

Studio 360

Fernando Aguiar, Lucas Gehlen e Lucas Gerlach Nachtigall

I. INTRODUÇÃO

A plataforma desenvolvida pela *start-up* Studio 360 tem por objetivo facilitar a comunicação do setor imobiliário, isto é, realizar o contato entre corretores, incorporadoras, urbanizadoras e o cliente final com maior agilidade e facilidade.

O presente trabalho busca analisar a existência de influência da sazonalidade na procura/acesso à empreendimentos imobiliários, através da seleção de características fornecidas pela base de dados.

II. BASE DE DADOS

A base de dados utilizada para a extração dos dados pertinentes ao objetivo do trabalho, é composta por 19 tabelas que descrevem empreendimentos, estados, municípios, acessos, dentre outras informações. Para a criação de um dataset com atributos relevantes à análise. Sendo cada um deles:

- data: Atributo do tipo *timestamp* formatado para retornar apenas o mês do acesso proveniente da tabela “empreendimento_acesso”;
- empreendimento_id: Atributo do tipo *integer* proveniente da tabela “empreendimento”, contém o identificador do empreendimento;
- cidade_id: Atributo do tipo *integer* proveniente da tabela “cidade”, contém o identificador da cidade;
- bairro_id: Atributo do tipo *integer* proveniente da tabela “bairro”, contém o identificador do bairro;
- acessos: Atributo do tipo *integer* proveniente do agrupamento dos atributos data, empreendimento_id, cidade_id e bairro_id

Estes atributos citados foram obtidos através do seguinte comando SQL, que realizou a query no base de dados com a utilização do SGBD PostgreSQL:

```
select to_char(date(ea.data),'MM') as data,
ea.empreendimento_id, c.id as cidade_id, b.id as bairro_id,
count(*) as acessos, c.nome as cidade, b.nome as bairro
from empreendimento_acessos ea join empreendimento e
on (e.id = ea.empreendimento_id) join cidade c on
(e.cidade_id = c.id) join estado est on(c.estado_id = est.id)
join bairro b on (e.bairro_id = b.id) join
empreendimento_planta ep on (e.id =
ep.empreendimento_id) join tipo_planta tp on (ep.tipo =
tp.id) where ( c.nome is not null and b.nome is not null and
ea.sequencia is not null) group by
to_char(date(ea.data),'MM'), ea.empreendimento_id, c.id,
c.nome, b.id, b.nome;
```

Ao final da seleção dos atributos foram definidos conjuntos de dados com x, y... mil registros. A Tabela 1 demonstra a quantidade de registros presentes no dataset categorizados por municípios.

Municípios	Amostras
Água Santa	7
Arroio do Sal	51
Bento Gonçalves	1822
Canela	2164
Capão da Canoa	5661
Carlos Barbosa	759
Caxias do Sul	66563
Farroupilha	2844
Gramado	11529
Igrejinha	46
Nova Petrópolis	278
Novo Hamburgo	486
Passo de Torres	62
Porto Alegre	1057
Santo Ângelo	271
São Francisco de Paula	519
Sarandi	158
Torres	5492
Tramandaí	231

Tabela 1: Distribuição das amostras por município em um conjunto de dados com 100 mil registros


É importante destacar que conforme o tamanho do dataset o número de municípios será diferente tal como o número de amostras.

III. REFERENCIAL BIBLIOGRÁFICO

Pelli Neto e Zárate[1] em seu trabalho optou por utilizar RNA (Redes Neurais Artificiais), sendo esta a Rede Neural Perceptron Multicamadas, junto com 172 conjuntos de dados, da qual todo foi utilizado para treinamento, para validação deste treinamento foi elaborado um modelo de regressão linear múltipla, através do aplicativo SisRen Windows – Sistema de Rede Neurais.

Por último Grybauskas, Pilinkienė e Stundžienė[3], utilizaram 15 métodos diferentes, para realizar seu estudo, sendo eles CatBoost Classifier, Extreme Gradient Boosting (XGB), Light Gradient Boosting Machine, Random Forest Classifier, Extra Trees Classifier, Gradient Boosting Classifier, Linear Discriminant Analysis, Logistic Regression, Ridge Classifier, Naive Bayes, Ada Boost Classifier, K-Neighbors Classifier, Decision Tree Classifier, Quadratic Discriminant Analysis and SVM—Linear Kernel, para a divisão de teste e treinamento optou-se por 30% e 70% respectivamente, os critérios de avaliação que foram obtidos foram acurácia, curva ROC, sensibilidade, especificidade, F1-score e Matthews coeficiente de correlação (MCC).

Após a seleção das características essenciais para a os datasets foram extraídos para arquivos de texto CSV. Depois disso, foi realizado o processamento os utilizando a linguagem de programação R. O deste processamento é facilitar a análise com o de responder o questionamento proposto. Após a ão de testes com bases de diferentes tamanhos, pela utilização de um arquivo CSV com 100.000 s com ordenação aleatória, para uma análise e contundente realizada na sequência.



range	count
1	8514
2	7483
3	6490

V. ANÁLISE DOS DADOS

Cidade	Accesses per cidade
Apucarana	14
Santa Cruz do Rio Pardo	46
Guarapuã	52
Guarapuã	104
Guarapuã	118
Guarapuã	138
Guarapuã	158
Guarapuã	178
Guarapuã	198
Guarapuã	218
Guarapuã	238
Guarapuã	258
Guarapuã	278
Guarapuã	298
Guarapuã	318
Guarapuã	338
Guarapuã	358
Guarapuã	378
Guarapuã	398
Guarapuã	418
Guarapuã	438

Também consideramos a relação entre o número total de amostras da base de acordo com os meses do ano. A partir disso, chegamos a conclusão de que o período do ano que obteve maior procura até o momento é entre os meses de junho e julho, conforme a figura 3.

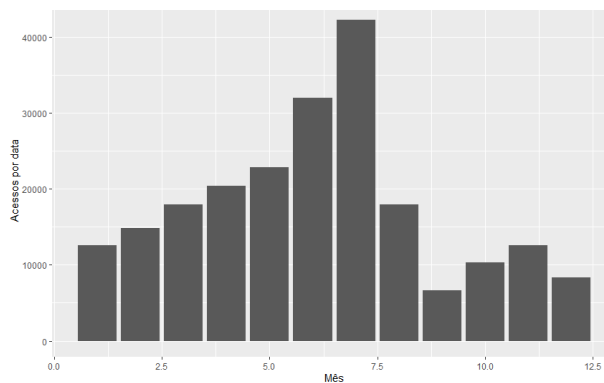


Fig. 3: Número de acessos por mês

Na sequência, realizamos a normalização dos dados via Z-Score, o que viabiliza um conjunto de dados com menor variação.

A figura 4 demonstra a matriz de correlação das variáveis. Foi observado que nenhuma variável possui uma alta correlação (acima de 95%) com outra. Bairro e empreendimento possuem uma correlação entre 40% e 60%. Este dado é um indicador da relação física entre as variáveis, considerando que um empreendimento está estritamente ligado a um bairro, e alguns bairros possuem maior ou menor número de empreendimentos. O mesmo ocorre entre cidade e bairro, considerando sua associação lógica via base de dados, em que um bairro está vinculado a uma cidade.

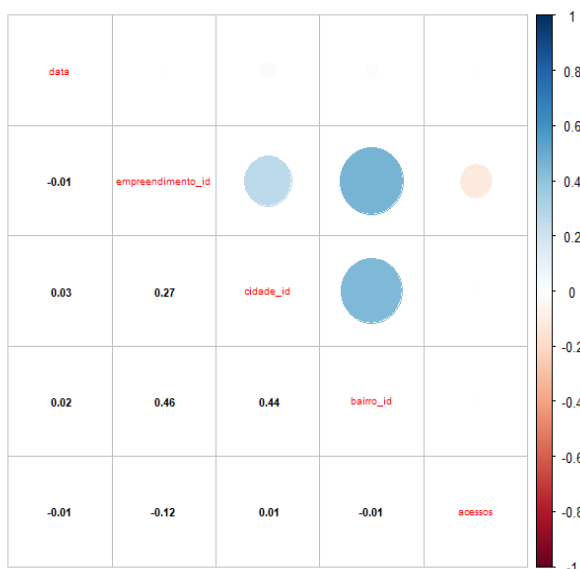


Fig. 4: Matriz de correlação das variáveis

Além da matriz de correlação das variáveis, outra visualização possível para demonstrar a correlação dos atributos se dá por meio da projeção dos dados, com a estimativa de autovetores da covariância.

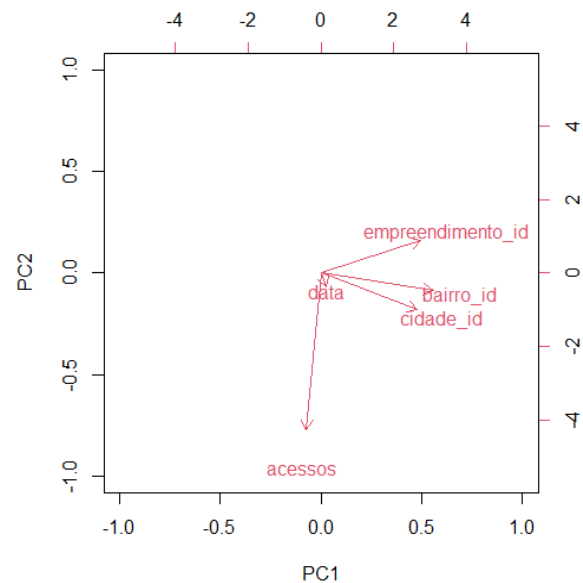


Fig. 5: Projeção dos dados para correlação dos atributos

Os autovetores confirmam a correlação entre bairro e cidade, bem como entre bairro e empreendimento. A data é o atributo que mais influencia no componente principal, o que pode ser um indicativo da influência da sazonalidade na procura por imóveis.

Também foi realizado o boxplot, a fim de observar as variâncias dos componentes principais. No exemplo da figura 6, a variância das amostras considerando o componente como a data. Podemos observar que há certa variabilidade, porém não em escala significativa a ponto de necessitar de uma redução de dimensionalidade.

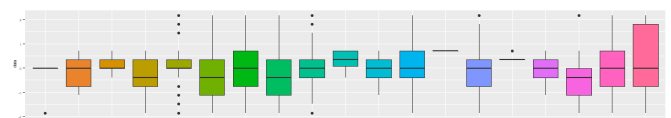


Fig. 6: Projeção dos dados

VI. CLASSIFICAÇÃO

Para um processo de classificação, como a existência de uma sazonalidade já foi observada pela análise da base de dados, optou-se por realizar uma predição do nível dos acessos, que seriam realizados pelos usuários, estes níveis foram adicionados com base no número de acessos que cada empreendimento teve, este processo foi detalhado durante o processo de preparação dos dados.

Nos classificadores paramétricos utilizados optou-se por escolher um grande número de métodos diferentes, para assim ter uma maior gama de possíveis resultados e maior garantia das análises e escolhas dos parâmetros utilizados, sendo os métodos escolhidos GMM 4 Componentes, GMM Completo, GMM Default, GMM Diagonal, MLP e MNET.

Os modelos de misturas Gaussianas (Gaussian Mixture Model – GMM) são uma função densidade de probabilidade paramétrica representada como uma soma ponderada de M densidades dos componentes Gaussianos ($p(x) = \sum_{k=1}^M \pi_k f(x|\mu_k, \Sigma_k)$) onde x é um vetor d -dimensional de variáveis contínuas, μ_i , $i = 1, \dots, M$, são as proporções da mistura (normalização ou contribuição), e $f(x|\mu_i, \Sigma_i)$, $i = 1, \dots, M$, são as densidades para cada componente. O que difere os tipos de GMM é que Σ_k é diagonal ($\sigma_{ij} = 0$, $i \neq j$) e as variâncias são todas iguais ($\sigma_{ii} = \lambda$) – Esférica, já Σ_k é diagonal ($\sigma_{ij} = 0$, $i \neq j$) e as variâncias são diferentes – Diagonal e Σ_k é o caso geral – Completa.

O MLP (Rede Neural Perceptron Multicamadas), esta por sua vez é caracterizada por possuir uma ou mais camadas intermediárias de neurônios e uma saída, a arquitetura mais comum para uma rede MLP é a completamente conectada, de forma que os neurônios de uma camada estão conectados a todos os neurônios da camada posterior. Na MLP, cada neurônio realiza uma função específica, sendo essa função implementada por um neurônio de alguma camada é uma combinação das funções realizadas pelos neurônios da camada anterior que estão conectados a ele, à medida que o processamento avança o processo realizado fica cada vez mais complexo até a camada de saída. Cada neurônio da camada de saída está associado a uma classe do conjunto de dados [7].

Para os classificadores não paramétricos utilizados também optou-se por escolher um grande número de métodos diferentes, para assim ter uma gama de possíveis resultados e maior garantia das análises e escolhas dos parâmetros utilizados, sendo os métodos escolhidos Bagging, KNN, Random Forest, Árvore de decisão, SVM Linear, SVM Polinomial e SVM Radial.

O método Bagging (Bootstrap Aggregating), responsável por gerar várias versões de um classificador e utilizá-las para obter, posteriormente, um classificador agregado. Este método realiza a criação de novos conjuntos de dados através de bootstrap (amostragem com reposição), utilizando-se da duplicação de dados para o mesmo [7].

O algoritmo *K nearest neighbours* (KNN) ou k -vizinhos que se trata de um método que consiste em memorizar os dados de treinamento para realizar predições para uma nova instância desconhecida, a partir dos valores observados para uma quantidade k de vizinhos mais próximos.[7].

O algoritmo *Random Forest* (RF), que faz uso da técnica de geração de várias árvores de decisão. O algoritmo estima árvores a partir de uma entrada aleatória do conjunto de treinamentos. Ao final, todas as árvores tomam a decisão individualmente, e a decisão mais votada dentre as árvores geradas é tida como o retorno do *Random Forest* [7].

O método árvores de decisão, realiza um processo de decisão multiestágio, i.e., em vez de utilizar o conjunto completo de características de uma vez só para decidir, diferentes subconjuntos de características são utilizados em diferentes níveis da árvore, Classes são rejeitadas

sequencialmente até que obtenha-se uma única classe aceita.

O SVM é um algoritmo que busca uma linha de separação entre duas classes distintas, analisando os dois pontos, um de cada grupo, mais próximos da outra classe. Isto é, o SVM escolhe a reta — também chamada de hiperplano em maiores dimensões— entre dois grupos que se distanciam mais de cada um.

VI. CONFIGURAÇÃO EXPERIMENTAL

A preparação do experimento conta com uma etapa importante, que é a escolha de um método avaliativo de desempenho. Optou-se pela utilização do Stratified Holdout.

Stratified Holdout ou amostragem estratificada, segundo Faceli [6], consiste em realizar uma divisão entre dados que serão utilizados para treinamento e teste realizando o processo de balancear as amostras selecionadas de cada classe. Uma das abordagens de balanceamento possível é manter o número de objetos em cada classe proporcional ao número de objetos que a classe possuía no conjunto original. Ao realizar o balanceamento das classes garante-se que classes com objetos dominantes não interfiram com classificadores tendenciosos [3].

Neste caso, para o Stratified Holdout implementado, sobre um conjunto de dados X sendo dividido em dois subconjuntos mutuamente exclusivo, treinamento Y e teste Z ($X = Y \cup Z$ e $Y \cap Z = \emptyset$), com uma divisão entre dados utilizados em treinamento e testes foi feita da seguinte forma: 66,66% dos dados foram utilizados para treinamento, enquanto os outros 33,33% utilizados para testes. Para uma otimização do modelo demandado os dados do problema, o subconjunto de treinamento é dividido ainda em dois conjuntos: treinamento e validação ($X = Y \cup W \cup Z$, $Y \cap Z = \emptyset$, $Y \cap W = \emptyset$ e $W \cap Z = \emptyset$), para uma divisão entre os dados do treinamento foi realizada da seguinte forma: 80% para treinamento dos dados, enquanto os outros 20% utilizados para validação.

Estes métodos e porcentagens foram escolhidos com base nos estudos de Sonavni[2], Grybauskas, Pilinkienė e Stundžienė[3] que como já mostrado utilizaram os métodos de validação cruzada e stratified respectivamente, e os classificadores de Grybauskas, Pilinkienė e Stundžienė obtiveram um melhor accuracy em comparação com Sonavni[2].

VIII. RESULTADOS

Conforme a figura 7, após o treinamento dos métodos, não-paramétricos, apresentados todos obtiveram uma acurácia acima de 90% de acertos indicando que independente do algoritmo de classificação, a base de dados garante a efetividade de classificação de acessos com base na sazonalidade.

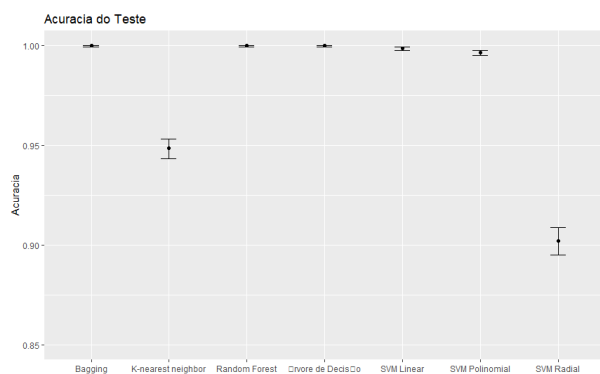


Fig. 7: Acurácia modelos Não paramétricos

Para os métodos de paramétricos conforme a figura 8, obtiveram resultados inferiores em relação aos métodos não paramétricos, mas apesar de inferiores todos os algoritmos obtiveram uma acurácia superior a 90%. Sendo os métodos MLP e MNET destacando-se com quase 100% de acurácia.

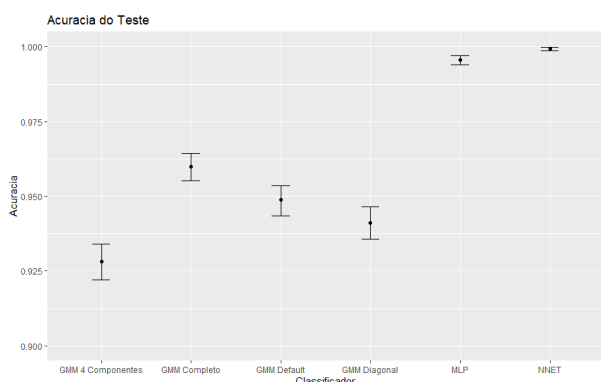


Fig. 8: Acurácia modelos paramétricos

Entre os melhores algoritmos destacam-se os Cart, Bagging e Random Forest, dos métodos não paramétricos, e o MLP e MNET, dos métodos paramétricos, que obtiveram os seguintes resultados de acurácia Cart, Bagging, Random Forest e MNET com uma Acurácia de 1, ou 100% de eficácia, e o MLP com uma Acurácia de 0.9957, ou 99,57% de eficácia. Entre os piores métodos estiveram o algoritmo SVM Radial, não paramétrico, e o algoritmo GMM 4 Componentes, paramétrico, que obtiveram resultados respectivamente de 90,21% e 92,82%, embora estes sejam os piores resultados, vale ressaltar que em nenhum trabalho utilizado para referência foi obtido um resultado tão satisfatório com quase ou 100% de eficácia de classificação.

IX. CONCLUSÃO

Neste projeto, foi possível perceber as dificuldades encontradas em um projeto real de preparação de dados. Percebemos a imensa quantidade de tabelas e atributos em bases de dados reais. Neste caso, e na maioria das situações reais de extração de dados, os dados não possuem um padrão adequado para passar por análises estatísticas complexas ou treinamentos. Por isso, uma de nossas maiores dificuldades foi entender a partir de quais dados poderíamos chegar a resposta do questionamento

proposto inicialmente: “Existe sazonalidade ao acesso de empreendimentos (inverno / verão)?”.

Como resultado da análise, obtemos algumas surpresas. Chegamos por exemplo à conclusão de que Junho e Julho são sem sombra de dúvidas os meses com maior número de acessos a empreendimentos. A partir da análise destes gráficos e também da estimativa de autovetores da covariância, também concluímos que a data interfere diretamente nos outros atributos.

Vale ressaltar também que, por trabalharmos com poucos atributos selecionados pontualmente, todos eles possuem um considerável poder de discriminação quando pensamos no número de acessos.

Com base nos comparativos entre os classificadores não-paramétricos e paramétricos, é possível concluir que o algoritmo que atingiu melhores resultados para a classificação é o classificador não-paramétrico, visto que conseguiu atingir métricas melhores para todos os casos de teste realizados, para o conjunto de dados do “Studio 360”.

Foi possível perceber também que os classificadores de melhor resultado são, os métodos não-paramétricos, Cart, Bagging e Random Forest, diferente do método SVM Radial, paramétrico, que obteve o pior resultado, entretanto este obteve uma acurácia de aproximadamente 0.9021 ou 90,21% de acerto.

X. REFERÊNCIAS

- [1]Pelli Neto, Antônio e Zárate, Luis Enrique. CONGRESSO BRASILEIRO DE ENGENHARIA DE AVALIAÇÕES E PERÍCIAS, BELO HORIZONTE/MG, XII., 2012, Belo Horizonte. AVALIAÇÃO DE IMÓVEIS URBANOS COM UTILIZAÇÃO DE REDES NEURAIIS ARTIFICIAIS [...]. [S. l.]: IBAPE, 2012. Disponível online em: <https://ibape-nacional.com.br/biblioteca/wp-content/uploads/2012/12/Avaliacao-de-Imoveis-Urbanos-com-Utilizacao-de-Redes-Neurais-Artificiais.pdf>. Acesso em: 5 nov. 2022.
- [2]Sonavni, Varun. Real Estate House Price Prediction Using Data Science. **Python in Plain English**, [S. l.], 15 set. 2021. Disponível online em: <https://python.plainenglish.io/data-science-project-real-estate-house-price-prediction-website-df71ac98a132>. Acesso em: 5 nov. 2022.
- [3]Grybauskas, A., Pilinkienė, V. & Stundžienė, A. Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic. J Big Data 8, 105 (2021). Disponível online em: <https://doi.org/10.1186/s40537-021-00476-0>. Acesso em: 5 novembro. 2022.
- [4]CHAKRABORTY, Amitabha. **Bengaluru House price data**. Kaggle, 10 abr. 2018. Disponível online em: <https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data>. Acesso em: 6 nov. 2022.
- [5]CHEN, Dehua; NIGRI, Eduardo; OLIVEIRA, Gibram; SEPULVENE, Luis; ALVES, Tiago. Métricas de Avaliação

em Machine Learning: Classificação. **Kunumi Blog**, Medium, p. 1-2, 10 jun. 2020. Disponível online em: <https://medium.com/kunumi/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-em-machine-learning-classifica%C3%A7%C3%A3o-49340dcd198>. Acesso em: 30 out. 2022.

[6] BORDE, Dr. Swapna; RANE, Aniket; SHENDE, Gautam; SHETTY, Sampath. Real Estate Investment Advising Using Machine Learning. **International Research Journal of Engineering and Technology (IRJET)**, Wwww.irjet.net, ano Mar-2017, v. 04, n. 03, p. 1821-1825, 22 mar. 2017. Disponível online em: https://d1wqtxts1xzle7.cloudfront.net/53503829/IRJET-V4I3499-with-cover-page-v2.pdf?Expires=1668198504&Signature=d~Dea5ggc1XRApdFBauzffBsLAb37GCWU5sceFz4JKC7BLiKTisP4PE~aSXy5hGyg-3RLVYNbfGWMhkM-DXMYjTnMDWnJXY~DRpbdGq7hmTNZibmLDl5wE007uq3n7yHJvwEVvy2Bw8FxoGLy9P50JKY0AzRbisajMgo8-yuIVTT~n30BREJr~ZfbReD6Yc7eweOOhTFi14rZgAoD76DM9WGJeDxnjiCbQEDN-qnr8ugQYLgh86Pqh~BO93WF93HpbnS6A1F1OTEXGjDtv63pR5bNLmzFvjI2LWJKdlRYIfpbXHV~YXhnlFNN5l3BKd~zKoipF85NcfVdGgfKg1GYw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA. Acesso em: 6 nov. 2022.

[7] FACELI, Katti; LORENA, Ana C.; GAMA, João; AL, et. **Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina**. Grupo GEN, 2021. 9788521637509. Disponível online em: <https://integrada.minhabiblioteca.com.br/#/books/9788521637509/>. Acesso em: 3 dez. 2022