

INVESTIDORES DO TESOUREIRO DIRETO: EXPLORAÇÃO DO COMPORTAMENTO DOS INVESTIDORES UTILIZANDO MINERAÇÃO DE DADOS

Raphael Tavares Sales

IFF - Instituto Federal de Educação, Ciência e Tecnologia Fluminense-RJ.
Rua Dr. Siqueira, 273 - Parque Dom Bosco, Campos dos Goytacazes, RJ - CEP: 28030-130.
raphaelsales@hotmail.com

Henrique Rego Monteiro da Hora

IFF - Instituto Federal de Educação, Ciência e Tecnologia Fluminense-RJ.
Rua Dr. Siqueira, 273 - Parque Dom Bosco, Campos dos Goytacazes, RJ - CEP: 28030-130.
Henrique.dahora@iff.edu.br

RESUMO

Do ponto de vista do investidor comum (pessoa física), o tesouro direto é uma alternativa atrativa de investimento na renda fixa com menores custos de transação uma vez que elimina intermediários. A partir de uma base de dados do cadastro dos investidores do Tesouro Direto no Brasil do site: dados.gov.br, este trabalho buscou utilizar técnicas de mineração de dados como a clusterização e árvore de indução a fim de traduzir o comportamento dos investidores do Tesouro Direto. As técnicas realizadas no banco de dados foram aplicadas de modo que identificassem o comportamento e o perfil do investidor através da mineração de dados, sendo esse usuário inativo ou ativo, identificando características recorrentes entre eles. Dentre os meses analisados, podemos chegar a um perfil comum como sendo do sexo masculino, casado e com idade média de aproximada de 45 anos.

Palavra-chave: Tesouro direto; Weka; Mineração de Dados.

ABSTRACT

From the point of view of the average investor (individual), direct treasury is an attractive alternative to fixed income investments with lower transaction costs once it eliminates intermediaries. From a database of the Treasury Direct Investors Register in Brazil from the site: data.gov.br, this work sought to use data mining techniques such as clustering and induction tree in order to translate the investor behavior of Treasury Direct. The techniques performed in the database were applied in order to identify the investor behavior and profile through data mining, whether this user is inactive or active, identifying recurring characteristics among them. Among the months analyzed, we can reach a common profile as being male, married and with an average age of approximately 45 years.

Keywords: Direct treasury; Weka; Data mining.

1. INTRODUÇÃO

O tesouro direto (TD) tem se mostrado um excelente investimento, tanto do ponto de vista da lucratividade quanto da seguridade e constância da renda, características da renda fixa. Segundo Pimentel *et al.* (2015), é normal que no Brasil as pessoas não possuam mais informações a respeito das variadas formas de investimento, sendo assim, muitas vezes tem existe um certo receio em investir o dinheiro poupado.

Implantado em 2002, a criação do TD foi benéfica para ambos os lados Governo e Investidores. Segundo Perlin (2016), do ponto de vista do governo, trata-se de uma nova fonte direta (e interna) de capital para financiar suas operações, além de ser também uma ferramenta para a política monetária do país. Do ponto de vista do investidor comum (pessoa física), o tesouro direto é uma alternativa atrativa de investimento na renda fixa com menores custos de transação uma vez que elimina intermediários.

Além da rentabilidade líquida dos instrumentos serem atrativas, a variedade de títulos de dívidas oferecidos pelo TD ao público comum é bastante alta, satisfazendo as mais diversas necessidades. Para Filho e Alves (2017), o TD se apresenta como uma forma de autofinanciamento do Estado a partir da emissão direta de títulos da dívida pública federal para pessoas físicas, como características destes investimentos os autores citam: a boa rentabilidade, o baixo custo e risco e a comodidade para realizar a aplicação.

Em setembro de 2018, as vendas do Tesouro Direto chegaram a atingir R\$ 1.761,2 milhões e seus regates totalizaram R\$ 921,9 milhões segundo dados do Tesouro Nacional (“Balanço e Estatísticas - Tesouro Direto - STN”, 2018). O título mais demandado pelos investidores foi o indexado à Selic (Tesouro Selic), cuja participação nas vendas atingiu 40,8%. Os títulos indexados à inflação (Tesouro IPCA+ e Tesouro IPCA+ com Juros Semestrais) corresponderam a 32,9% do total e os prefixados, 26,3%.

Louro *et al.* (2017), afirmam que a tomada de decisão é um processo que exige embasamento e conhecimento. Para o autor em seu trabalho sobre a análise preditiva do perfil de investidores no tesouro direto com vistas às ações de marketing, o indivíduo quando deparado com um cenário de mercado cada vez mais competitivo deve tomar todas as suas decisões com base em dados concretos, que traduzem a realidade e com análises que revelam o que estes escondem.

Para Ferreira (2016), a análise de investimento em títulos públicos do Tesouro Nacional em um cenário de crise econômica faz o seguinte questionamento: “em momentos de crise, os títulos públicos do Tesouro Nacional podem ser uma boa opção de investimento, frente às opções de produtos tradicionais de renda fixa oferecidas pelos bancos? ” E acrescenta: “crises econômicas normalmente geram grandes oscilações nos mercados financeiros e, conseqüentemente, preocupações aos investidores”.

Ainda segundo Ferreira (2016), o investimento em títulos públicos em um momento de crise econômica pode sim ser mais rentável do que investimento em produtos tradicionais de renda fixa. Ademais, é possível afirmar que este retorno pode ser substancialmente superior aos títulos tradicionais de renda fixa ofertados pelos bancos.

Segundo Larose (2005), a Mineração de Dados é comumente classificada pela sua capacidade em realizar determinadas tarefas, dentre estas podemos citar a de descrição que visa descrever os padrões e tendências revelados pelos dados. A descrição geralmente oferece uma possível interpretação para os resultados obtidos e é muito utilizada em conjunto com as técnicas de análise exploratória de dados, para comprovar a influência de certas variáveis no resultado obtido. Através da mineração de dados de perfis podem ser previstas tendências de mercado, além de predição de possíveis concorrentes dentro de um mesmo, com a mineração padrões podem ser descobertos para uma posterior tomada de

decisão.

Este trabalho tem como objetivo expor o comportamento dos investidores do Tesouro Direto no Brasil, tendo como referência o corte temporal. O primeiro (Dezembro de 2015) e o último mês (Agosto de 2016) do processo de impedimento ocorrido recentemente no Brasil de modo que possa ser explorado o perfil desses investidores nesse período de grande incerteza na economia do país. Uma série de perguntas nos resultados foram respondidas a fim de caracterizar o comportamento do investidor naquele momento caótico que o país atravessava.

2. METODOLOGIA

Esta pesquisa pode ser caracterizada, de acordo com o exposto por Vergara (2009), como sendo descritiva e quantitativa, pois procurou através da mineração de dados descreverem um comportamento comum de investidores no período do processo de impedimento que se transcorreu no Brasil entre os anos de 2015 e 2016.

Diversos processos definem e padronizam as fases e atividades da Mineração de Dados. Apesar das particularidades, todos em geral contêm a mesma estrutura. No presente trabalho, foi escolhido o CRISP-DM (*Cross-Industry Standard Process of Data Mining*).

As fases do processo CRISP-DM de acordo com Olson e Delen (2008) são:

1. Entendimento dos Negócios: Nessa etapa, o foco é entender qual o objetivo que se deseja atingir com a mineração de dados. O entendimento do negócio irá ajudar nas próximas etapas.

2. Entendimento dos Dados: As fontes fornecedoras dos dados podem vir de diversos locais e possuírem diversos formatos. Segundo Olson e Delen (2008), após definir os objetivos, é necessário conhecer os dados visando:

- a) Descrever de forma clara o problema;
- b) Identificar os dados relevantes para o problema em questão;
- c) Certificar-se de que as variáveis relevantes para o projeto não são interdependentes;
- d) Normalmente as técnicas de agrupamento e de exploração visual também são utilizadas nesta etapa.

3. Preparação dos Dados: Devido às diversas origens possíveis, é comum que os dados não estejam preparados para que os métodos de Mineração de Dados sejam aplicados diretamente. Dependendo da qualidade desses dados, algumas ações podem ser necessárias. Este processo de limpeza dos dados geralmente envolve filtrar, combinar e preencher valores vazios, além de eliminar determinados caracteres incompatíveis.

4. Modelagem: É nesta fase que as técnicas (algoritmos) de mineração serão aplicadas. A escolha da técnica depende dos objetivos os quais se deseja atingir.

5. Avaliação: Considerada uma fase crítica do processo de mineração, nesta etapa é necessária a participação de especialistas nos dados, conhecedores do negócio e tomadores de decisão. Diversas ferramentas gráficas são utilizadas para a visualização e análise dos resultados (modelos). Testes e validações, visando obter a confiabilidade nos modelos, devem ser executados (*cross validation, supplied test set, use training set, percentage split*) e indicadores para auxiliar a análise dos resultados precisam ser obtidos (matriz de confusão,

índice de correção e incorreção de instâncias mineradas, estatística kappa, erro médio absoluto, erro relativo médio, precisão, *F-measure*, dentre outros).

6. Distribuição: Após executado o modelo com os dados reais e completos é necessário que os envolvidos conheçam os resultados.

Tradicionalmente, os métodos de mineração de dados são divididos em aprendizado supervisionado (preditivo) e não supervisionado (descritivo). Apesar do limite dessa divisão ser muito tênue (alguns métodos preditivos podem ser descritivos e vice-versa), ela ainda é interessante para fins didáticos.(Fayyad, 1996)

Para Camilo e Silva (2009), a diferença entre os métodos de aprendizado supervisionados e não supervisionados reside no fato de que os métodos não-supervisionados não precisam de uma pré-categorização para os registros, ou seja, não é necessário um atributo alvo. Tais métodos geralmente usam alguma medida de similaridade entre os atributos.

Segundo Mccue (2014), As tarefas de agrupamento e associação são consideradas como não supervisionadas. Já no aprendizado supervisionado, os métodos são providos com um conjunto de dados que possuem uma variável alvo pré-definida e os registros são categorizados em relação a ela. As tarefas mais comuns de aprendizado supervisionado são a classificação (que também pode ser não supervisionado) e a regressão.

Durante o processo de mineração, diversas técnicas devem ser testadas e combinadas afim de que comparações possam ser feitas e então a melhor técnica (ou combinação de técnicas) seja utilizada. O Presente estudo se deu através das seguintes etapas adaptadas do modelo CRISP-DM:

- a) Aquisição dos dados através do site: www.dados.gov.br;
- b) Pré-processamento e Higienização dos dados brutos e do banco de dados utilizado;
- c) Transformação em arquivos CSV. e posteriormente ARRF;
- d) Definição e utilização das seguintes variáveis de dados contidos na base de estudo: profissão, cidade de residência, estado civil, idade e Status de investimento (ativo ou inativo).
- e) Mineração através do Weka;
- f) Classificação dos dados através do algoritmo J48 utilizando a árvore de indução;
- g) Clusterização com *Simplekmeans*;
- h) Análise dos Resultados e Conclusão.

2.1. BASE DE DADOS UTILIZADA E SUA COMPREENSÃO

Segundo o arquivo de referência dos metadados dos investidores do TD, no site dados.gov.br O conjunto de dados utilizados contém a lista de investidores cadastrados no programa

Tesouro Direto. A listagem inclui dados do perfil do investidor como data de adesão, profissão, cidade de residência e estado civil, dentre outros.

Há também informação se o investidor está ativo ou não, o que significa que ele ainda realiza operações no programa, assim como existe um indicador se ele operou nos últimos 12 meses. Os investidores são identificados por um código único. Quando um investidor possui cadastro em mais de uma instituição financeira, esse outro cadastro é registrado em uma nova linha com o mesmo código de investidor.

2.2. PREPARAÇÃO E HIGIENIZAÇÃO DOS DADOS

Devido ao grande número de dados e a limitação do Excel para manipulação de um número de linhas acima de 1.048.576, foi utilizado o programa Microsoft Access do pacote office para que assim fosse possível a filtragem dos meses selecionados para o estudo e análise dos perfis de investidores. Posteriormente os mesmos foram copiados para uma planilha em Excel.

Foram excluídas 1.067 linhas de um total 26.475 no mês de dezembro de 2015 e 3.024 linhas de um total de 62.829 em agosto de 2016 devido ao fato de que quando um investidor possui cadastro em mais de uma instituição financeira, esse outro cadastro é registrado em uma nova linha com o mesmo código de investidor. Caso tal exclusão de dados não fosse realizada causaria interferência direta no resultado final do perfil de investidor.

Em seguida a exclusão dos dados repetidos o arquivo foi transformado em formato csv, a fim de permitir a utilização do mesmo no Weka e posteriormente transformado em formato arrf.

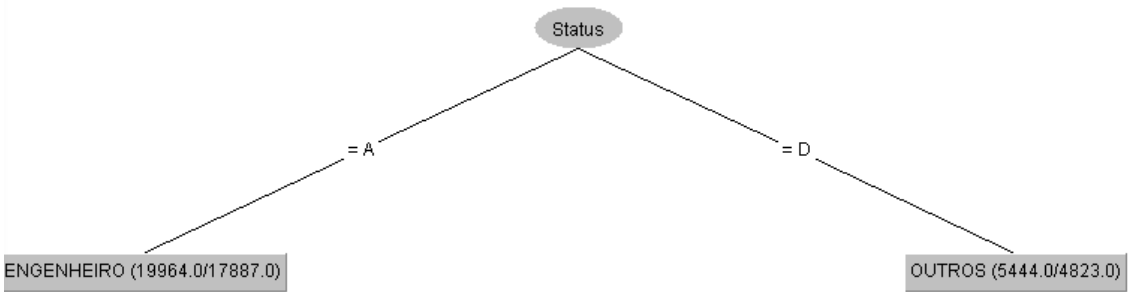
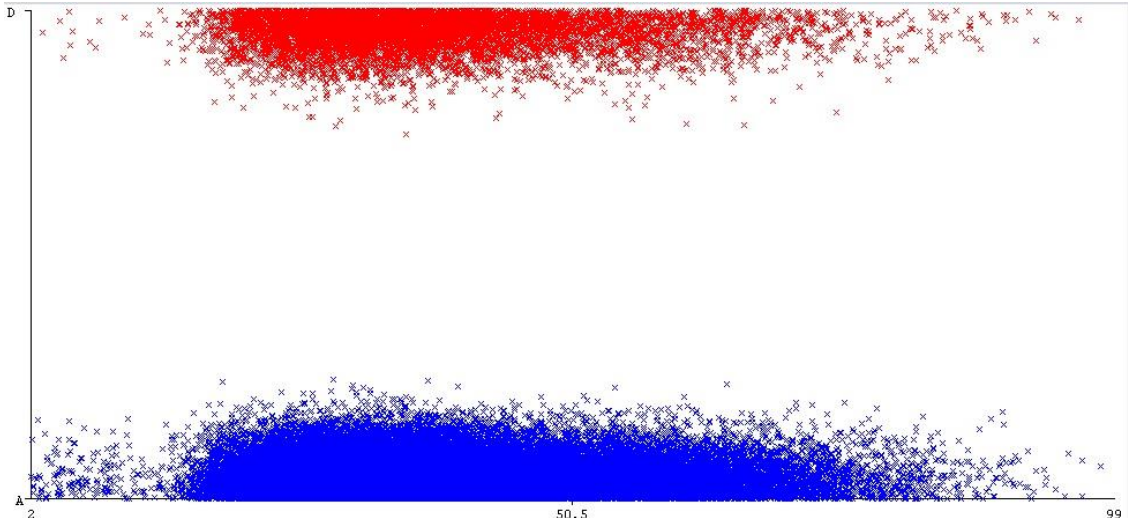
Após a etapa de limpeza, os registros do campo “Data de Adesão” passou pelo processo de transformação de dados. O procedimento realizado foi modificar o formato do campo de “dia/mês/ano” para “mês”, afim de que após formatação, as análises a serem realizadas fossem otimizadas e fizessem mais sentido na etapa de interpretação dos dados.

3. RESULTADOS

As técnicas realizadas no banco de dados foram aplicadas de modo que identificassem o perfil do investidor, tanto inativo quanto ativo, de forma a identificar características recorrentes entre eles. Dessa forma, os métodos utilizados foram: classificação, clusterização e algoritmo de árvore de indução. As técnicas utilizadas se prontificaram em responder as seguintes questões:

- 1) Qual a relação entre a profissão e a situação da conta?
- 2) Qual a relação entre a idade e a situação da conta e a média de idade dentre elas?
- 3) Qual a relação entre o estado civil e atividade da conta?
- 4) Quais fatores influenciam a conta estar ativa ou inativa?
- 5) Quais os principais grupos formados dentro das contas ativas?
- 6) Quais os principais grupos formados dentro das contas inativas?
- 7) Qual a média de idade daqueles que aderiram ao programa do TD naqueles meses de incerteza nacional?
- 8) Qual unidade federativa teve mais adesões em cada um dos meses analisados?
- 9) Quais profissões tiveram mais adesões dentre os meses analisados?
- 10) Qual o perfil mais recorrente dentre os 2 meses estudados e o perfil mais recorrente de cada mês?
- 11) Quais os fatores induzem ao estado civil dos investidores que aderiram ao programa no mês de dezembro de 2015?

R1	<p>Quando analisado o mês de dezembro de 2015, é possível notar uma relação de atividade e a profissão engenheiro, enquanto a inatividade está mais relacionada à classe de outras profissões conforme a Figura 1 abaixo. No mês de agosto de 2016 foi realizada uma clusterização a fim de definir essa relação e os resultados obtidos foram:</p> <p>Cluster #0: Contador, Ativo;</p>
----	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

	<p>Cluster #1: Servidor público federal, Ativo; Cluster #2: Estudante, Ativo; Cluster #3: Advogado, Inativo; Cluster #4: Vendedor Pracista Representante Comercial Caixeiro Viajante, Ativo.</p>  <p>Figura 1. Árvore de Decisão - Profissão x Atividade – Dezembro de 2015. Fonte: Elaborado pelo autor.</p>
R2	<p>No mês de dezembro de 2015 a relação e a média de idade tanto quanto relacionado à atividade ou inatividade da conta são bem similares, com média de 40 anos aproximadamente para ambos. No mês de agosto de 2016 obteve-se uma média de 38 anos para as contas ativas e inativas. É possível notar pela Figura 2 que no mês de agosto de 2016 a inatividade diminuiu, devido ao fato do impeachment estar praticamente consolidado naquele momento.</p>  <p>Figura 2. Gráfico - Idade x Atividade – Agosto de 2016. Fonte: Elaborado pelo autor. Legenda: A – Atividade (Eixo Y); D – Inatividade (Eixo Y); Idade (Eixo X).</p>
R3	<p>Utilizando o algoritmo SimpleKMeans de clusterização em dezembro de 2015 foram detectados 3 grupos:</p> <p>#0: Casado (a) com brasileiro (a) nato (a), ativo; #1: Solteiro (a), ativo; #2: Solteiro (a), inativo.</p> <p>Em agosto 2016 temos:</p>

	#0: Casado (a) com brasileiro (a) nato (a), Ativo; #1: Casado (a) com brasileiro (a) nato (a), Inativo; #2: Divorciado (a), Ativo.																		
R4	A partir dos resultados obtidos é possível concluir que um dos fatores que influencia a inatividade da conta é a faixa etária abaixo de 40 anos, sexo masculino, e o estado civil ser solteiro. Com relação à atividade pode-se chegar à seguinte conclusão: o grande fator de influência da atividade é o sexo masculino, sendo esse fator preponderante para atividade da conta correspondendo a 76% destes.																		
R5	<p>Os 2 principais grupos formados nas contas ativas foram:</p> <table><tr><td></td><td>Grupo 1</td><td>Grupo 2</td></tr><tr><td>Estado Civil</td><td>Solteiro (a)</td><td>Casado (a) com brasileiro (a) nato (a)</td></tr><tr><td>Sexo</td><td>M</td><td>M</td></tr><tr><td>Profissão</td><td>OUTROS</td><td>ENGENHEIRO (a)</td></tr><tr><td>Idade</td><td>33</td><td>45</td></tr><tr><td>UF</td><td>SP</td><td>SP</td></tr></table>		Grupo 1	Grupo 2	Estado Civil	Solteiro (a)	Casado (a) com brasileiro (a) nato (a)	Sexo	M	M	Profissão	OUTROS	ENGENHEIRO (a)	Idade	33	45	UF	SP	SP
	Grupo 1	Grupo 2																	
Estado Civil	Solteiro (a)	Casado (a) com brasileiro (a) nato (a)																	
Sexo	M	M																	
Profissão	OUTROS	ENGENHEIRO (a)																	
Idade	33	45																	
UF	SP	SP																	
R6	<p>Os 2 principais grupos formados nas contas inativas foram:</p> <table><tr><td></td><td>Grupo 1</td><td>Grupo 2</td></tr><tr><td>Estado Civil</td><td>Solteiro (a)</td><td>Casado (a) com brasileiro (a) nato (a)</td></tr><tr><td>Sexo</td><td>M</td><td>M</td></tr><tr><td>Profissão</td><td>OUTROS</td><td>ADMINISTRADOR</td></tr><tr><td>Idade</td><td>33</td><td>46</td></tr><tr><td>UF</td><td>SP</td><td>SP</td></tr></table>		Grupo 1	Grupo 2	Estado Civil	Solteiro (a)	Casado (a) com brasileiro (a) nato (a)	Sexo	M	M	Profissão	OUTROS	ADMINISTRADOR	Idade	33	46	UF	SP	SP
	Grupo 1	Grupo 2																	
Estado Civil	Solteiro (a)	Casado (a) com brasileiro (a) nato (a)																	
Sexo	M	M																	
Profissão	OUTROS	ADMINISTRADOR																	
Idade	33	46																	
UF	SP	SP																	
R7	A média de idade nos meses estudados foi de 39 anos, com idade mínima de 2 e a máxima de 99 anos, a mediana foi de 36 anos.																		
R8	O estado com mais adesões no mês de dezembro de 2015 foi o de São Paulo com 11.499, seguido pelo Rio de Janeiro com 3.483. Em agosto de 2016 o mesmo se repete com SP com 25.357 e RJ com 6.621 uma proporção de um do RJ para cada 4 de SP aproximadamente.																		
R9	Em dezembro de 2015 a categoria outras profissões, seguida de Engenheiro (a). No mês de agosto de 2016, outras profissões seguida de Administrador (a).																		
R10	<p>Por meio de clusterização podemos identificar dois tipos de perfis no mês de dezembro de 2015: casado (a) com brasileiro (a) nato (a), M, engenheiro, 39 anos; SP e Solteiro (a), F, outras profissões, 43 anos, SP.</p> <p>No mês de agosto de 2016 temos as seguintes associações: Solteiro (a), M, outras profissões, 37 anos, SP e Solteiro (a), F, administrador, 41 anos, SP.</p>																		
R11	Análise feita através da árvore de indução utilizando o algoritmo J48. De acordo com a figura 3 a seguir utilizando como critério alvo o estado civil do investidor. É notório que quando o investidor tem uma idade inferior a 33 anos tende a ser um investidor de estado civil igual a: solteiro (a).																		

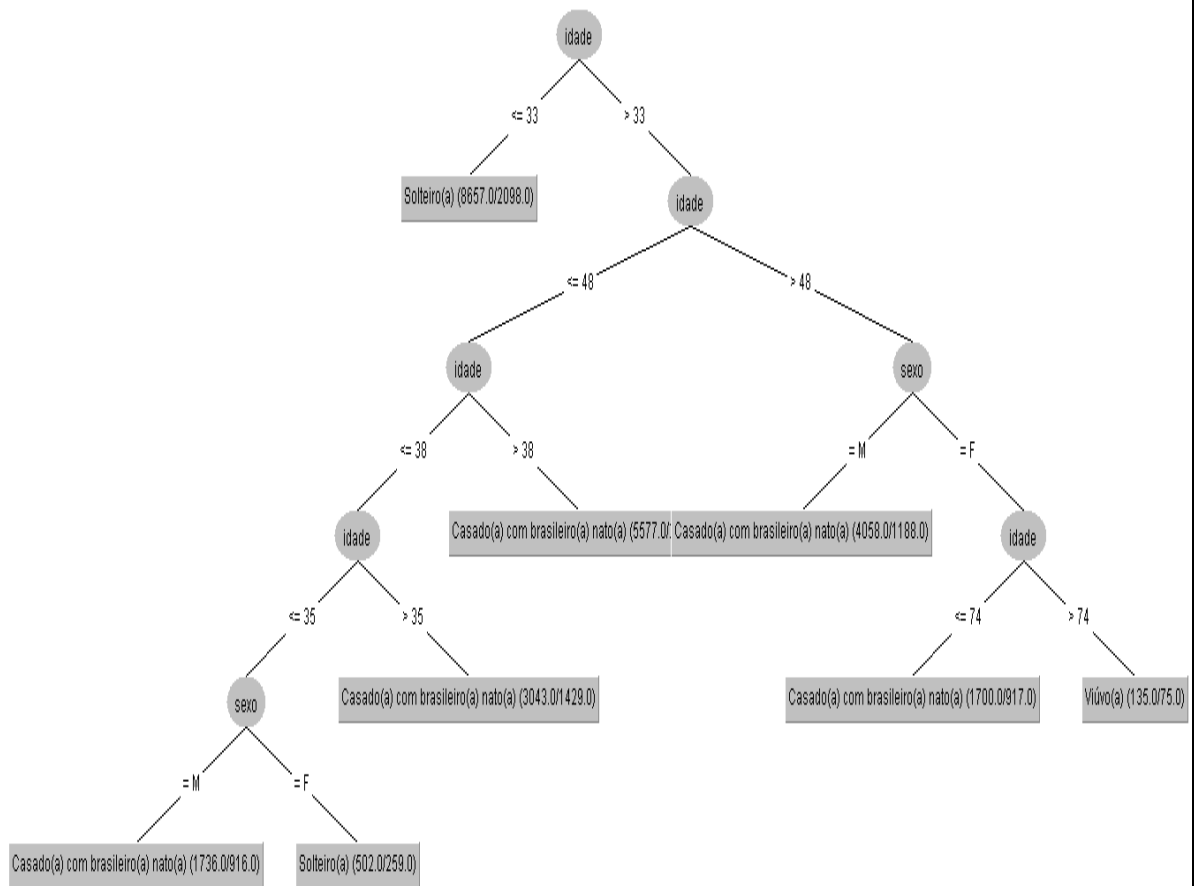


Figura 3. Árvore de indução da variável alvo – Estado Civil – Dezembro de 2015.
Fonte: Elaborado pelo autor.

4. CONCLUSÃO

Através dos resultados obtidos é notório que se possa chegar a algumas conclusões ao comparar a análise dos perfis em cada um dos meses estudados. O estado de São Paulo e Rio de Janeiro são os estados onde o TD teve mais adesões em ambos os meses (Dezembro de 2015 e Agosto de 2016). Houve uma melhora considerável no ponto de vista da atividade das contas uma vez que em agosto já havia sido votado o impeachment pelo senado, o que fez com que os proprietários tivessem mais segurança ao negociar os títulos.

No que diz respeito à faixa etária dos investidores conclui-se de que se trata de um público que está na faixa dos 30 até 45 anos em sua grande maioria, o que nos faz chegar a um consenso de que são pertencentes a uma classe que já está consolidada no mercado de trabalho e pretende fazer planos para o futuro em longo prazo, já que também em sua grande maioria estão casados (as) e em profissões como administrador (a) ou engenheiro (a).

Com relação aos perfis relacionados à atividade ou inatividade dos meses estudados, não se pode notar uma diferença entre os perfis dos grupos formados pela clusterização com o algoritmo *Simplekmeans* do Weka. Nos dois casos obtiveram-se perfis bem similares conforme R5 e R6. Outro fator de notoriedade é que a aderência se deu em sua grande maioria (76%) pelos investidores do sexo masculino.

Em futuras pesquisas, sugere-se que sejam realizadas análises comparativas nos meses posteriores, a fim de se comparar se houve uma melhora efetiva na atividade das

contas do TD, assim como na movimentação das mesmas, uma vez que o vice-presidente já havia sido empossado e o país começou a ter uma melhora no âmbito econômico fazendo com que os investidores tivessem mais estabilidade para uma tomada de decisão.

5. REFERÊNCIAS BIBLIOGRÁFICAS

- [1] BALANÇO E ESTATÍSTICAS - TESOURO DIRETO - STN. (2018). Recuperado: 8 de novembro de 2018, de <http://www.tesouro.gov.br/pt/balanco-e-estatisticas>.
- [2] CAMILO, C. O. (2009). **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**, 29.
- [3] FAYYAD, U. (1996). **From Data Mining to Knowledge Discovery in Databases**, 18.
- [4] FERREIRA, L. V. B. (2016). **Investimento em títulos públicos: Uma análise de investimento em títulos públicos do Tesouro Nacional em um cenário de crise econômica**. Universidade Federal Do Rio Grande Do Sul Escola De Administração Departamento De Ciências Administrativas, 114.
- [5] FILHO, P. J. R., & ALVES, D. F. (2017). **Peso dos títulos públicos brasileiros negociados via tesouro direto na composição da Dívida Pública Mobiliária Federal Interna (DPMFI)**, 11.
- [6] LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining**. John Wiley and Sons, Inc, 2005.
- [7] LOURO, A., ANDRADE VIEIRA, C., & DA SILVA BIZZI, M. (2017). **Análise preditiva do perfil dos investidores do tesouro direto para ações de marketing**.
- [8] MCCUE, C. (2014). **Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis**. Butterworth-Heinemann.
- [9] OLSON, D. L., & DELEN, D. (2008). **Advanced Data Mining Techniques**. Springer Science & Business Media.
- [10] PERLIN, M. (2016). **The Microstructure of Tesouro Direto: The Sazonality of the Order Flow and the Formation of Spreads**. *Economia Aplicada*, 20(3), 253.
- [11] PIMENTEL, B. C. *ET AL.* **TÍTULOS PÚBLICOS: Uma alternativa de investimento com ganhos reais e seguro**. Simpósio de excelência em gestão e tecnologia. 2015.
- [12] VERGARA, S. C. **Projetos E Relatórios De Pesquisa Em Administração**. [S.L.] Atlas, 2009.