[](https://moodle.mackenzie.br/moodle)

**Universidade Presbiteriana Mackenzie**

**BIG DATA & ANALYTICS**

**TURMA H**

**TRABALHO FINAL DISCIPLINA**

**MINERAÇÃO DE DADOS**

**Hermes Luiz Bolinelli Junior RA: 71661573**

**ABRIL/2018**

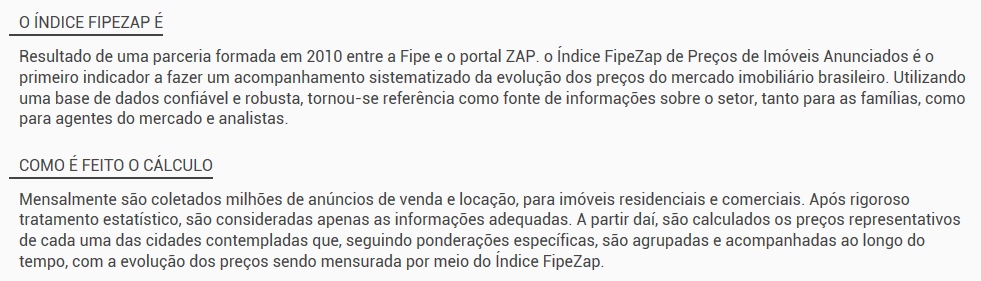
**Projeto Data Mining - Avaliação Imobiliária**

<https://github.com/HermesJunior/DM>

1. **Introdução**

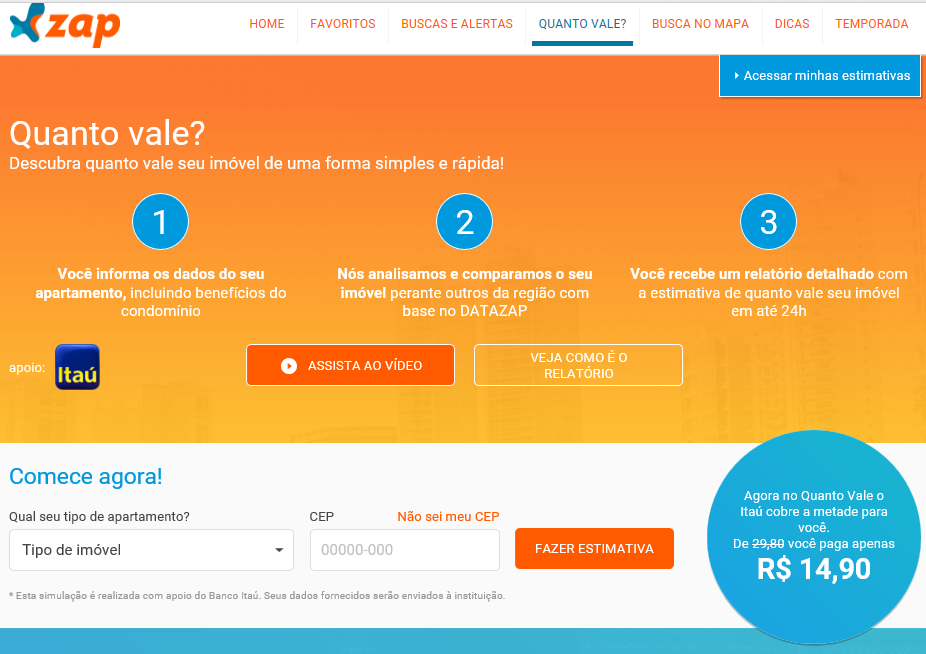
O presente trabalho objetiva classificar e estimar o valor de mercado de um imóvel através do Método Comparativo de Dados de Mercado (método este preconizado pelas normas de Engenharia de Avaliações ABNT NBR 14.653).

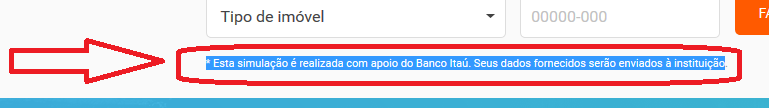
A utilização de grandes Banco de Dados vem cada vez mais sendo utilizada pelas grandes instituições financeiras de moda a permitir um aumento de eficiência nas avaliações de imóveis. Os métodos tradicionais vêm, pouco a pouco, perdendo espaço com a utilização de grandes datasets. Dentro deste contexto algumas instituições já vêm empregando Big-Data nas suas análises, veja, por exemplo, o índice Fipe-Zap:



http://fipezap.zapimoveis.com.br/

e, ainda no mesmo site:





Dentro deste cenário um dos desafios é a classificação do padrão de um imóvel e a sua clusterização em grupos homogêneos, como preconizado pela norma ABNT NBR 14.653, que se torna inviável de ser feito manualmente em grandes datasets, e as consequentes valorações dos imóveis após esta classificação/agrupamento.

O IBAPESP possui publicações que auxiliam nesse processo de classificação e homogeneização, em que procura tirar subjetividade destes processos como a cartilha Valores de Edificações de Imóveis Urbanos 2017 (<http://www.ibape-sp.org.br/normas_estudos/Default.aspx>).

Para maior compreensão deste tema recomenda-se a leitura desta cartilha, em que princípios são aplicados neste trabalho, assim como a norma ABNT NBR 14653, mas que foge ao escopo deste a sua explanação, que tornaria inviável no espaço deste texto.

1. **Objetivo**

Isto posto, este trabalho tem como objetivo a classificação, o agrupamento em dos padrões homogêneos de imóveis urbanos, tipologia apartamentos, com base em aspectos objetivos e quantitativos, e sua consequente valoração.

Como método não supervisionada para a classificação dos imóveis foi utilizado o algoritmo k-means, para as classificações (método supervisionado) foi utilizado os métodos Knn e Regressão Logística. Já para o modelo de Regressão para obtenção dos valores foi utilizado a função lm.

Todo o script desenvolvido em R está anexo a este relatório, bem como sua versão MarkDown processada.

1. **Análise dos dados**

Para a análise de dados foi utilizado originalmente um banco de dados com pouco mais de 70.000 registros e 390 features.

Muitas dessas features possuíam dados correlacionados com outras features, dados incompletos, dados preenchidos de forma incorreta por diversos motivos (exemplo: unidade habitacional em apartamento com 32 banheiros??), ... ou seja um tratamento na base de dados original foi necessário. A base de dados se mostrou bastante heterogênea e com muitos erros de preenchimento. Isto talvez se deva ao fato de ser construída por vários profissionais, com critérios e níveis de experiência diferentes. A padronização e métodos que eliminem a subjetividade aqui já se mostraram importantes.

Para esse tratamento inicialmente foi utilizado o software [openRefine](http://openrefine.org/). Neste software foi realizada um amplo tratamento dos dados originais, diminuindo o dataset para os objetivos deste trabalho que é um trabalho de final de disciplina, o escopo limitado do trabalho e poder de processamento tanto dos equipamentos utilizados como do próprio R.

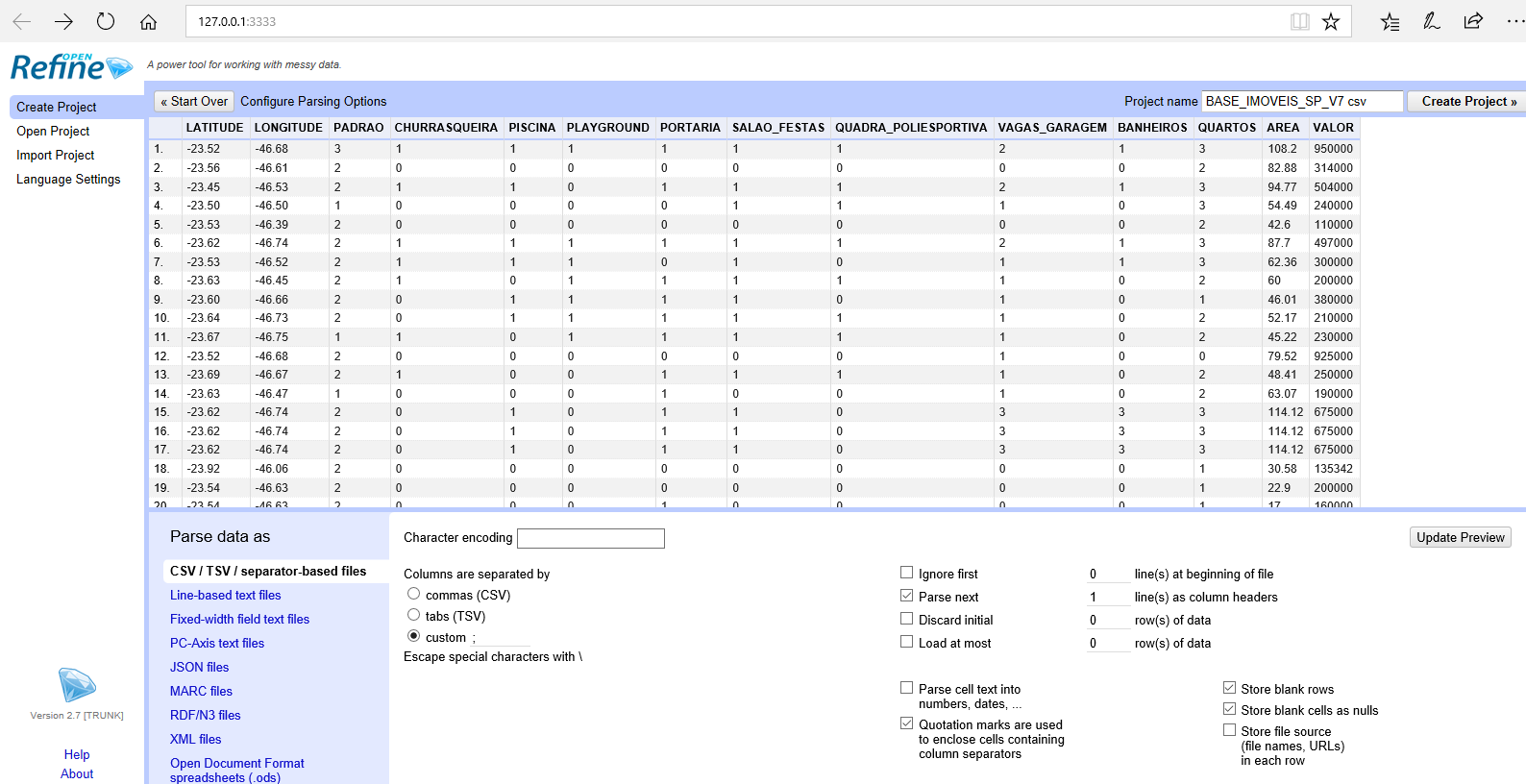
****

Imagem software openRefine durante o tratamento dos dados.

Ao fim desse processo obteve-se a base de dados nomeada como “BASE\_IMOVEIS\_SP\_V7” que foi a base efetivamente utilizada no trabalho, com 13 features e 18950 registros.

No script do R é possível notar que foi realizada ainda uma continuidade desse tratamento, então verificou-se neste script a ocorrência de NA´s (que foram eliminados). Também foi realizada a normalização dos dados nesta etapa.

1. **Normalização dos dados**

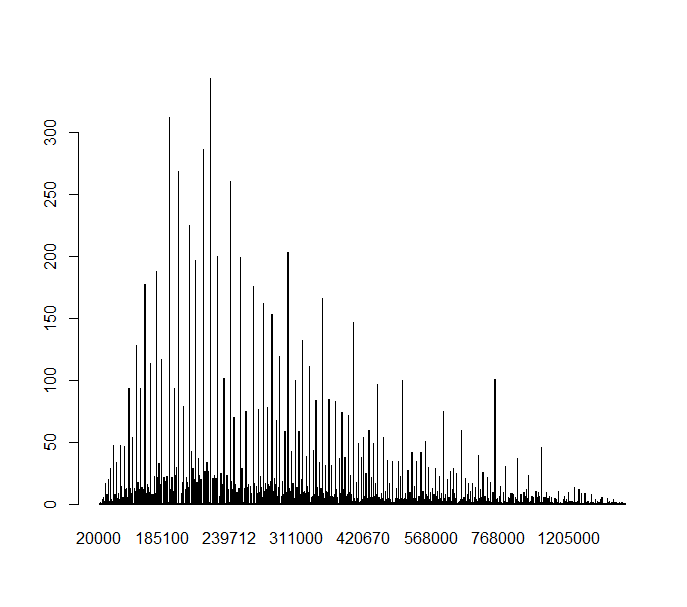
Os dados tratados foram normalizados. Manteve-se o dataset original, este mesmo datasetr foi normalizado e criou-se já nesta etapa os datasets train, t (treinamento original e normalizado) e os datasets test e tst (treinamento original e normalizado).

1. **Análise dos dados**

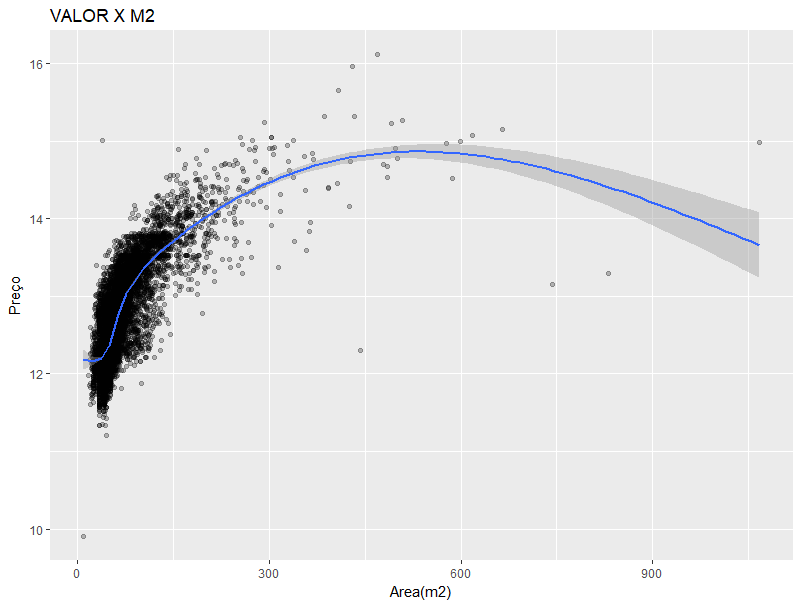
Após o tratamento realizou-se a análise dos dados. Nesta etapa utilizou-se de gráficos e cálculo das correlações para melhor entendimento do dataset utilizado.

Abaixo seguem os gráficos obtidos e na sequência as análises realizadas:

* 1. barplot(table(train$VALOR))

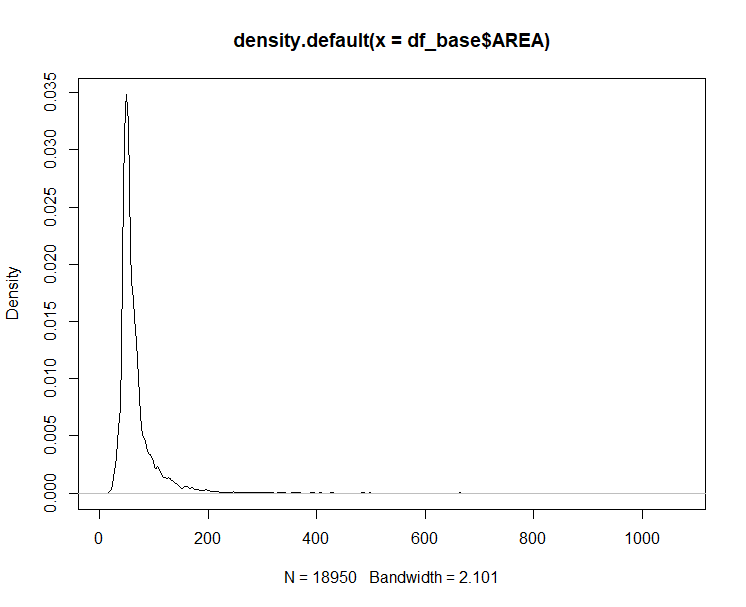
****

* 1. ggplot(train, aes(x = AREA, y = log(VALOR))) + geom\_point(position = "jitter", alpha = 0.25)

****

* 1. d <- density(df\_base$AREA)

plot(d)



Dos gráficos acima nota-se uma concentração de imóveis área até 150 m2 com média aproxima a de 70m2. Já os valores estão na maioria na faixa de R$ 300.000,00.

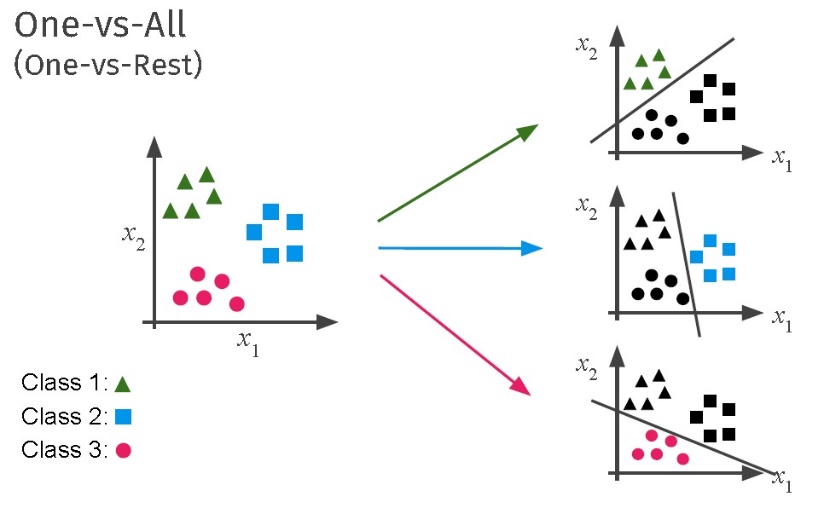
Os imóveis maiores que os acima indicados estão em uma proporção menor, ficando esparsos no gráfico, já os imóveis menores estão em concentração maior.

1. **Aplicação do algoritmo K-nn**

O algoritmo Knn foi utilizado, nos dataset treinamento e test normalizados, para verificar a coluna “Padrão” da base de dados. O índice de acurácia mostrou-se alto, com quase 95% de acerto.

1. **Aplicação do algoritmo Regressão Logística**

O algoritmo Regressão Logística foi utilizado, nos dataset treinamento e test normalizados, para verificar a coluna “Padrão” da base de dados. Como esse métodos retorna valores entre 0/1 e a feature “Padrão” possui 3 valores (1,2,3) optou pela estratégia “One x All”:



Em que verificou-se cada ocorrência contra as outras duas ocorrências, como segue:

|  |  |
| --- | --- |
|  | % ACERTO |
| GRUPO 1 | 100% |
| GRUPO 2 | 98.75% |
| GRUPO 3 | 100% |

1. **Aplicação do algoritmo K-means e Regressão Multivariada**

Inicialmente foi calculado o valor do imóvel para todos os registros e features do dataset utilizado. Este modelo inicial foi chamado de MODELO 01 BASE. Após o cálculo do Modelo 01 Base foi realizado o agrupamento em 3 grupos, etapa em que foi empregado o algoritmo k-means.

Apesar de já destacado no início do trabalho, esta simulação é interessante pois na prática esse agrupamento não terá uma base prévia para classificação, pois depende das característica intrínsecas e extrínsecas dos imóveis, da data em que foi realizada, das condições de conservação e manutenção, reformas, do próprio mercado, entre inúmeras outras. Muitas das quais são extremamente subjetivas e que fica na dependência do arbítrio do avaliador.

Claro que aqui a experiência conta, mas poder contar com um método que possibilite através de determinadas features eliminar a subjetividade aumentando a padronização dos trabalhos em critérios técnicos e objetivos, pode em muito contribui com a área de avaliações.

Isto posto, oara a aplicação do algoritmo k-means foi definido k=3 por termos 3 padrões no dataset, pois aé uma ideia acessória a comparação entre essas informações. Os dados foram tratados e normalizados como já descrito, sendo utilizado toda a base de dados. Para cada um destes grupos será processado o modelo Regressão “lm” para a obtenção dos valores de cada grupo e posterior comparação com o Modelo 01 Base.

Tanto para o modelo base como para os modelos de cada um dos grupos calculou-se também o MAE (Erro Médio Absoluto) para comparação. Depois comparamos cada um dos erros médios (MAE) de cada grupo com o modelo 01 base. O desejado seria o modelo de cada grupo ter seu erro menor do que o modelo base, pois o modelo base possui todos os registros, enquanto que cada um dos grupos são mais homogêneos entre si (em tese). Mas não foi o que não ocorreu na prática (ver tabela abaixo).

Fatores que podem ter contribuído para isso é a qualidade do dataset, que verificou-se ser, ao longo do trabalho, bastante comprometido na sua qualidade das informações, bem como as features efetivamente utilizadas que poderiam ter sido outras. A escolha das features adequadas e representativas do que se espera obter é importante. Mas isso não aconteceu justamente por não poder utilizar as features que muitas vezes eram selecionada, pois estas estavam bastante comprometidas na sua qualidade (faltando registros, erros outros), tendo-se que optar por outras.

Desta forma chegamos aos seguintes resultados:

|  |  |  |
| --- | --- | --- |
| **MODELO** | **VALOR(R$)** | **MAE** |
| BASE | 368.175.60 | 88.482,06 |
| A | 405.611.20 | 97.441.43 |
| B | 274.088.80 | 168.419.80 |
| C | 354.328.80 | 87.998.49 |

1. **Distribuição dos imóveis por Grupos Homogêneos**

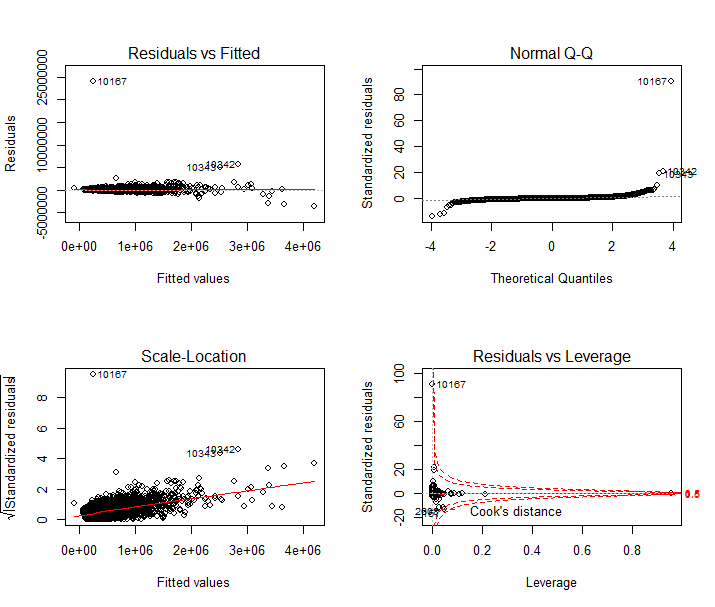
Para demonstrar a distribuição da classificação obtida foi gerado um mapa dentro do script R utilizando como variáveis a latitude e longitude dos imóveis dentro do município de São Paulo.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Legenda: Classificação**   |  |  | | --- | --- | |  | A | |  | B | |  | C | |

* 1. Regressão: Gráficos

Os gráficos abaixos foram obtidos para o Modelo 01 Base.:

par(mfrow=c(2,2))plot(ylm)



* O gráfico "*Residual x Fitted*" mostra se os resíduos têm padrões não lineares, podendo haver uma relação não linear entre variáveis preditoras e uma variável de resultado, e o padrão pode aparecer nesse gráfico se o modelo não capturar o relacionamento não linear.

Se você encontrar resíduos igualmente espalhados ao redor da alinha horizontal sem padrões distintos, isso é uma boa indicação de que você não tem relacionamentos não lineares.

* O gráfico "Normal-QQ" mostra se os resíduos são normalmente distribuídos e se os resíduos seguem uma linha reta ou desviam-se severamente.

Podemos observar que os resíduos são alinhados na linha reta tracejada.

* O gráfico "Scale-Location" mostra se os resíduos são distribuídos igualmente ao longo dos intervalos de preditores.

Desta forma podemos verificar a suposição de iqual variância.

* O gráfico "Residuals vs Leverage" ajuda a encontrar variáveis influentes, caso hajam.

Nem todos os outliers são influentes na análise de regressão linear. Mesmo que os dados tenham valores extremos, eles podem não ser influentes para determinar uma linha de regressão. Isso significa que os resultados não seriam muito diferentes se os incluíssemos ou os excluíssemos da análise.

1. **Crux**

A grande dificuldade enfrentada neste projeto foi o ajuste da base de dados, pois existiam muitos dados em não conformidade de tipo e muitos erros de semântica no quais a base estava tanto com caracteres armazenados em UTF-8 e ISO-8859-1 os quais foram necessário certo esforço na limpeza e correção dos mesmos.

Outra dificuldade foi o tamanho da base original e a quantidade excessiva de variáveis (features) onde tivemos que escolher as melhores, do que a base disponibiliza, para as análises.

Nem sempre a feature que era considerada necessária para este trabalho tinha os registros adequados ou completos, o que obrigou a sua não seleção. Isso comprometeu, a nosso ver, bastante nos resultados esperados.

1. **Conclusão**

Vamos dividir essa conclusão em 2 etapas.

1)

Em relação aos resultados obtidos neste trabalho as técnicas se mostraram adequadas, visto que são técnicas já utilizadas pelas grandes empresas. Independente dos resultados abaixo discutidos, o cenário é bastante promissor e os erros precisam ser melhor entendidos e corrigidos.

Os algoritmos de classificação apresentaram resultados como esperado, com altos percentuais entre os padrões definidos em campo e os obtidos pelo processamento das demais features em cada método. Isso talvez ocorra em função do padrão ser aplicável a benfeitoria em si, e as variáveis utilizadas são basicamente as intrínsecas ao imóvel, com exceção da latitude e longitude, que no caso da definição do padrão contribuiu bastante.

Já no agrupamento pelo método k-means, em que buscou-se os grupos mais homogêneos e não o padrão, os resultados foram bastante diversos do que se esperava no início, ao se propor o trabalho. Isso se deve a alguns fatores:

* Qualidade do dataset efetivamente utilizado (desbalanceado; qualidade dos dados, seleção das features que foram possíveis e não as que objetivadas inicialmente);
* O fato que as variáveis são basicamente intrínsecas ao imóvel, sendo que para a formação dos grupos homogêneos isto não é positivo, ao contrário, prejudica. Neste caso as variáveis extrínsecas também são extremamente relevantes (exemplo: área terreno, localização, vizinhança, aspectos valorizantes e/ou desvalorizantes (feiras, favelas, parques,...) entre outros) mas que não foram utilizadas por motivos já expostos.
* O acima impacta na homogeneização dos grupos, o que contribuiu para aumento do erro
* Isso influiu também no MAE e nos valores dos imóveis obtidos.

2)

Como trabalho de final de disciplina, entendemos que o trabalho atingiu os objetivos propostos na disciplina, onde pode-se explorar os conceitos de Mineração de Dados, com a utilização da ferramenta R, aplicadas a um caso real.

As dificuldades encontradas foram as mesmas que se enfrentam no dia a dia. Diferencia-se nestas apenas em alguns aspectos de gestão do tempo de dedicação dos profissionais, visto que aqui precisou-se achar tempo e no caso real a dedicação é total. Diferencia-se também no uso de equipamentos mais robustos, dentre outras diferenças. Mas todo o restante é equivalente ao dia a dia.

Isso permitiu, portanto, uma discussão da aplicação das técnicas, do conhecimento e das dificuldades que se impõem, da necessidade de investimento contínuo tanto na obtenção dos *skills* como de maior experiência que só advém com o tempo.

Hermes Luiz Bolinelli Junior

Abril, 2018.