagem computacional desenvolvidos e a sua respetiva análise.**

***Keywords*— Inteligência Artificial, Aprendizagem Computacional Supervisionada, Regressão Linear e Polinomial, *K–Nearest Neighborhood*, Regressão por Vetores de Suporte, Redes Neuronais**

I. Introdução

A aprendizagem computacional (ou *machine learning*) é uma área da inteligência artificial utilizada por programas para realizar uma tarefa com o mínimo de instruções possível, baseando-se num conjunto de dados para tomar decisões.

As tarefas de aprendizagem são classificadas em três categorias: supervisionada, não supervisionada e por reforço. Neste trabalho, iremos-nos focar a primeira, que consiste na análise de um conjunto de dados previamente fornecidos para determinar a relação entre os dados de entrada e de saída.

O objetivo do projeto é analisar um conjunto de dados que descrevem a popularidade de notícias em redes sociais - ***Google+***, ***Linkedin*** e ***Facebook*** - e desenvolver um modelo para cada uma delas que seja capaz de prever qual será a popularidade de uma notícia publicada na rede social.

Este documento encontra-se dividido nas seguintes secções:

- **Trabalho relacionado -** trabalhos semelhantes e relevantes que foram consultados durante a elaboração do projeto.
- **Trabalho desenvolvido -** um resumo do trabalho que foi desenvolvido e uma breve análise dos resultados.
- **Conclusões e Perspetivas de Desenvolvimento -** sumário do trabalho desenvolvido e perspetivas sobre trabalho futuro neste tema.
- **Referências Bibliográficas**.

II. Trabalho Relacionado

O artigo ***Multi-Source Social Feedback of Online News Feeds***, no qual se enquadra o conjunto de dados sobre o qual o nosso trabalho vai incidir, é de especial interesse. Nele pode ser encontrado uma boa descrição das variáveis que compõe o *dataset*, assim como as razões que tornam a análise da popularidade de notícias em redes sociais apelativa.

No repositório ***News-Popularity-Prediction*** de *Axel Perez*, podem ser encontrados alguns modelos já implementados em *Python* de regressão linear e árvores de decisão para o mesmo conjunto de dados. [1]

No repositório ***newspopularity*** de *Kenelly Rodrigues*, pode ser encontrado uma análise e extração de conhecimento efetuada em *Java* para o mesmo conjunto de dados. [2]

III. Trabalho Desenvolvido

Neste trabalho foram elaborados modelos de *Machine Learning* para a previsão da popularidade de notícias sobre os tópicos Palestina, *Microsoft*, Economia e *Obama* nas redes sociais referidas anteriormente. Um modelo foi desenvolvido para cada rede social e para cada um dos seguintes métodos de aprendizagem supervisionada: regressão linear e polinomial, *K–Nearest Neighborhood*, regressão por vetores de suporte e redes neuronais

Cada modelo tem em conta o tópico da notícia, hora e dia da semana da sua publicação, assim como a pontuação do sentimento do título e do sub-título. Será produzida uma pontuação de popularidade da notícia para a dada rede social na qual o modelo se enquadra. Os modelos são de regressão e não de classificação, uma vez que o domínio da pontuação da popularidade é contínuo.

O trabalho foi desenvolvido em *Python*, recorrendo à ferramenta *Jupyter Notebook*, uma aplicação interativa de criação de documentos que são compostos por uma lista de campos que, por sua vez, podem conter código ou texto formatado em *Markdown*. Foi ainda utilizada a biblioteca *scikit-learn*, que conta com a implementação de muitos algoritmos de aprendizagem computacional, assim como outras funcionalidades

relacionadas com pré-processamento de dados e análise de modelos.

Conjunto de Dados

O conjunto de dados utilizado é da autoria de Nuno Moniz e Luís Torgo e está descrito no artigo ***Multi-Source Social Feedback of Online News Feeds***, elaborado pelos mesmos. [3]

Estão presentes cerca de 93 mil notícias e existem os seguintes campos:

- **Título da notícia**
- **Sub-título da notícia**
- **Fonte da Notícia** (jornal responsável pela publicação da mesma)
- **Tópico** (que pode ser *Obama*, Palestina, Economia ou *Microsoft*)
- **Data de Publicação**
- **Sentimento do Título** (pontuação do sentimento do título; pode varia -1, sentimento negativo, e 1 sentimento positivo)
- **Sentimento do Sub-título** (pontuação do sentimento do sub-título; varia entre -1 e 1)
- **Popularidade nas Redes Sociais** (número de partilhas no *Facebook*, *GooglePlus* e *LinkedIn*)

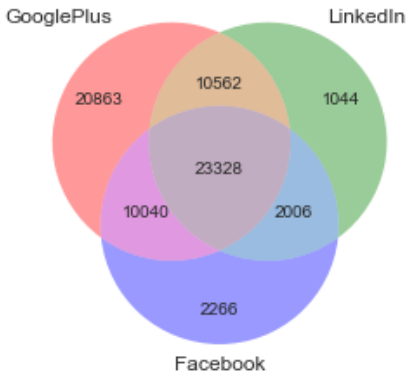

Figura 1: Distribuição das notícias pelas redes sociais.

Pré-Processamento dos Dados

Foram criados 2 novos campos, **Dia da Semana de Publicação** e **Hora de Publicação**, derivados da Data de Publicação de cada notícia. De seguida, foi investigada a hipótese de ser plausível a existência de alguma relação entre a popularidade de uma notícia e a altura em que foi publicada. Para validar a hipótese, foram criados, para cada rede social, gráficos que traduzem a popularidade média da notícia em função do Dia da Semana de Publicação e da Hora de Publicação. Foi confirmado que existia uma correlação entre a popularidade atingida por uma notícia e estes atributos.

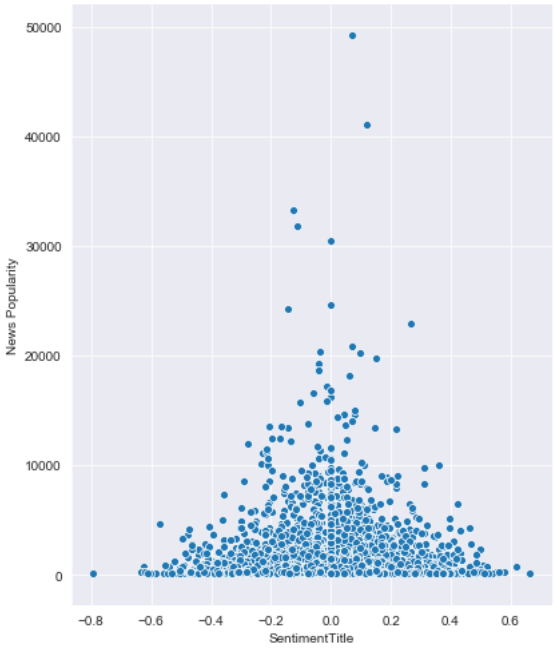

Figura 2: Popularidade das notícias segundo o sentimento do título.

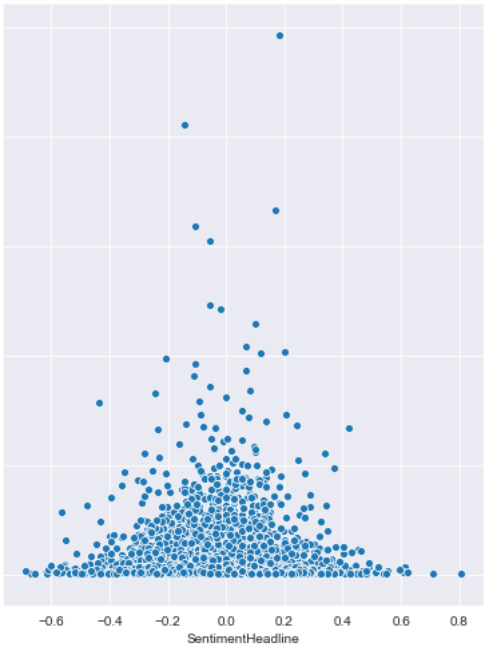

Figura 3: Popularidade das notícias segundo o sentimento do sub-título.

Estes 2 campos foram, de seguida, traduzidos em números inteiros e normalizados para pertencerem ao intervalo entre 0 e 1, de forma a tornar a sua escala mais semelhante à dos outros.

Algoritmos utilizados

Neste projeto, foram utilizados quatro algoritmos de aprendizagem supervisionada para regressão, cujo obje-

tivo é, utilizando um conjunto de dados de entrada, prever qual será o valor de saída, neste caso a popularidade de uma notícia. Segue-se uma breve descrição de cada algoritmo utilizado.

Regressão linear e polinomial

Na regressão polinomial, a relação entre a variável dependente, que irá ser a popularidade das notícias, e as independentes, que poderão ser, o dia e a hora de publicação e o tópico abordado pela notícia, para nomear alguns, é modelada como uma função polinomial de grau maior que um. Já na linear, a relação é modelada como uma função polinomial de grau um. Durante o processo de treino do modelo, são encontrados os coeficientes do polinómio que conduzem ao menor erro do modelo.

K-Nearest Neighbors

Este algoritmo atribui um valor a um elemento, tendo como base os valores dos k elementos de treino mais próximos (mais semelhantes). Primeiro, é calculada a distância euclidiana entre o novo elemento e cada elemento de treino. De seguida, são selecionados os K elementos mais próximos do elemento inserido. Finalmente, é calculada a média desses elementos selecionados, e esse valor será o previsto para o novo elemento.

Regressão por Vetores de Suporte

Este é um algoritmo semelhante ao *Support Vector Machine* (SVM). No entanto, em contraste com o SVM que é utilizado em tarefas de regressão, o SVR é usado para regressão. O objetivo do algoritmo é a criação de uma área centrada num hiperplano, a linha utilizada para separar o conjunto de dados em dois de maneira a utilizar os valores que aí se encontram para efetuar a previsão da variável dependente.

Redes Neuronais

As Redes Neuronais é um algoritmo inspirado no sistema nervoso humano, mais particularmente no cérebro. O processamento dos dados encontra-se distribuído por pequenas unidades interligadas, que se assemelham a, e por essa razão são denominadas, neurónios. A aprendizagem acontece quando é atingida uma solução generalizada nos vários neurónios da rede.

IV. Experiências e Resultados

O conjunto de dados conta com cerca de 93 mil notícias. Estas foram divididas num conjunto de treino dos modelos, onde se encontram 70% das notícias, e num conjunto para avaliar a performance destes mesmo, que é constituído pelos restantes 30%. Foram utilizadas as seguintes medidas durante o processo de avaliação:

- **Raiz Quadrada do Erro Quadrático Médio** (RMSE)
  Avalia o precisão do modelo. É definido como a raiz quadrada da média do erro quadrático.
- **Coeficiente de Determinação** (R^2)
  Avalia a proporção da variância nos dados que o modelo consegue explicar. Varia usualmente entre 0 e 1, podendo tomar valores inferiores a 0 em casos que o modelo explica muito mal os dados.

Regressão linear e polinomial

Os modelos desenvolvidos para a regressão linear e polinomial demonstraram resultados (constantes na Tabela I, no anexo) razoáveis quando comparados com os outros modelos. No geral, o erro diminuiu com o aumento do grau do polinómio. No entanto isso não se verificou para o grau 4, tendo o erro aumentado e o coeficiente de determinação diminuído para os modelos de todas as redes sociais.

K-Nearest Neighbors

Os resultados dos modelos para este algoritmo encontram-se nas figuras 4, 5 e 6 do anexo. Como seria de esperar, o erro diminuiu e o coeficiente de determinação aumentou em função do parâmetro K. No entanto, não se deve concluir que o aumento de K torna o modelo mais preciso, uma vez que ocorre o fenómeno de *overfitting*, isto é, o modelo pode estar muito bem ajustado ao conjunto de dados utilizado, mas pode não ser generalizável para dados nunca testados.

Regressão por Vetores de Suporte

Os resultados dos modelos para este algoritmo encontram-se na Tabela II do anexo. O facto do coeficiente de determinação ser inferior a 0 para estes modelos revela que eles explicam muito mal os dados. O erro dos modelos é maior que os modelos dos restantes algoritmos. Este método de aprendizagem não é adequado para a tarefa de prever a popularidade de notícias neste conjunto de dados.

Redes Neuronais

Foram testados vários modelos para cada rede social com diferente número de neurónios nas camadas intermediárias. Os resultados obtidos encontram-se nas figuras 7, 8, e 9 do anexo. A função de ativação utilizada foi a *relu*.

O erro diminuiu e o coeficiente de determinação aumentou em função do número de iterações (*epochs*) na fase de treino do modelo. Neste caso, foram realizadas 100. Para os modelos do *Facebook* não parece que este número tenha sido adequado porque o erro ainda estava a diminuir nas últimas iterações. No entanto, para os restantes modelos parece que foi suficiente porque o valor do erro e do coeficiente de determinação convergiram.

Não se pode concluir, tendo em conta os dados, que exista um número ideal de neurónios, uma vez que para cada rede social os resultados obtidos foram drasticamente diferentes.

V. Conclusões e perspetivas de desenvolvimento

Foram desenvolvidos vários modelos para prever a popularidade de notícias em redes sociais. Os métodos de aprendizagem supervisionada utilizados foram Regressão Linear e Polinomial, *K-Nearest Neighbors*, Regressão por Vetores de Suporte e Redes Neuronais.

Os resultados obtidos foram satisfatórios. No entanto, reconhece-se que existe grande espaço para melhorar os mo-

delos, uma vez que o coeficiente de determinação é próximo do 0 para todos eles.

Os melhores modelos obtidos, tendo em conta o erro e o coeficiente de determinação, foram os de Redes Neuronais, apesar dos modelos dos outros algoritmos estarem bastantes próximos.

Trabalho futuro sobre este tema poderá incidir sobre o desenvolvimento de novos modelos, tendo em conta outros algoritmos como regressão *Ridge* e *Lasso*, aprofundamento da exploração de Redes Neuronais (testando com diferentes funções de ativação e diferente número de camadas intermediárias e neurónios), e a re-avaliação dos modelos apresentados recorrendo a validação cruzada.

REFERÊNCIAS

[1] Prediction of News Popularity in Multiple Social Media Platforms by *axperez*. Available at: `https://github.com/axperez/News-Popularity-Prediction` Last accessed 24 May 2019

[2] newspopularity by *kenellysra*. Available at: `https://github.com/kenellysra/newspopularity` Last accessed 24 May 2019

[3] Nuno Moniz and Luís Torgo (2018), *Multi-Source Social Feedback of Online News Feeds*.

Anexos

Tabela I: Resultados dos Modelos de Regressão Linear (grau 1) e Polinomial (grau 2, 3 e 4)

Grau	Rede Social	Raiz Quadrada do Erro Quadrático Médio (RMSE)	Coeficiente de Determinação (R^2)
1	Facebook	664.101	0.013
	GooglePlus	20.399	0.005
	Linkedin	98.318	0.001
2	Facebook	662.407	0.018
	GooglePlus	20.336	0.012
	Linkedin	98.241	0.002
3	Facebook	656.796	0.035
	GooglePlus	20.322	0.013
	Linkedin	98.159	0.004
4	Facebook	657.055	0.034
	GooglePlus	20.328	0.012
	Linkedin	98.307	0.001

Tabela II: Resultados dos Modelos de Regressão por Vetores de Suporte (*kernel rbf*)

Rede Social	Raiz Quadrada do Erro Quadrático Médio (RMSE)	Coeficiente de Determinação (R^2)
Facebook	675.860	-0.021
GooglePlus	20.761	-0.039
Linkedin	95.545	-0.010

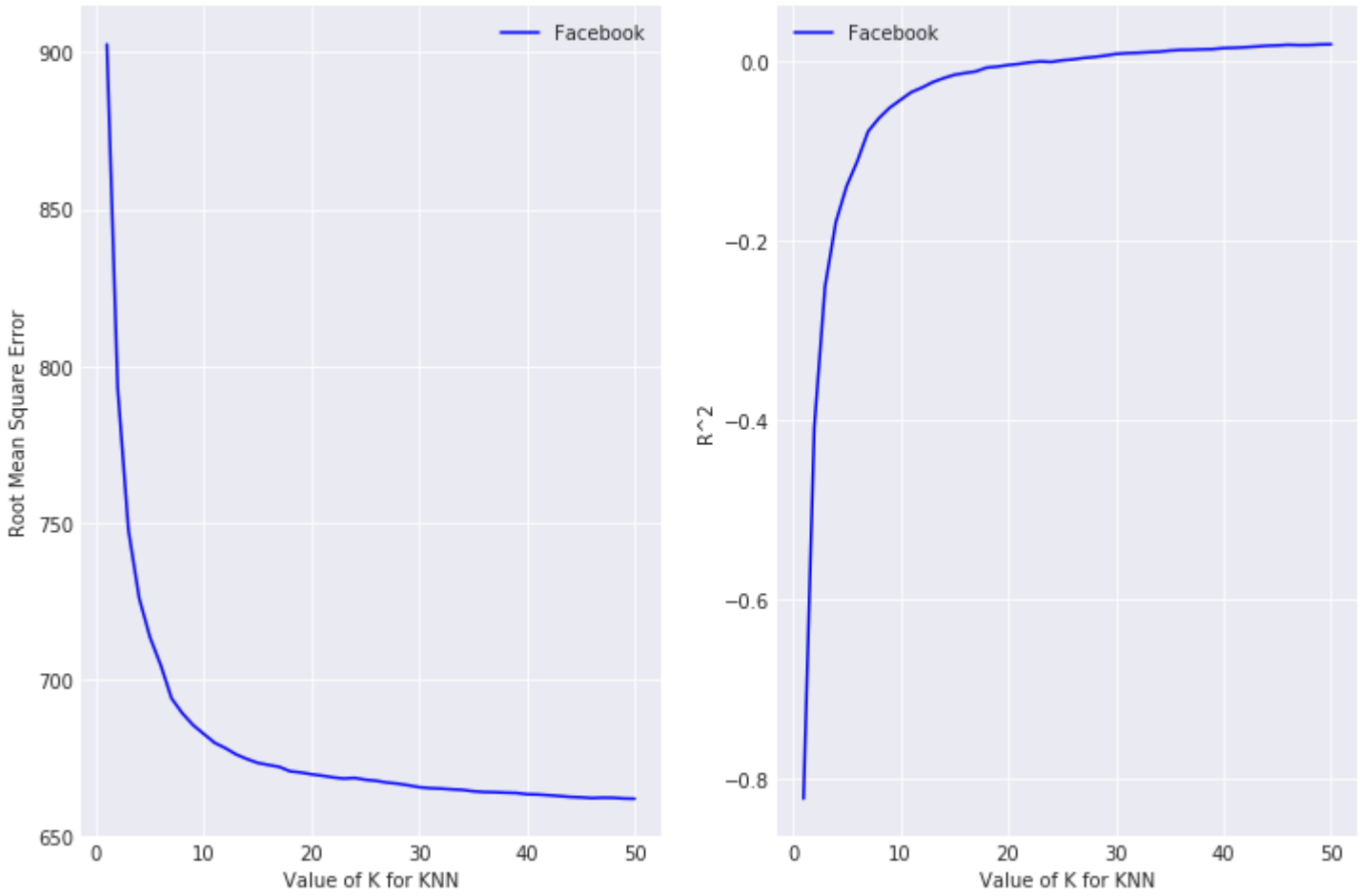

Figura 4: Resultados do Modelo de *KNN* para o Facebook em função do valor de K.

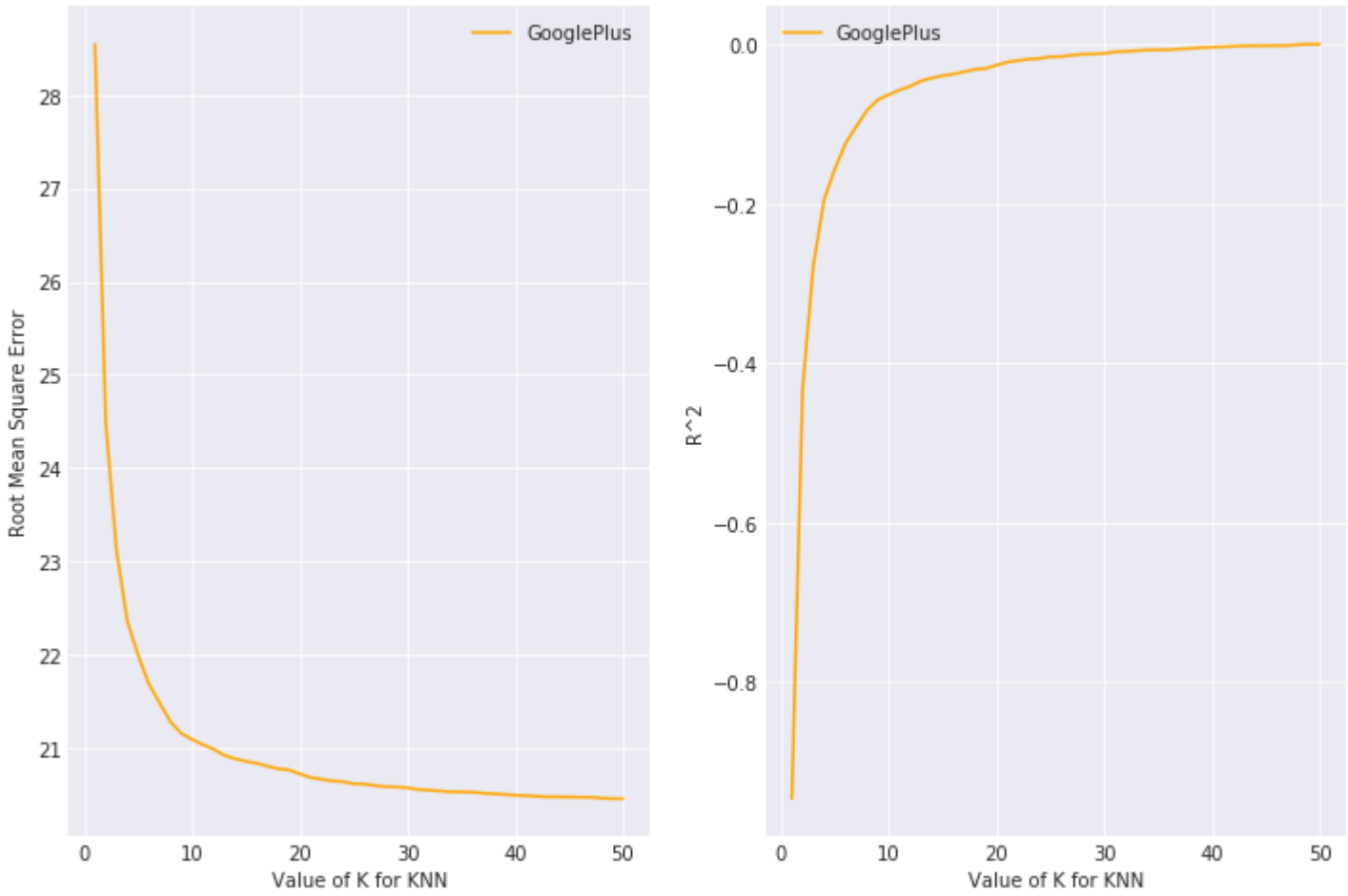

Figura 5: Resultados do Modelo de *KNN* para o GooglePlus em função do valor de K.

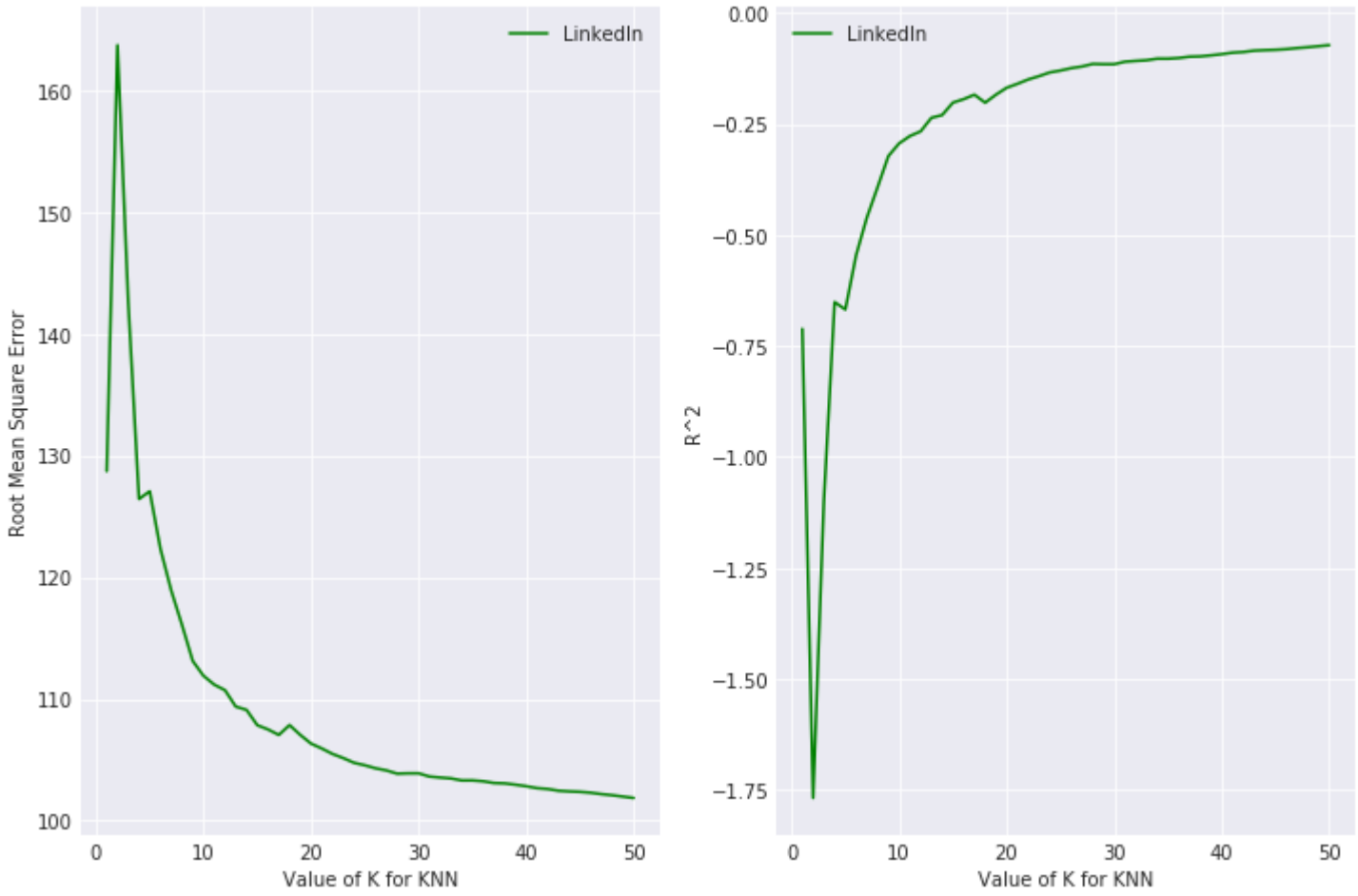

Figura 6: Resultados do Modelo de *KNN* para o LinkedIn em função do valor de K.

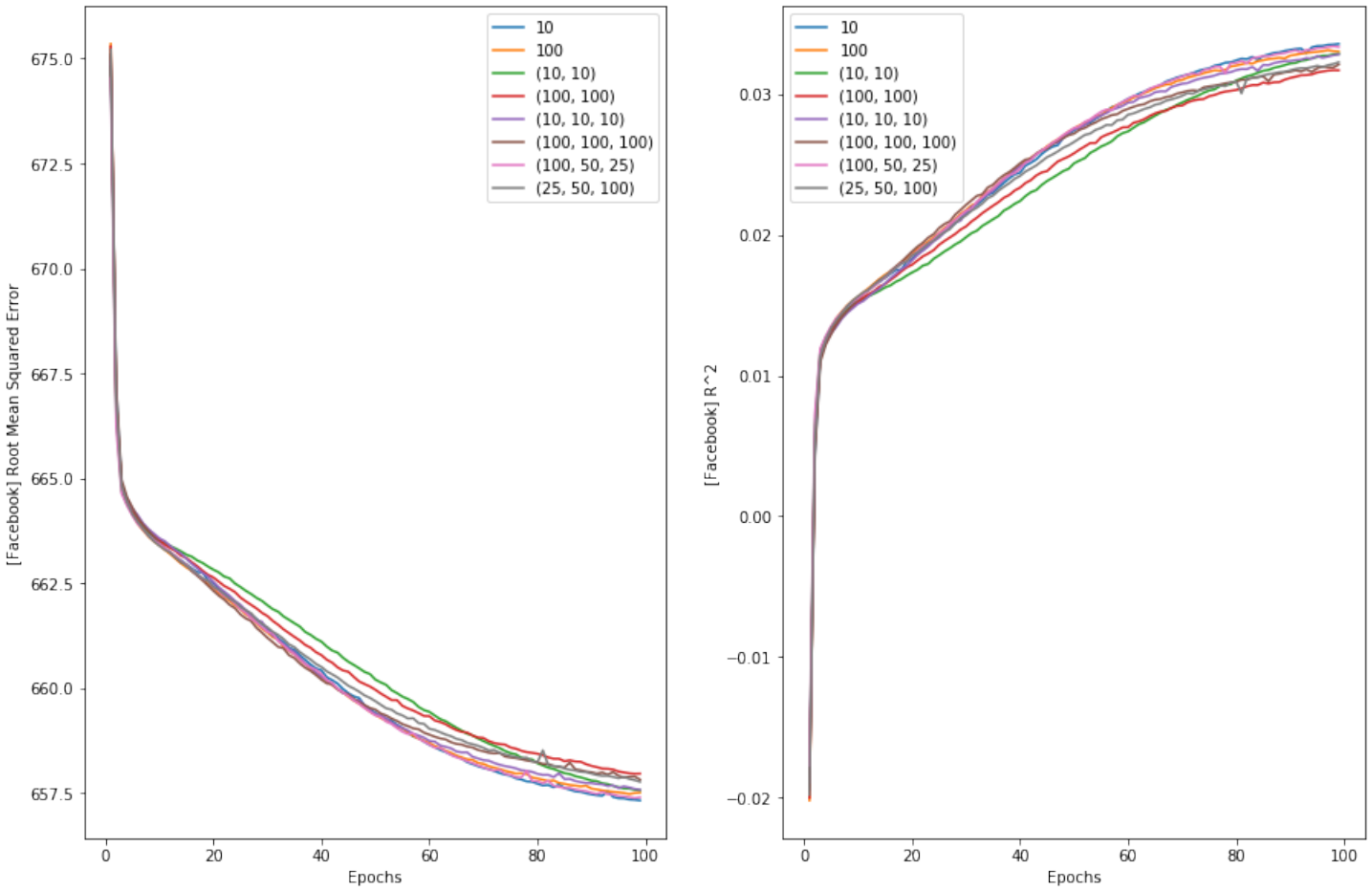

Figura 7: Resultados dos Modelos de Redes Neuronais para o Facebook em função do número de camadas e neurónios intermediários.

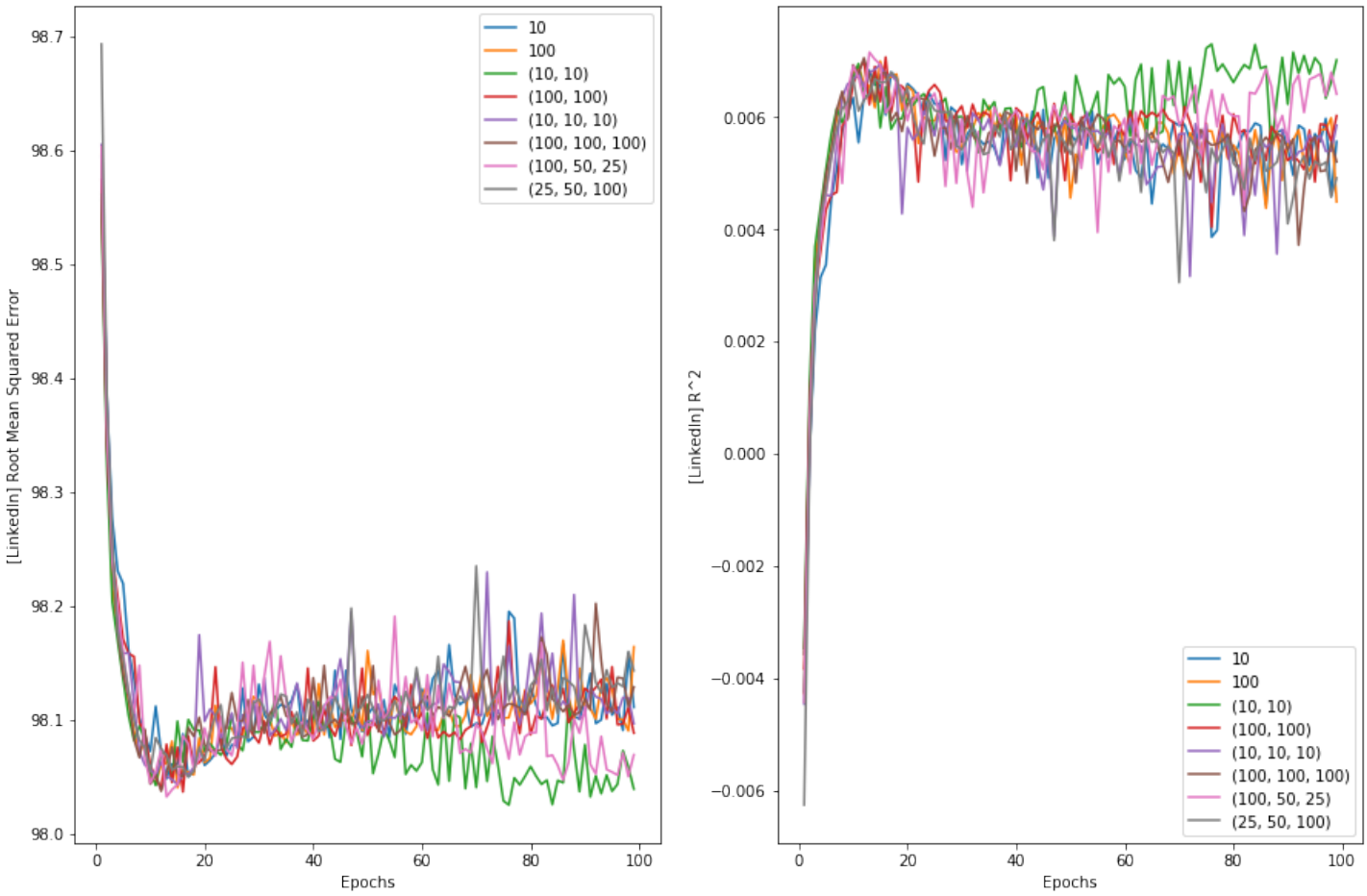

Figura 8: Resultados dos Modelos de Redes Neuronais para o GooglePlus em função do número de camadas e neurónios intermediários.

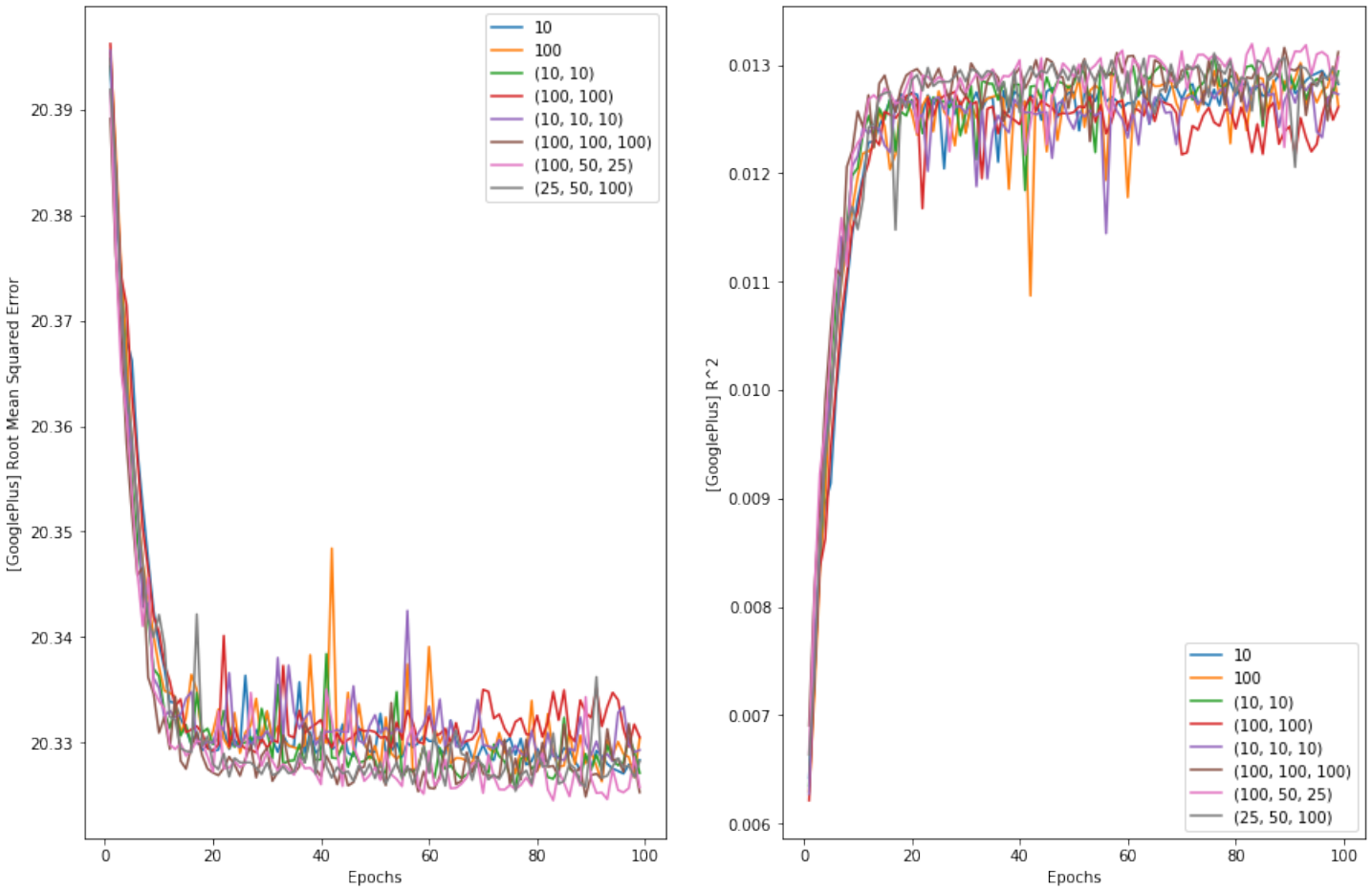

Figura 9: Resultados dos Modelos de Redes Neuronais para o LinkedIn em função do número de camadas e neurónios intermediários.