Aplicação de Técnicas de Classificação e Regressão para Análise de Preços Habitacionais em Portugal

Application of Classification and Regression Techniques for Housing Price Analysis in Portugal

Inês Nascimento
Rita Dias
Escola Superior de Tecnologia e Gestão
Instituto Politécnico de Beja
Beja, Portugal
secretariado.estig@ipbeja.pt

Resumo — **O mercado habitacional português tem registado uma evolução significativa nos últimos anos, com fortes assimetrias regionais e crescente pressão sobre os preços. Neste trabalho, aplicam-se técnicas de Machine Learning a dados oficiais do Instituto Nacional de Estatística (INE), com o objetivo de analisar e prever o preço da habitação por metro quadrado (€/m²) ao nível das regiões NUTS III. Inicialmente, o problema é formulado como uma tarefa de regressão, tratando o preço como uma variável contínua, recorrendo aos algoritmos Regressão Linear e k-Nearest Neighbors (kNN). Posteriormente, o preço é discretizado em três classes (baixa, média e alta), permitindo reformular o problema como uma tarefa de classificação e avaliar modelos como Random Forest, Decision Tree e kNN. A metodologia seguida enquadra-se no processo Knowledge Discovery in Databases (KDD), incluindo etapas de seleção, pré-processamento, transformação, modelação e validação dos dados. Os resultados demonstram que o kNN apresenta o melhor desempenho na tarefa de regressão, com erros significativamente mais baixos e elevado coeficiente de determinação. Na classificação, o Random Forest destaca-se como o modelo mais robusto, apresentando valores elevados de F1, precisão e recall. O estudo evidencia o potencial das técnicas de Machine Learning como ferramentas de apoio à análise e segmentação do mercado habitacional português.**

Palavras Chave - Machine Learning; Preços da Habitação; Regressão; Classificação; KDD; INE; Mercado Imobiliário; NUTS III

Abstract — ***The Portuguese residential housing market has experienced significant changes in recent years, with strong regional disparities and increasing price pressure. This work applies Machine Learning techniques to official data from Portuguese Statistics Institute (INE), aiming to analyse and predict housing prices per square metre (€/m²) at the NUTS III regional level. The problem is initially addressed as a regression task, treating price as a continuous variable and applying Linear Regression and k-Nearest Neighbors (kNN) algorithms. Subsequently, prices are discretised into three classes (low, medium and high), reformulating the problem as a classification task and enabling the evaluation of models such as Random Forest, Decision Tree and kNN. The adopted methodology follows the Knowledge Discovery in Databases (KDD) process, including data selection, preprocessing, transformation, modelling and validation. The results show that kNN achieves superior performance in the regression task, with lower error metrics and a high coefficient of determination. In the classification task, Random Forest emerges as the most robust model, achieving high F1, precision and recall values. Overall, the study highlights the importance of Machine Learning techniques as effective tools for analysing and segmenting the Portuguese housing market.***

Keywords - Machine Learning; Housing Prices; Regression; Classification; KDD; Official Statistics; Real Estate Market; NUTS III

I. INTRODUÇÃO

O mercado imobiliário residencial desempenha um papel central na economia, não apenas pela sua influência no investimento e no rendimento das famílias, mas também pelo impacto que pode exercer sobre a estabilidade financeira e o crescimento económico. A habitação representa uma parte significativa da riqueza e do consumo das famílias, sendo simultaneamente um bem de uso e de investimento [1]. Estes efeitos são particularmente relevantes em economias desenvolvidas, como a portuguesa, onde o setor imobiliário é um dos principais indicadores macroeconómicos e financeiros [1].

Em Portugal, a evolução recente dos preços da habitação tem sido influenciada por diversos fatores, como o turismo, o investimento estrangeiro e mudanças nas condições de crédito, contribuindo para uma crescente pressão no acesso à habitação em várias regiões do país. A variabilidade dos preços entre as diferentes NUTS III evidencia a necessidade de analisar e compreender as dinâmicas territoriais do mercado [1].

A avaliação imobiliária tem sido feita sobretudo com análises manuais e com base na experiência dos profissionais, bem como com modelos estatísticos tradicionais. No entanto, estes métodos podem ter limitações, como dificuldade em escalar para muitos imóveis, alguma subjetividade e pouca capacidade para identificar relações mais complexas entre os vários fatores que influenciam o valor de um imóvel [2].

Com o crescimento da disponibilidade de dados e de ferramentas computacionais, os métodos de Machine Learning (ML) emergiram como uma alternativa robusta e eficiente, apresentando melhorias comprovadas na previsão de preços imobiliários e permitindo uma análise mais completa dos fatores que determinam o valor dos imóveis [3].

Este trabalho utiliza dados oficiais do Instituto Nacional de Estatística (INE) sobre preços de habitação por metro quadrado (€/m²), uma variável de natureza contínua [4]. Deste modo, a primeira abordagem adotada consiste na aplicação de algoritmos de regressão, nomeadamente *Linear Regression* e *k-Nearest Neighbors (kNN)*, com o objetivo de prever o valor do imóvel em função das características disponíveis.

Contudo, para além da previsão direta do preço, torna-se pertinente analisar o mercado habitacional numa perspetiva segmentada, uma vez que diferentes níveis de valorização podem refletir realidades socioeconómicas distintas e apoiar decisões de investimento mais informadas. Assim, procedeu-se à discretização do atributo preço em três classes: baixa, média e alta, através da divisão do intervalo de valores em partes iguais. Esta transformação permitiu reformular o problema como uma tarefa de classificação, possibilitando a avaliação de algoritmos como *Random Forest*, *Decision Tree* e *kNN* na identificação de padrões e distinção entre segmentos do mercado.

Em ambas as abordagens, os modelos desenvolvidos foram avaliados com recurso às métricas adequadas a cada tipo de problema, tais como R^2, RMSE e MAE para a tarefa de regressão e Recall, F1, matriz de confusão e análise ROC para a tarefa de classificação. O processo completo seguiu as etapas do ciclo KDD, incluindo seleção, preparação, modelação e validação dos dados, garantindo coerência metodológica e qualidade nos resultados obtidos.

O artigo encontra-se estruturado da seguinte forma: a Secção II apresenta a revisão de literatura; a Secção III descreve a metodologia adotada, enquadrada no processo KDD; a Secção IV apresenta os resultados obtidos; a Secção V discute as conclusões.

II. Revisão de Literatura

A. Avaliação Imobiliária e Modelos Tradicionais

A literatura destaca que o valor da habitação é influenciado simultaneamente por características físicas, localização e condições macroeconómicas. Métodos tradicionais de avaliação incluem o método comparativo e o método do rendimento, ambos amplamente utilizados no contexto europeu e português [2].

Estudos sobre o mercado português apontam ainda a importância de fatores regionais e assimetrias espaciais, reforçando a necessidade de considerar estruturas territoriais como as NUTS III na modelação de preços [1].

B. Machine Learning na Previsão de Preços Habitacionais

O recurso a algoritmos de ML tem ganho destaque devido à sua capacidade de identificar relações entre múltiplas variáveis e de processar grandes volumes de dados. A investigação internacional demonstra ganhos significativos de performance quando comparados com modelos estatísticos convencionais [3].

Modelos baseados em árvores, como Random Forest e Gradient Boosting, têm apresentado resultados particularmente competitivos na previsão de valores imobiliários [5][6].

C. Estudos Aplicados ao Mercado Português

Em Portugal, tem surgido investigação que recorre a dados mais granulares e atualizados para melhorar a previsão de preços imobiliários. Num estudo recente, vários modelos de machine learning - como Decision Trees, K-Nearest Neighbors, Random Forest, Gradient Boosting, Support Vector Machines e XGBoost foram comparados na estimação do valor de venda de imóveis. Entre estes, os modelos de ensemble, particularmente o XGBoost e o CatBoost, demonstraram o melhor desempenho ao apresentarem valores inferiores de MAE e RMSE (indicadores de erro, onde valores mais baixos representam maior precisão), evidenciando maior capacidade para captar relações complexas no mercado habitacional português [5].

D. KDD e Métodos de Descoberta de Conhecimento

O processo KDD: Knowledge Discovery in Databases integra etapas de seleção, pré-processamento, transformação de dados, modelação e validação de resultados, garantindo rigor na transformação de dados em conhecimento útil [7].

A adoção deste processo em trabalhos de data mining imobiliário contribui para análises sistemáticas, replicáveis e transparentes, reduzindo riscos de enviesamento e garantindo que os resultados são interpretáveis e relevantes para os decisores.

III. Metodologia

A metodologia adotada neste estudo segue então o processo KDD: Knowledge Discovery in Databases, que estabelece um conjunto estruturado de etapas para a transformação de dados em conhecimento útil, garantindo rigor, reprodutibilidade e qualidade analítica [7].

O desenvolvimento do trabalho integrou duas abordagens complementares: por um lado, as etapas de pré-processamento e transformação dos dados foram realizadas em Python, recorrendo à biblioteca pandas e ao ambiente Jupyter Notebook, pela sua eficiência e flexibilidade; por outro lado, a modelação, avaliação e validação dos modelos foram conduzidas no software Orange, cuja interface visual orientada a workflows facilita a experimentação e a comparação de diferentes algoritmos [8][9][10].

A. Processo KDD

O processo KDD aplicado inclui as etapas:

1. Seleção dos Dados

   Foram utilizados dados oficiais do Instituto Nacional de Estatística (INE), disponibilizados em formato Excel, contendo informação sobre preços da habitação por metro

quadrado. A análise foi restringida ao nível territorial NUTS III, assegurando coerência espacial e permitindo a comparação regional dos valores imobiliários [4].

2. Pré-processamento e Limpeza dos Dados

O pré-processamento foi integralmente realizado em Python com recurso à biblioteca *pandas*. Inicialmente, procedeu-se à remoção de linhas completamente vazias, garantindo que apenas observações com informação relevante fossem consideradas. De seguida, foram eliminadas linhas correspondentes a metadados e cabeçalhos não estruturados, bem como colunas que continham exclusivamente valores nulos.

Posteriormente, os cabeçalhos do dataset foram reconstruídos a partir de múltiplas linhas, combinando informação temporal e categorial num único identificador por coluna. Esta operação permitiu transformar a estrutura original do ficheiro, que não se encontrava diretamente adequada a tarefas de análise e modelação.

Adicionalmente, a coluna que continha simultaneamente o código e o nome da região foi desagregada em dois atributos distintos, RegiaoCodigo e RegiaoNome, garantindo uma representação mais limpa e normalizada da informação geográfica.

3. Transformação dos Dados

Após a limpeza, o dataset foi transformado de formato largo (wide) para formato longo (long) através da operação *melt*, permitindo que cada observação representasse um único registo regional, temporal e categorial com o respetivo preço por metro quadrado.

Para além das variáveis Ano e Trimestre, foi criada uma variável temporal composta designada AnoTrimestre, no formato *AAAA.T* (por exemplo, 2020.1, 2020.2, 2020.3, 2020.4). O objetivo desta transformação foi representar explicitamente a evolução temporal numa única variável ordinal e contínua, permitindo capturar a noção de progressão do tempo.

Esta abordagem revelou-se particularmente relevante no contexto da modelação em Orange, uma vez que a ferramenta não disponibiliza mecanismos diretos para especificar relações de ordem temporal ou tendências crescentes entre observações. Assim, a variável AnoTrimestre permite incorporar implicitamente a informação de que o tempo é uma dimensão crescente, assumindo-se que, em termos gerais, a evolução temporal está associada a um aumento progressivo do valor da habitação em euros por metro quadrado. Deste modo, esta variável contribui para que os modelos de Machine Learning consigam captar tendências temporais subjacentes aos dados

As categorias associadas ao tipo de habitação foram igualmente normalizadas, distinguindo imóveis novos e existentes, bem como a categoria total.

O resultado deste processo foi um dataset estruturado, limpo e consistente, adequado à aplicação de algoritmos de Machine Learning, posteriormente exportado para um novo ficheiro para utilização na fase de modelação.

4. Modelos de Machine Learning

Dado que a variável alvo corresponde ao preço por metro quadrado (€/m²), uma variável contínua, o problema foi inicialmente formulado como uma tarefa de regressão. Para esta finalidade foram utilizados os algoritmos *Linear Regression* e *k-Nearest Neighbors (kNN)*.

Numa segunda etapa, procedeu-se à discretização do preço em três classes (baixa, média e alta) através da divisão do intervalo de valores em partes iguais. Esta transformação permitiu reformular o problema como uma tarefa de classificação, possibilitando a análise da capacidade dos modelos em distinguir diferentes segmentos do mercado habitacional.

5. Avaliação e Validação

A avaliação dos modelos foi realizada no software Orange através do módulo *Test & Score*, recorrendo a cross validation e a métricas adequadas à natureza de cada tarefa.

B. Ferramentas Utilizadas

1. Python e Pandas

A opção por Python deveu-se à sua eficiência no tratamento de dados, rapidez de execução e elevado suporte em bibliotecas orientadas a análise de dados [8].

2. Orange Data Mining

A escolha do Orange assenta na sua capacidade para modelar visualmente fluxos analíticos, aplicar algoritmos de machine learning, apresentar resultados de forma imediata e adaptar-se a análises exploratórias rápidas. Estudos comparativos destacam-no como uma ferramenta eficaz para tarefas de classificação e regressão, acessível a utilizadores não especialistas e suportada pela linguagem Python [9][10].
Neste trabalho foram estruturados dois workflows: um dedicado à regressão, recorrendo a Linear Regression e kNN ligados ao módulo Test & Score e Scatter Plot, e outro orientado à classificação, utilizando Random Forest, kNN e Decision Tree, complementados com Test and Score, Confusion Matrix e ROC Analysis.

3. Modelos de Machine Learning

A tarefa de regressão recorreu aos algoritmos Linear Regression e kNN, uma vez que a variável alvo é contínua e estes modelos permitem avaliar diferentes níveis de complexidade na relação entre os atributos e o preço dos imóveis.

Na tarefa de classificação foram utilizados Random Forest, kNN e Decision Tree, possibilitando uma comparação entre técnicas baseadas em vizinhança e abordagens assentes em árvores de decisão. A inclusão do Random Forest é particularmente pertinente neste domínio, dada a sua robustez em problemas imobiliários e a capacidade de captar relações não lineares sem exigir uma parametrização extensa [3][5].

4. Segmentação do Mercado (Classificação)

A segmentação do mercado foi obtida através da discretização automática do preço por metro quadrado (€/m²) em três classes com intervalos iguais, com base na distribuição dos valores observados no dataset. O processo resultou nos seguintes limiares de segmentação:

- Classe baixa: valores inferiores a 853,5 €/m²
- Classe média: valores entre 853,5 €/m² e 1196 €/m²
- Classe alta: valores iguais ou superiores a 1196 €/m²

Esta discretização permitiu distinguir regiões com diferentes níveis de valorização imobiliária, correspondendo, respetivamente, a zonas de menor acessibilidade, oferta intermédia e mercados com preços mais elevados. A transformação foi realizada diretamente no Orange com recurso ao widget *Discretize*, assegurando consistência entre os dados transformados e os modelos de classificação utilizados na fase de modelação [11].

5. Validação dos Modelos

As avaliações dos modelos seguiram métricas distintas conforme a natureza da tarefa: na regressão foram utilizados o MAE, o RMSE e o R^2, enquanto na classificação se recorreu à Precision, Recall, F1, matriz de confusão e análise ROC. Estes indicadores permitem comparar o desempenho, a interpretabilidade e a capacidade de generalização dos diferentes algoritmos aplicados [3][5].

IV. DISCUSSÃO DOS RESULTADOS

Nesta secção discutem-se os resultados obtidos para os modelos de regressão e de classificação aplicados ao conjunto de dados de preços da habitação, relacionando o desempenho quantitativo com a sua utilidade prática para apoio à decisão.

Numa primeira fase experimental, a modelação foi realizada utilizando como variáveis de entrada RegiaoNome, TipoCategoria, Ano e Trimestre, tendo como variável alvo o Preço por metro quadrado (€/m²). Esta abordagem permitiu avaliar o comportamento inicial dos algoritmos de regressão e classificação com base na informação temporal tal como disponibilizada nos dados originais.

No entanto, verificou-se que a utilização das variáveis Ano e Trimestre em separado introduz uma limitação metodológica, uma vez que o Orange trata estes atributos como independentes, não reconhecendo explicitamente a sua relação temporal nem a natureza sequencial e crescente dos períodos observados. Como consequência, a progressão temporal entre trimestres consecutivos não é diretamente capturada pelos modelos.

Com o objetivo de diminuir esta limitação, foi criada uma variável temporal composta, designada AnoTrimestre (por exemplo, 2020.1, 2020.2, 2020.3 e 2020.4), que agrega a informação do ano e do trimestre numa única variável numérica crescente. A substituição de Ano e Trimestre por AnoTrimestre, mantendo como restantes variáveis de entrada RegiaoNome e TipoCategoria, conduziu a resultados ligeiramente superiores de forma consistente para os diferentes algoritmos testados.

Face a esta melhoria sistemática de desempenho, foi tomada a decisão de adotar a variável composta AnoTrimestre como representação temporal definitiva na modelação. Assim, a análise detalhada dos resultados apresentada na secção seguinte baseia-se exclusivamente nesta configuração, garantindo maior coerência temporal e melhor capacidade discriminativa dos modelos de classificação.

A. Regressão

Na tarefa de regressão, o objetivo foi prever diretamente o preço por metro quadrado (€/m²) como variável contínua. Foram avaliados dois modelos: Regressão Linear e kNN, utilizando como variáveis de entrada RegiaoNome, TipoCategoria e AnoTrimestre, e como variável alvo o Preço (€/m²).

A avaliação dos modelos de regressão foi efetuada através de validação cruzada com 10 folds, uma prática amplamente reconhecida como standard na literatura por proporcionar um equilíbrio adequado entre robustez estatística e eficiência computacional [12]. Este método permite obter estimativas mais fiáveis do desempenho dos modelos, reduzindo a influência de uma única divisão treino–teste.

De forma complementar, foram testadas diferentes configurações de validação cruzada (3, 5 e 20 folds), tendo-se observado que a utilização de 10 folds conduziu a resultados ligeiramente melhores e mais consistentes em termos das métricas de erro e do coeficiente de determinação. Assim, esta configuração foi selecionada como abordagem definitiva para a avaliação dos modelos de regressão.

Os resultados quantitativos evidenciam um desempenho claramente superior do modelo kNN. Conforme apresentado na Tabela II, o kNN obteve um MSE de 4262.409, RMSE de 65.287, MAE de 35.829 e MAPE de 3.078, com um R^2 de 0.987. Em contraste, a Regressão Linear apresentou um MSE de 23939.421, RMSE de 154.724, MAE de 110.985, MAPE de 11.010 e um R^2 de 0.926, indicando erros significativamente mais elevados e menor capacidade explicativa. Estes valores podem ser observados na tabela I.

TABLE I. TABELA DE MÉTRICAS DE REGRESSÃO

Model	MSE	RMSE	MAE	MAPE	R2
kNN	4262.409	65.287	35.829	3.078	0.987
Linear Regression	23939.421	154.724	110.985	11.010	0.926

A análise gráfica dos valores reais versus valores previstos reforça estas diferenças. A Figura 1, correspondente ao modelo kNN, mostra uma forte concentração dos pontos ao longo da diagonal, com coeficiente de correlação próximo de r ≈ 0.99, evidenciando elevada precisão ao longo de praticamente todo o intervalo de preços. Por outro lado, a Figura 2, relativa à Regressão Linear, apresenta maior dispersão em torno da diagonal, sobretudo para valores de preço mais elevados, revelando limitações na modelação de relações não lineares presentes nos dados.

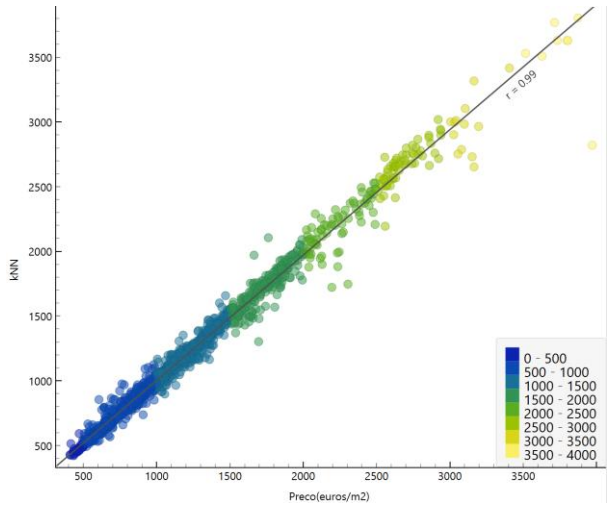

Figure 1. Scatter Plot do modelo kNN

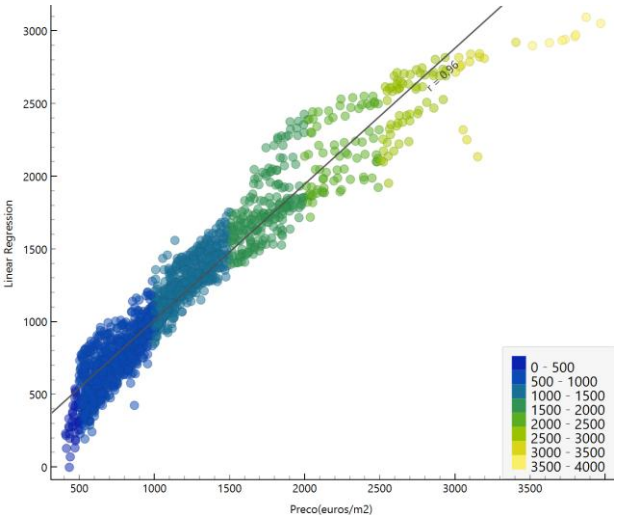

Figure 2. Scatter Plot do modelo regressão linear

Estes resultados mostram que o kNN se adapta melhor aos dados, conseguindo representar com maior precisão a relação entre as variáveis e o preço por metro quadrado. A Regressão Linear, embora capture a tendência geral, apresenta erros mais elevados, sobretudo para valores extremos. Assim, o kNN revela-se o modelo mais adequado para a tarefa de regressão neste estudo.

B. Classificação

A tarefa de classificação do preço por metro quadrado em três classes (baixo, intermédio e elevado) apresentou resultados globalmente muito elevados para todos os algoritmos testados (kNN, Árvore de Decisão e Random Forest), o que evidencia a adequação da discretização do preço em €/m² para suportar modelos de classificação com elevado desempenho.

A avaliação dos modelos de classificação foi realizada através de validação cruzada estratificada com 10 folds, uma abordagem amplamente adotada na literatura como configuração standard em tarefas de Machine Learning, por representar um compromisso equilibrado entre viés, variância e custo computacional [12]. Este procedimento garante uma avaliação mais robusta e reduz a dependência de uma única divisão treino–teste, assegurando simultaneamente uma distribuição equilibrada das classes em cada iteração.

Adicionalmente, foram realizados testes preliminares com diferentes números de folds (3, 5 e 20), tendo-se verificado que a configuração com 10 folds apresentou resultados ligeiramente superiores e mais estáveis ao nível das métricas de classificação. Com base nesta análise, a validação cruzada estratificada com 10 folds foi adotada como abordagem final para a tarefa de classificação.

As métricas F1, precisão e recall apresentam valores elevados para todos os modelos, indicando um desempenho consistente na classificação das três classes de preço. O Random Forest destaca-se com valores de 0,943 nas três métricas, superando ligeiramente a Árvore de Decisão e o kNN, cujos resultados se situam em torno de 0,93. Esta diferença, embora reduzida, evidencia uma melhor capacidade do Random Forest em equilibrar corretamente a identificação das classes, justificando a sua escolha como modelo de referência para a tarefa de classificação. Os valores das métricas de classificação do Test and Score podem ser observados na Tabela II.

TABLE II. TABELA DE MÉTRICAS DE CLASSIFICAÇÃO

Model	AUC	CA	F1	Prec	Recall	MCC
kNN	0.988	0.929	0.929	0.930	0.929	0.894
Random Forest	0.990	0.943	0.943	0.943	0.943	0.915
Tree	0.960	0.930	0.930	0.930	0.930	0.896

As matrizes de confusão mostram que o Random Forest comete menos erros globais e apresenta melhor separação da

classe intermédia. O modelo Tree e o kNN mantêm desempenhos elevados, mas com maior confusão entre classes adjacentes, sobretudo entre os intervalos intermédio e elevado. As matrizes de confusão do modelo Random Forest, kNN e Tree podem ser observadas nas figuras 3, 4 e 5, respetivamente.

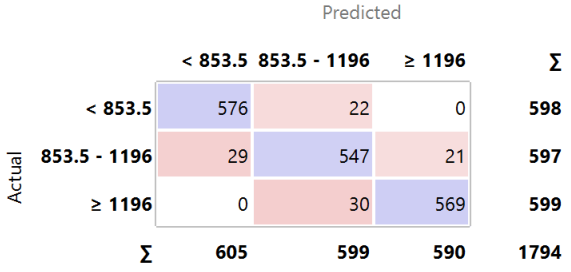

Figure 3. Matriz de Confusão do modelo Random Forest

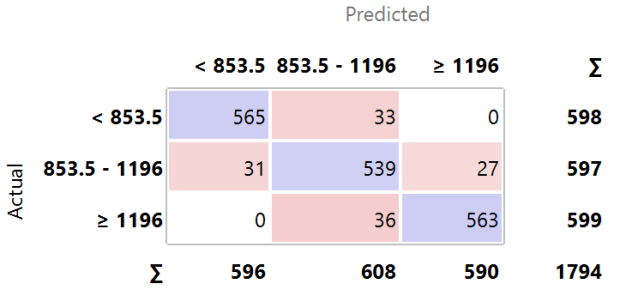

Figure 4. Matriz de Confusão do modelo kNN

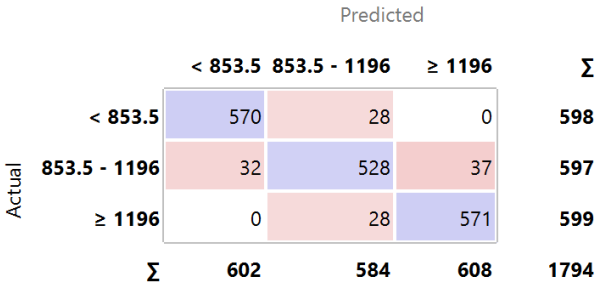

Figure 5. Matriz de Confusão do modelo Tree

A Figura 6 apresenta as curvas ROC para a classe intermédia (853.5–1196 €/m²). A linha verde representa o modelo Random Forest, a linha castanha o modelo kNN e a linha azul o modelo Tree. Os três modelos exibem curvas próximas do canto superior esquerdo, evidenciando elevada capacidade discriminativa. O Random Forest apresenta a curva globalmente dominante, seguido de muito perto pelo kNN, enquanto o modelo Tree revela um desempenho ligeiramente inferior. Ainda assim, todos os modelos superam claramente o comportamento aleatório, confirmando a robustez da tarefa de classificação.

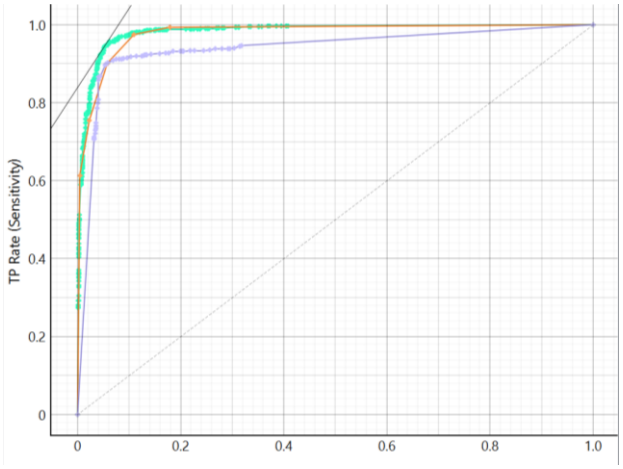

Figure 6. Matriz de Confusão do modelo kNN

V. Conclusões

Neste trabalho foi realizada uma análise do mercado habitacional português com base em dados oficiais do INE, recorrendo a técnicas de Machine Learning para abordar o problema sob duas perspetivas complementares: regressão do preço por metro quadrado e classificação do mercado em diferentes faixas de preço. A adoção do processo KDD permitiu estruturar de forma sistemática todas as fases do estudo, desde o tratamento dos dados até à avaliação dos modelos.

Na tarefa de regressão, os resultados demonstraram que o modelo k-Nearest Neighbors apresenta um desempenho significativamente superior à Regressão Linear, evidenciado por valores mais baixos de erro e por um elevado coeficiente de determinação. Na tarefa de classificação, o Random Forest destacou-se como o modelo mais robusto, alcançando elevados valores de precisão, recall e F1, e demonstrando uma elevada capacidade discriminativa entre as classes definidas.

A utilização da variável temporal composta AnoTrimestre revelou-se uma decisão metodológica relevante, permitindo representar de forma mais adequada a progressão temporal dos dados e melhorar o desempenho dos modelos. De forma global, os resultados obtidos confirmam o potencial das técnicas de Machine Learning como ferramentas de apoio à análise e segmentação do mercado habitacional, contribuindo para uma melhor compreensão das dinâmicas regionais dos preços da habitação.

Importa, no entanto, salientar que os valores elevados das métricas observadas podem estar parcialmente associados à elevada granularidade espacial dos dados, uma vez que a análise foi realizada ao nível das regiões NUTS III. Em Portugal, os preços da habitação apresentam uma forte segmentação geográfica, com diferenças marcadas entre regiões, o que pode facilitar a separação entre classes e contribuir para um desempenho elevado dos modelos. Assim, embora os resultados sejam consistentes, a sua interpretação deve ter em conta esta característica estrutural dos dados.

Como trabalho futuro, destaca-se a possibilidade de incorporar conjuntos de dados adicionais, nomeadamente variáveis socioeconómicas, demográficas ou financeiras, de modo a enriquecer a análise e aumentar a capacidade explicativa dos modelos. A inclusão de séries temporais mais longas poderá igualmente permitir o desenvolvimento de modelos mais avançados e a aplicação de técnicas específicas de previsão temporal.

REFERÊNCIAS BIBLIOGRÁFICA

[1] Tavares, Fernando & Pereira, Elisabeth & Moreira, Antonio. (2014). The Portuguese Residential Real Estate Market. An Evaluation of the Last Decade. Panoeconomicus. 61. 739-757. 10.2298/PAN1406739T.

[2] Pagourtzi, Elli & Assimakopoulos, Vassilis & Hatzichristos, Thomas & French, Nick. (2003). Real estate appraisal: A review of valuation methods. Journal of Property Investment & Finance. 21. 383-401. 10.1108/14635780310483656.

[3] Ja'afar, Nur & Mohamad, Junainah & Ismail, Suriatini. (2021). Machine Learning for propoerty price prediction and price valuation: a systematic literature review. Planning Malaysia. 19. 10.21837/pm.v19i17.1018.

[4] Instituto Nacional de Estatística (INE). (2025). Valor mediano das vendas de alojamentos familiares por m². Disponível em: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&contecto=pi&indOcorrCod=0012234&selTab=tab0&xlang=pt (Acedido em dezembro 2025).

[5] Ferreira, J. (2024). A Machine Learning Approach to Forecasting House Prices in the Portuguese Market.

[6] Kapoor, Akash. (2020). A Comparative Study on House Price Prediction. International Journal for Modern Trends in Science and Technology. 6. 103-107. 10.46501/IJMTST061220.

[7] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) Knowledge Discovery and Data Mining: Towards a Unifying Framework. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 2-4 August 1996, 82-88.

[8] McKinney, Wes. (2022). Python for Data Analysis: Data Wrangling with pandas, NumPy & Jupyter (3rd ed.). O'Reilly Media. ISBN 978-1-098-10403-0.

[9] Kaur, Ritu & Gulia, Preeti. (2020). Experimental Evaluation of Open Source Data Mining Tools (WEKA and Orange). International Journal of Engineering Trends and Technology. 68. 30-35. 10.14445/22315381/IJETT-V68I8P206S.

[10] V. Padmavaty, C. Geetha, N. Priya 2020. Analysis of Data Mining tool Orange. *International Journal of Modern Agriculture*. 9, 4 (Dec. 2020), 1146-1150.

[11] Orange Data Mining. (2025). Discretize - Orange Visual Programming 3 documentation. Disponível em: https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/data/discretize.html (Acedido em dezembro 2025).

[12] Kohavi, Ron. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence. 14.