FACULDADE PROFESSOR MIGUEL ÂNGELO DA SILVA SANTOS – FeMASS
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

PROCESSO DE KDD UTILIZANDO ALGORITIMOS DE RANDOM FOREST E DECISION TREE PARA ANÁLISE DE DADOS

POR:
MATHEUS SOARES ROCHA

MACAÉ
2020

FACULDADE PROFESSOR MIGUEL ÂNGELO DA SILVA SANTOS – FeMASS
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

MATHEUS SOARES ROCHA

PROCESSO DE KDD UTILIZANDO ALGORITIMOS DE RANDOM FOREST E DECISION TREE PARA ANÁLISE DE DADOS

Trabalho Final apresentado ao curso de graduação em Sistemas de Informação, da Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS), para obtenção do grau de BACHAREL em Sistemas de Informação.

Professor Orientador: Sérgio Eduardo Corrêa Netto, M.e

MACAÉ/RJ

2020

MATHEUS SOARES ROCHA

PROCESSO DE KDD UTILIZANDO ALGORITIMOS DE RANDOM FOREST E DECISION TREE PARA ANÁLISE DE DADOS

Trabalho de Conclusão de Curso apresentado ao curso de graduação em Sistemas de Informação, da Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS), para obtenção do grau de BACHAREL em Sistemas de Informação.

Aprovada em ___ de _____ de 20___

BANCA EXAMINADORA

Sérgio Eduardo Corrêa Netto, M.e

Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS)

1º Examinador

Isac Mendes Lacerda, M.e

Faculdade Professor Miguel Ângelo da Silva Santos (FeMASS)

2º Examinador

EPÍGRAFE

“A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê.” (Arthur Schopenhauer)

RESUMO

Atualmente, na era da informação, obter conhecimento através dos dados é de grande importância para as diversas áreas da sociedade, o que justifica a relevância do tema dissertado. Este trabalho explica o processo de Descoberta de Conhecimento em Banco de Dados, o conceito geral sobre Mineração de Dados e sua aplicação em um banco de dados sobre empréstimos bancários, na qual foram utilizados os algoritmos de Random Forest e Decision tree para prever quais futuros clientes poderiam receber a aprovação da carta de crédito com base em registros anteriores.

Palavras-chave: Mineração de Dados. Descoberta de Conhecimento. KDD.

ABSTRACT

Currently, in the information age, obtaining knowledge through data have a lot importance for the different areas of society, which justifies the existence of the thesis being discussed. This work explains the process of Knowledge Discovery in Database, the general concept about Data Mining, the application in a database about bank loans, which was used the algorithms of Random Forest and Decision tree to predict which future customers could receive the letter of credit approval based on previous records.

Key words: Data Mining. knowledge discovery. KDD.

LISTA DE FIGURAS

LISTA DE TABELA

LISTA DE ABREVIATURAS E SIGLAS

SGBD	Sistema de gerenciamento de banco de dados
KDD	Knowledge Discovery in Databases
DM	Data Mining
MD	Mineração de dados
IOT	Internet Of Things (Internet das Coisas)
CRISP-DM	Cross-Industry Standard Process of Data Mining
E-Comerce	Comércio Eletrônico
ML	Machine Learning

SUMÁRIO

1 INTRODUÇÃO

Com o passar dos anos a quantidade gigantesca de dados de diversos formatos oriundos de pesquisas na internet, interações em redes sociais, dispositivos IOT, automóveis e satélites têm sido muito maiores do que os possibilidade humana de interpretá-los. Considerando também que com o passar do tempo as tecnologias de armazenamento começaram a ficar mais baratas, tornando mais viável o armazenamento de uma quantidade enorme de dados, o que há alguns anos era impossível de se fazer. Também foi necessário criar formas de se armazenar dados que fossem capazes de guardar esses dados tanto em tamanho quanto em estrutura, dados não estruturados, que são diferentes dos modelos tradicionais, utilizados pelos Sistemas de Gerenciamento de Banco de Dados (SGBD).

Segundo Gartner, Cerca de 2,2 milhões de terabytes – o suficiente para armazenar 88 milhões de filmes em qualidade Blu-Ray – é a quantidade de novos dados criados todos os dias no mundo e esse número deve crescer para 40 trilhões de terabytes diários já no ano de 2020. Vivemos a Era do Big Data.

Com tanta informação disponível hoje em dia, o que fazer com elas? Como aproveitá-las? Como utilizar de maneira efetiva toda essa quantidade de dados que está disponível?

Essa abundante e acessível quantidade de dados e a urgência de converter esses dados em conhecimento útil tornou a mineração de dados um assunto importante no setor da informação e da sociedade. As informações e conhecimentos conseguidos são empregados em várias áreas como na análise de mercado, detecção de fraude e retenção de clientes, controle de produção e exploração científica (HAN; KAMBER, 2006).

Devido ao grande volume de dados e por serem não estruturados, as ferramentas tradicionais não conseguem fazer uma análise adequada e com isso, foram se popularizando métodos e ferramentas capazes de trabalhar com essa enorme quantidade de dados e que tivesse a possibilidade de apoiar decisões humanas baseada nesses dados. Este conceito começou a ser chamado *Knowledge Discovery in Databases*, ou somente KDD.

O processo de KDD que é mostrado ao longo deste documento é composto por cinco etapas, que são: Seleção, Pré-Processamento, Formatação, Mineração de Dados e Avaliação/Interpretação. Ao final de todo esse processo, é possível gerar conhecimento sobre os dados que foram utilizados como entrada.

A expressão *Data Mining* (DM) ou Mineração de dados (MD), mais popular, é, na verdade, uma das etapas da Descoberta de Conhecimento em Bases de Dados, mas é muito referida também como se fosse o processo em si, mais adiante explicaremos a diferença.

A mineração de é definida, ultimamente, como um processo de descoberta de padrões em grandes quantidades de dados, de forma automática ou, na maioria das vezes, semiautomática, para a extração de informação previamente desconhecida, válida que gera ações úteis, e onde que os padrões descobertos são significativamente vantajosos para a tomada de decisões estratégicas (CABENA et al., 1998). Esses atributos têm atraído uma boa parte das atenções da indústria da tecnologia da informação, pois a mineração dos dados é apresentada como um resultado da evolução natural desta industria (HAN e KAMBER, 2001).

O objetivo deste trabalho é focar mais atentamente a etapa de MD utilizando–se de algoritmo de *Machine Learning* sobre base de dados com a intenção de obter informação, que por métodos convencionais seriam quase impossíveis de conseguir.

Os objetivos gerais deste trabalho é demonstrar como o processo de *Knowledge Discovery in Databases* (KDD) é utilizado em grandes quantidades de dados e a aplicação de mineração de dados para extração de conhecimento.

Os objetivos específicos deste trabalho são:

- Apresentar e explicitar como funciona o processo de KDD;
- Apresentar e descrever a base de dados escolhida;
- Demonstrar a utilização dos algoritmos de Machine Learning e realizar simulações até que a etapa esteja madura;
- Analisar e Descrever os resultados obtidos após os testes.

É notório a importância que informação tem em todos os setores da sociedade desde os tempos antigos, mas atualmente, esse acúmulo de dados vem ocorrendo em uma escala imensa e com isso abre espaço para muita extração de conhecimento que esteja implícita nos dados gerados pela sociedade.

Organizar dados em grupos é uma das formas mais naturais de compreensão e aprendizagem (JAIN; DUBES, 1988).

Porém, esse problema torna-se complexo à medida que o volume de dados a serem agrupados aumenta, o problema ganha complexidade e com isso as técnicas tradicionais e

manuais de análise de dados se tornam ineficientes ou mesmo impraticáveis. Sendo assim, é necessário o auxílio de ferramentas computacionais para se trabalhar sobre dados.

Segundo Fayyad (1996), esse acúmulo de dados necessita de uma nova geração e teorias e ferramentas computacionais para ajudar os humanos a extrair informações (conhecimento).

A aplicação de métodos de KDD tem obtido bastante sucesso em várias áreas do conhecimento. A pioneira delas foi na astronomia, era utilizado para processar 3 terabytes de imagens do *Second Palomar Observatory Sky Survey*. Outras áreas como bancos, servindo para controle de fraudes, ou análise para investimentos usando algoritmos neurais como também no Marketing, para procurar grupos de pessoais que compram certos produtos e as possibilidades de eles comprar determinados outros produtos.

Utilizar as técnicas de KDD na tentativa de se buscar conhecimento nesta nova realidade é de interesse de muitas áreas não só de negócio, mas de pesquisas e projetos por diversos motivos:

- Volume de dados: DM aplica-se a grandes quantidades de dados. Empresas como de Televisão, Bancos, Telecom, e-Comerce, e redes sociais tem produzido quantidades enormes de dados sobre seus negócios, de origens e formatos distintos, o que se caracteriza como *Big Data*, combinados ao processo de KDD e Mineração de Dados podem trazer bastante conhecimento aos seus respectivos negócios.
- Armazenamento de dados: Os novos modelos de armazenamento de dados como os *Data Lakes* ou *Data Warehouses* tem disponibilizado os dados de formas mais fáceis e acessíveis para a análise de dados.
- Avanços tecnológicos em *hardware*: Etapa de DM requer bastante recurso computacional para executar seus algoritmos e com esses avanços o acesso a *hardwares* que deem conta desse processamento se tornou muito mais fácil.
- Bancos de dados Distribuídos: Esse modelo de banco de dados incrementa mais ainda as possibilidades de assertividade na etapa de DM.
- Informação é o Poder: Atualmente as empresas mais valiosas do mundo são empresas que detém acessos a muitos dados, como também os utilizam para adquirir conhecimento.

Estre trabalho tem como base a pesquisa exploratória, de cunho bibliográfica, pois busca conceitos e teorias a partir de revisão de literatura em artigos científicos, livros, teses de mestrado e doutorado.

Os dados coletados para este trabalho são de base de dados relacionados a empréstimos bancários, e o método de análise foi o processo de KDD, juntamente com a aplicação dos algoritmos de machine learning. Os dados levantados, foram analisados e apresentados de forma quantitativa e qualitativa.

Este trabalho há uma introdução a qual contextualiza todo o ambiente a qual este processo é aplicado bem como o que o tornou viável sua aplicação atualmente. Os objetivos deste trabalho é demonstrar o processo de KDD em si e sua aplicação em bases de dados.

A justificativa para este trabalho é quanto este processo, que foi criado há tempos é tão atual e pode ser um caminho muito bem definido para a aplicação de algoritmos de machine learning independentemente de qual linguagem de programação ou *software* utilizar.

Inicialmente foi explorado o conceito do processo de KDD, como funciona a obtenção de conhecimento bem como as etapas que o compões e os elementos que o apoiam. Posteriormente teve um foco especial na etapa de Data Mining em que é apresentado alguns conceitos sobre algoritmos aprendizado de máquina e as melhores aplicações para cada modelo. O capítulo três foi aplicado o processo seguindo cada etapa definida no referencial teórico, aqui optou-se por utilizar o processo de KDD e não o de CRISP-DM, e ao final houve a avaliação de como funcionou o modelo e as predições que ele fez.

Nas considerações finais se obteve sucesso ao que foi proposto neste trabalho, apresentar e aplicar o todo o processo descrito no referencial teórico bem como todos os objetivos foram alcançados nesta aplicação.

2 PROCESSOS DE ANÁLISE DE DADOS

2.1 Processo De KDD

Segundo Fayyad (1996), a terminologia *"knowledge discovery in database"* foi utilizada pela primeira vez em 1989 (por Piatetsky-Shapiro) com objetivo de destacar que o conhecimento é o produto de uma descoberta baseada em dados. É popular no mercado da inteligência artificial e Aprendizado de Máquina.

O termo *Data Mining* (DM) ou Mineração de dados (MD) é mais utilizado como referência a KDD por estatísticos, analista de dados e profissionais da área da tecnologia da informação.

Em *"From Data Mining to Knowledge Discovery in Databases",* Fayyad, Piatetsky-Shapiro e Smyth (1996), o seu cerne é diferenciar Knowledge Discovery in Databases (KDD) de *Data Mining* ou Mineração de Dados (DM). Para eles, o processo de KDD se refere a uma coleção de passos e ou processos que visam a descoberta de conhecimento que seja relevante a partir dos dados, enquanto DM seria apenas um passo dentro deste processo de descoberta que envolve a fase de modelagem de dados.

"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados". (FAYYAD et al. ,1996, p. 40). Segundo Goldschimidt & Passos (2005) o termo iterativo sugere a que pode haver possibilidade de repetições parciais ou totais do processo de KDD, e a expressão não trivial atenta para a complexidade normalmente presente na execução do processo. Com relação a expressão padrão válido aponta que o conhecimento deve ser verdadeiro e apropriado ao contexto da aplicação de KDD e o termo padrão novo deve adicionar novos conhecimentos aos existentes, para que todo esse processo crie conhecimento útil que pode ser aplicado de forma a proporcionar benefícios ao contexto de aplicação de KDD.

O processo de KDD aglomera uma sequência de cinco etapas: Seleção, Pré-processamento ou Limpeza, Transformação, Mineração de Dados e Interpretação ou Avaliação dos resultados. No conjunto de obras literárias sobre KDD há definições apresentadas sob a perspectiva de diversos autores. Houve a escolha por fazer um estudo da

descrição das etapas tendo como base duas obras sobre o tema e em seguida é apresentada uma comparação entre as definições.

2.1.1 Etapas KDD Segundo Fayyad

A figura 1 representa as etapas do processo de KDD segundo Fayyad (1996).

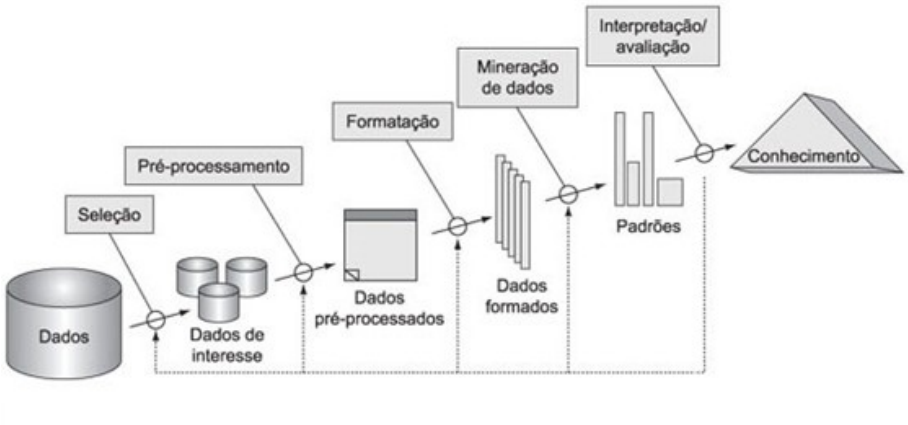


Figura 1: Etapas do processo KDD Fayyad et al. (1996). Fonte: Adaptação de Fayyad et al. (1996)

Saber o qual é o problema a ser resolvido é a parte crucial do processo de descoberta de conhecimento em banco de dados. Pois a partir desta premissa, é possível utilizar as ferramentas do processo de KDD de forma correta chegando ao final de todo o processo com uma solução a ser utilizada.

- **Seleção de dados**

Após definido o problema a ser resolvido, o processo é iniciado com a seleção dos dados. Esta etapa pode ser crítica pois os dados podem não estar disponíveis em uma forma correta para se utilizar no processo de KDD. Ou mesmo, precisar de um especialista que esclareça como os dados estão, uma coisa que é muito comum em organizações. Os analistas de dados que vão coletar dados (podem não entender ou) não entendem como funciona o armazenamento dos dados e como estão disponibilizados e acabam recorrendo a quem sabe.

Outro problema nesta etapa é descobrir a localização em que os dados estão no banco de dados. A *engenharia de software* é uma padronização de criação de sistemas que está sendo utilizada de forma mais ampla recentemente, portanto, sistemas gerenciais mais antigos podem não ter uma documentação ou mesmo um padrão a ser seguido quanto a armazenamento de dados. Isto acaba tornando o processo de coleta de dados mais difícil.

Na fase de seleção é escolhido o conjunto de dados, armazenados em um domínio, onde são encontradas os registros e variáveis que serão utilizados. É um processo difícil, pois os dados podem vir de diversas fontes como planilhas em Excel, formato CSV, dos sistemas de informação, dados vindos da internet, de *Data Warehouses,* ou mesmo dados que não seguem uma estrutura como imagens e arquivos de áudio.

- **Pré-Processamento**

É muito comum encontrar uma base de dados na qual os dados estejam desorganizados, com valores incorretos, dados que não servem para serem utilizados nos modelos preditivos, disparidade grande entre valores de uma coluna, dentre outros problemas. Para os modelos que serão utilizados na etapa de MD tenham resultados superiores ao esperado ou mesmo que desenvolvam seu correto funcionamento, é necessário que estes dados passem por uma limpeza. É nesta etapa que os dados inconsistentes, redundantes e desnecessários para o modelo são limpos do conjunto.

"Nesta fase busca-se aprimorar a qualidade dos dados" (BATISTA,2003, pág.35).

- **Formatação**

Após a limpeza, esse conjunto de dados ainda pode não estar no padrão ideal para a utilização dos algoritmos de MD, portanto, esta etapa visa transpassar as limitações do algoritmo que será utilizado. Por exemplo, aqui o conjunto de dados está totalmente correto e de acordo com a solução do problema, porém por questões de localidade pode se ter problemas como comumente se vê em arquivos de dados, o algoritmo para funcionar de forma correta, a data deve estar no formato MM-dd-yyyy e não como se utiliza no Brasil que é o formado dd-MM-yyyy.

É nesta etapa que os dados são formatados e estruturados para que atendam por completo as exigências do modelo de algoritmo e os dados obtidos de diversas fontes são postos em um único repositório. Desta forma, Silva (2000) indica que já se tenha definido a

técnica de mineração e o algoritmo minerador que serão utilizados para a partir de então transformar os dados para o formato adequado.

- **Mineração de Dados**

Esta é a etapa em que se coloca os algoritmos mineradores para buscar conhecimento implícito nos dados pré-processados e formatado, com finalidade de encontrar padrões nos mesmos.

Esta etapa é que recebe maior destaque em todo o processo. É aplicado o algoritmo de mineração mais apropriado, que tem como entrada os dados do repositório gerado através da etapa anterior, com a finalidade de se obter algum resultado proveitoso e que será enfim interpretado e avaliado na última fase (REZENDE, 2003).

- **Avaliação/Interpretação**

O principal objetivo desta etapa é melhorar a compreensão do conhecimento adquirido, geralmente em modelos de relatórios explicativos, com uma documentação de todo o processo e explicação das informações relevantes encontradas no processo de KDD. Após a análise do conhecimento, caso este não seja de interesse do usuário final ou não cumpra com os objetivos propostos, o processo de extração pode ser repetido ajustando-se os parâmetros ou melhorando o processo de escolha dos dados para a obtenção de resultados melhores numa próxima iteração. (REZENDE, 2003).

2.1.2 Etapas do CRISP-DM Segundo Chapman

Cross Industry Standard Process for Data Mining (CRISP-DM) é uma metodologia que surgiu anos após a criação de KDD com objetivos semelhantes, mas voltado a resolver problemas de negócio nas organizações.

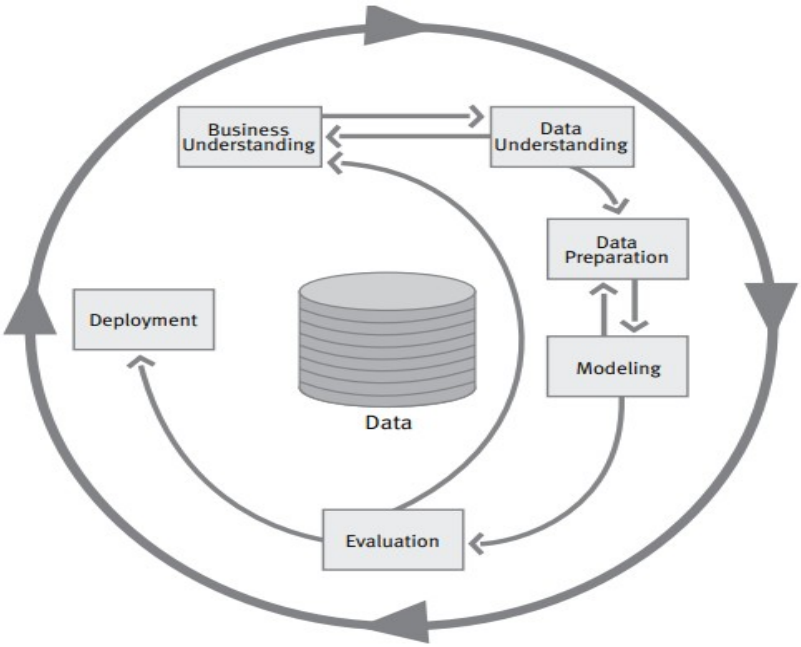


Figura 2: Fases do modelo de referência CRISP-DM (Chapman et al., 2000)

- **Entendimento de Negócio**

Essa fase inicial concentra-se no entendimento dos objetivos e requisitos do projeto sob uma perspectiva de negócio, transformando esse conhecimento em uma definição de problema de mineração de dados e em um plano preliminar projetado para atingir os objetivos.

A ideia geral deste passo é entender como é a organização, como funcionam os processos, como está a situação do negócio no início do projeto, o que os clientes esperam desse projeto e não só os objetivos a serem conquistados, mas também a identificação dos recursos financeiros e humanos que possam ser usados no decorrer do projeto.

- **Entendimento dos Dados**

Esta fase começa com a coleta inicial de dados e prossegue com as atividades que permitam que o analista se familiarize com os dados, identifique problemas de qualidade dos dados, descobrir as primeiras ideias sobre os dados e / ou detectar subconjuntos interessantes para formar hipóteses sobre informações implícitas.

- **Preparação dos Dados**

A fase de preparação de dados engloba todas as atividades necessárias para construir o conjunto de dados, dados esses que serão alimentados na ferramenta de modelagem. Provavelmente esta etapa será executada uma certa quantidade de vezes. Tarefas que incluem a seleção de tabelas, registros e atributos, além de transformação e limpeza dos dados.

- **Modelagem**

Nesta etapa várias técnicas de modelagem, que no caso são os algoritmos de mineração, são escolhidas e aplicadas. Aqui se busca obter o algoritmo que tenha a melhor performance para o problema a ser resolvido. Geralmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados, é comum se usar também diversos algoritmos. Algumas técnicas têm bastante peculiaridades quanto ao formato em que os dados estão. Portanto, muitas vezes é necessário voltar à fase de preparação de dados.

- **Avaliação**

Esta etapa tem o objetivo de analisar os resultados do modelo implementado, e se necessário, fazer revisões para garantir que ele atinja de forma adequada os objetivos do negócio da empresa. É importante avaliar o modelo de forma cautelosa antes de sua implementação a fim de verificar se ele atinge de forma adequada os objetivos da empresa. No final desta fase, uma decisão sobre o uso dos resultados da mineração de dados deve ser alcançada.

- **Implementação**

Nesta etapa, o modelo, após todas as etapas anteriores estiverem sido totalmente satisfeitas é de fato implementado na organização. Este resultado pode ser complexo como implementar um algoritmo de mineração que fique sendo executado na empresa, ou simples como a criação de um relatório. Nesta etapa é que também são feitas as modificações quando necessárias do modelo implementado.

2.1.3 Estudo Comparativo Entre Os Modelos De KDD e CRISP-DM

O processo KDD (Fayyad et al., 1996) se estrutura como um modelo de processo porque se estabelece todas as etapas que devem ser seguidas para desenvolver um projeto de DM, mas não é uma metodologia porque ele não define como se deve ser feita cada uma das propostas no modelo.

O modelo CRISP-DM (Chapman et al., 2000) define quais tarefas devem ser executadas para se obter um projeto de DM, é, portanto, um modelo de processo também. Diferentemente do KDD este processo é em modelo cascata, pois em uma certa etapa não se pode voltar a anterior, em tese. O CRISP-DM possui uma metodologia porque fornece recomendações sobre como se executar as tarefas. Mesmo assim, essas recomendações se limitam a propor outras tarefas e não oferecem um guia sobre como fazê-las. O que ainda o caracteriza como um modelo também.

Tabela 1 - Um estudo comparativo de modelos de processos de mineração de dados mineração Sinônimos de mineração Substantivo terraplenagem (KDD e CRISP-DM). Fonte: Adaptação Shafique, Umair & Qaiser, Haseeb. (2014)

Modelo	**KDD**	**CRISP-DM**
Nº Passos	9	6
Nome dos passos	Desenvolvimento e entendimento da aplicação	Entendimento do negócio
	Criação do conjunto de dados	Entendimento dos dados
	Limpeza e pré-processamento de dados	
	Transformação dos dados	Preparação dos dados
	Escolher a tarefa de mineração de dados adequada	Modelagem
	Escolher o algoritmo de mineração de dados adequado	
	Empregar o algoritmo de mineração de dados	
	Interpretação de padrões	Avaliação
	Utilizar conhecimento adquirido	Implementação

Como não se pretende aplicar o processo para resolver um problema de negócio em uma empresa ou organização, acabou optando se por dar continuidade na demonstração do processo de KDD.

2.2 Mineração De Dados (Data Mining)

A mineração de dados pode ser entendida como a exploração e análise de grandes quantidades de dados, de maneira automática ou semiautomática, com o objetivo de descobrir

padrões e regras relevantes (BERRY; LINOFF, 1997, P. 5). O principal objetivo do processo de DM é fornecer às empresas conhecimento que as possibilitem montar melhores estratégias de marketing, vendas e suporte, e com isso melhorando seus negócios.

Atualmente as organizações utilizam Sistemas de Gerenciamento de Banco de Dados (SGBD), conseguem armazenar dados oriundos de operações diárias, porém, grande parte dessas empresas não conseguem utilizar esses dados para obter conhecimento sobre seu negócio.

O conceito de mineração é gradativamente mais utilizado como uma tarefa de descoberta de informações, ela possibilita retirar de grandes conjuntos de dados um resultado que melhore e facilite as escolhas baseados nos dados minerados (HOSKING, 1997)

Com frequência a mineração de dados vem sendo considerada uma mistura de pesquisas e estatísticas, bancos de dados e inteligência artificial. MD é a parte de um processo de pesquisa denominado Busca de Conhecimento em Banco de Dados (*Knowledge Discovery in Database* - KDD), descrito na seção anterior, no qual possui própria metodologia para a preparação e exploração dos dados, interpretação dos seus resultados e assimilação dos conhecimentos minerados (HAN; KAMBER, 2011).

O processo de transformação da informação em conhecimento teve um crescimento em larga escala devido ao avanço tecnológico dos hardwares, a utilização do método de KDD proporcionou um grande avanço e uma maior facilidade e agilidade na extração de informação, possibilitando a análise de bases de dados antes inacessíveis e obtendo conhecimento que anteriormente não era possível. (PASSOS, 2006).

2.2.1 Métodos E Técnicas De Mineração De Dados

Tradicionalmente as técnicas de MD se dividem em algoritmos de modelo preditivo e descritivo, ou, como é chamado atualmente, aprendizado supervisionado e não-supervisionado. (HAN; KAMBER, 2006).

Segundo (FAYYAD et al., 1996), vai mais um pouco além onde os objetivos da descoberta de conhecimento podem ser definidos em dois: verificação onde o sistema se limita a analisar a hipótese do especialista e descoberta onde o objetivo é encontrar novos padrões. A descoberta é subdividida em predição e descrição. Preditivo, onde o sistema encontra padrões com o objetivo de prever o comportamento futuro de algumas entidades; e

descritivo, onde o sistema encontra padrões com a finalidade de apresentá-los a um usuário em uma forma compreensível.

A maioria dos métodos de mineração de dados é baseada em técnicas experimentadas e testadas de aprendizado de máquina (ML), reconhecimento de padrão e estatística: classificação, agrupamento, regressão dentre outros.

Embora os limites entre preditivo e descritivo não sejam claros (alguns dos modelos preditivos podem ser descritivos, na medida em que sejam compreensíveis e vice-versa), a distinção é útil para entender o objetivo geral da descoberta. A relativa importância do modelo preditivo e descritivo para as aplicações especificas de DM pode variar de forma considerável. Os objetivos da previsão e descrição pode ser obtida usando uma variedade de métodos específicos de mineração de dados (FAYYAD et al., 1996),

- **Classificação**

Classificação é aprender uma função que mapeia (classifica) um item de dados em uma das várias classes predefinidas.

A classificação é uma tarefa da mineração de dados que consiste em avaliar os dados processados, classificando-os de acordo com as suas características. Para isso criam-se classes caracterizadas, e os dados processados são relacionados a essa classe por meio das peculiaridades (SILVA, 2000).

A **figura 3** mostra uma repartição simples dos dados de empréstimo em duas regiões de classe; note que não é possível separar as classes perfeitamente usando um método linear de decisão. O banco pode querer usar as regiões de classificação para automaticamente decidir se os futuros solicitantes de empréstimos serão aprovados ou não.

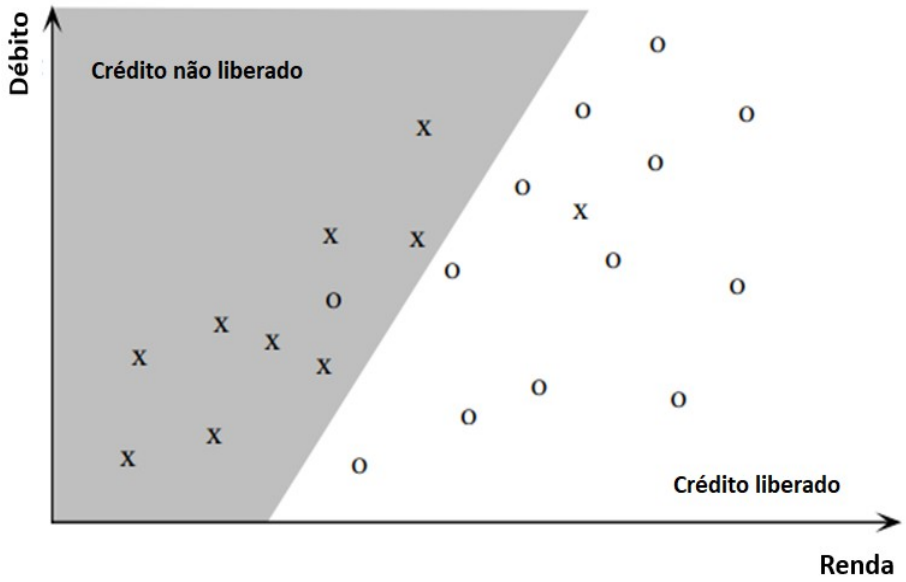


Figura 3: Classificação sobre um conjunto de dados de empréstimos. Fonte: Adaptação de Fayyad et al. (1996)

- **Regressão**

É uma função que mapeia um item de dados para uma variável de previsão com valor real, em outras palavras, ela é bem similar a classificação, porém não visa classificar identificar as variáveis por atributos numéricos.

Diferente da classificação, a estimação é usada quando o registro é identificado por um valor numérico e não um categórico (CAMILO; SILVA, 2009).

Um exemplo seria predição da soma da biomassa presente em uma floresta; estimativa da probabilidade de um paciente sobreviver, dado o resultado de um conjunto de diagnósticos de exames; predição do risco de determinados investimentos, definição do limite do cartão de crédito para cada cliente em um banco; dentre outros. Estatística, Redes Neurais, dentre outras áreas, oferecem ferramentas para implementação da tarefa de regressão (MICHIE et al., 1994).

A **figura 4** mostra o resultado de simples regressão linear em que a dívida total é ajustada como função linear da renda: o ajuste é ruim porque existe apenas uma fraca correlação entre as duas variáveis.

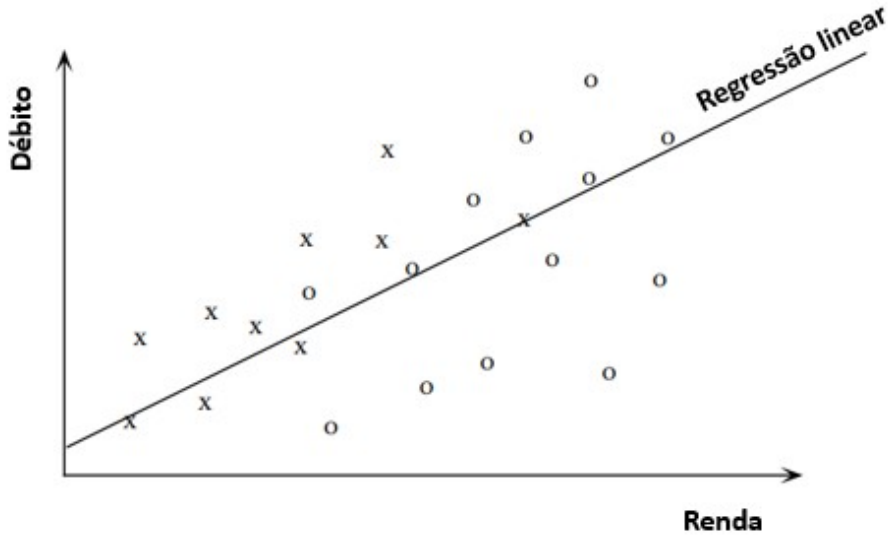


Figura 4: Regressão linear sobre um conjunto de dados de empréstimos. Fonte: Adaptação de Fayyad et al. (1996)

- **Segmentação (clusterização)**

Utilizada para dividir os registros de uma base de dados em grupos (clusters), de tal forma que os elementos dos grupos compartilhem propriedades comuns que os diferencie dos outros grupos. Tem como objetivo maximizar as similaridades dos elementos dos clusters e minimizar as similaridades dos clusters. Diferente da tarefa de classificação, que tem rótulos predefinidos, a segmentação precisa automaticamente identificar os grupos de dados aos quais o usuário deverá atribuir rótulos (Fayyad et al., 1996a)

Possui muitas aplicações geralmente na área de Marketing, como, por exemplo, agrupar clientes com comportamento de compra bem parecidos, a fim de recomendar de forma melhor novos produtos para esses clientes. Na implementação desta tarefa podem ser utilizados algoritmos tais como: K-Means, K-Modes, K-Prototypes, K-Medoids, Kohonen, dentre outros. (PASSOS, 2006).

A **figura 5** mostra um possível agrupamento do conjunto de dados do empréstimo em três grupos; observe que os clusters se sobrepõem, permitindo que os dados pontos para pertencer a mais de um cluster. Os rótulos de classe originais (indicados por **x** e **o** nas figuras anteriores) foram substituídos por um + para indicar que a associação à classe não é mais assumida conhecida.

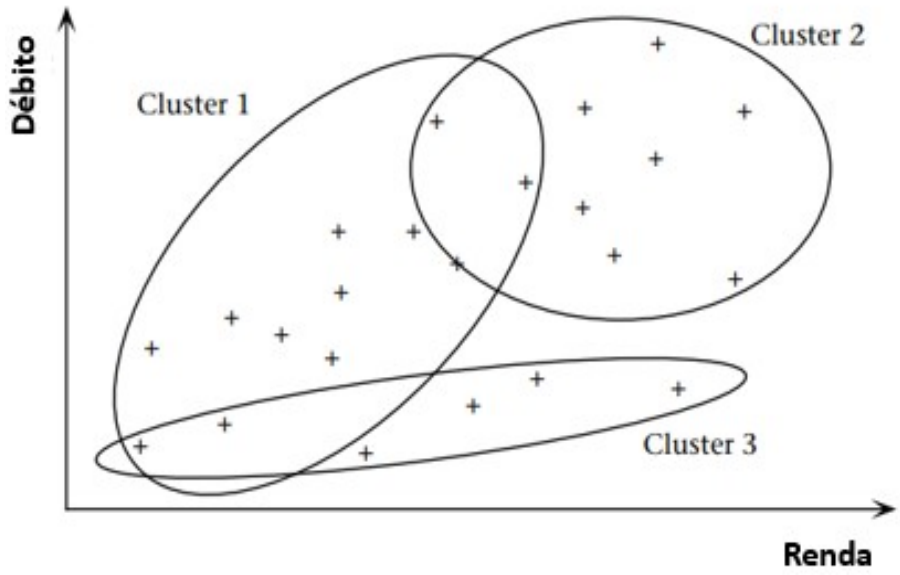


Figura 5: Segmentação sobre um conjunto de dados de empréstimos. Fonte: Adaptação de Fayyad et al. (1996)

- **Sumarização**

Essa tarefa, muito comum em KDD, consiste em procurar identificar e indicar características comuns entre conjuntos de dados (WEISS; INDURKHYA, 1998).

Segundo (PASSOS, 2006), considere um banco de dados com informações sobre clientes que tem a assinatura de uma revista. A tarefa de sumarização deve buscar por características que sejam comuns a boa parte dos clientes. Por exemplo: são assinantes da revista X, homens na faixa etária de 25 a 45 anos, com nível superior e que trabalham na área de finanças. Essa informação poderia ser utilizada pela equipe de marketing da revista para direcionar a oferta para novos assinantes. É muito comum aplicar a tarefa de sumarização a cada um dos agrupamentos obtidos pela tarefa de clusterização. Lógica Indutiva e Algoritmos Genéticos são alguns exemplos de tecnologias que podem ser aplicadas na implementação da tarefa de sumarização.

- **Modelagem de dependência**

Segundo (Fayyad et al., 1996a), consiste em encontrar um modelo que descreve dependências significativas entre variáveis. Existem modelos de dependência em dois níveis:

o primeiro o de nível estrutural do modelo especifica (geralmente em forma gráfica) quais variáveis são localmente dependentes uma da outra e o segundo, o nível quantitativo do modelo especifica os pontos fortes das dependências usando alguma escala numérica. Por exemplo, as redes de dependência probabilística usam independência condicional para especificar o aspecto estrutural do modelo e as probabilidades ou correlações para especificar os pontos fortes das dependências. Redes de dependência probabilística estão cada vez mais procurando aplicações em áreas diversas como a área médica, em bases de dados de recuperação de informações e modelagem do genoma humano

- **Detecção de Desvios (Outliers)**

Essa tarefa consiste em procurar identificar registros do banco de dados cujas características não atendam aos padrões considerados normais no contexto (WEISS; INDURKHYA, 1998). Esses registros são chamados de *outliers*. Estas técnicas podem detectar mudanças ou padrões que não são esperados em todo o conjunto de dados ou em um subconjunto. Um exemplo seria de empresas de cartão de crédito a qual elas podem bloquear compras que não seguem ao perfil do cliente.

2.2.2 Decision Tree

A indução da árvore de decisão é o aprendizado das árvores de decisão do treinamento rotulado por classe tuplas. Uma árvore de decisão é uma estrutura de árvore semelhante a um fluxograma, onde cada nó interno (nó não folha) denota um teste em um atributo, cada ramificação representa um resultado do teste, e cada nó folha (ou nó terminal) contém um rótulo de classe. O nó superior em uma árvore é o nó raiz. (HAN; KAMBER, 2006).

Uma árvore de decisão típica é mostrada na figura abaixo representa o conceito compra um computador, ou seja, prevê se um cliente de uma loja provavelmente comprará um computador. Os nós internos são denotados por retângulos e os nós folha são denotados por ovais. Alguns algoritmos de árvore de decisão produzem apenas árvores binárias (em que cada nó interno se ramifica para exatamente dois outros nós), enquanto outros podem produzir árvores não binárias.

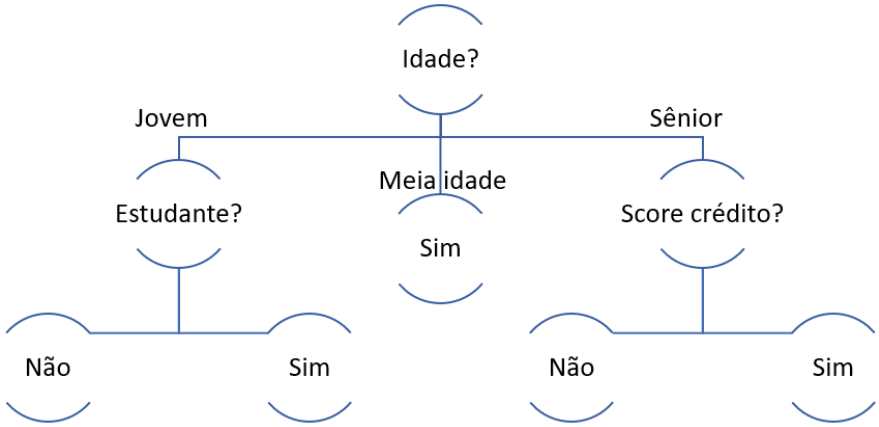


Figura 6: Árvore de decisão para venda de computador. Fonte: adaptação (HAN; KAMBER, 2006).

Como as árvores de decisão são usadas para classificação? Dada uma tupla, X, para a qual o rótulo de classe associado é desconhecido, os valores de atributo da tupla são testados em relação à árvore de decisão. Um caminho é traçado da raiz até um nó folha, que contém a previsão de classe para aquela tupla. As árvores de decisão podem ser facilmente convertidas em regras de classificação.

Por que os classificadores de árvore de decisão são tão populares? A construção de classificadores de árvore de decisão não requer nenhum conhecimento de domínio ou configuração de parâmetros e, portanto, é apropriada para a descoberta de conhecimento exploratório. As árvores de decisão podem lidar com dados multidimensionais. Sua representação do conhecimento adquirido na forma de árvore é intuitiva e geralmente fácil de assimilar pelos humanos. As etapas de aprendizagem e classificação da indução da árvore de decisão são simples e rápidas. Em geral, os classificadores da árvore de decisão têm boa precisão. No entanto, o uso bem-sucedido pode depender dos dados disponíveis. Os algoritmos de indução de árvore de decisão têm sido usados para classificação em muitas áreas de aplicação, como medicina, manufatura e produção, análise financeira, astronomia e biologia molecular. As árvores de decisão são a base de vários sistemas de indução de regras comerciais.

2.2.3 Random Forest

Um conjunto de *Random Forest* usa muitas árvores de decisão individuais não ajustadas que são criadas aleatoriamente pela divisão em cada nó da árvore de decisão [Breiman (2001)]. Cada árvore provavelmente será menos precisa do que uma árvore criada

com as divisões exatas. Mas, ao combinar várias dessas árvores “aproximadas” em um conjunto, podemos melhorar a precisão, geralmente fazendo melhor do que uma única árvore com divisões exatas. (ROKACH; MAIMON, 2005).

Imagine que cada um dos classificadores no conjunto seja um classificador de árvore de decisão de forma que a coleção de classificadores seja uma "floresta". As árvores de decisão individuais são geradas usando uma seleção aleatória de atributos em cada nó para determinar a divisão. Mais formalmente, cada árvore depende dos valores de um vetor aleatório amostrado de forma independente e com a mesma distribuição para todas as árvores da floresta. Durante a classificação, cada árvore vota e a classe mais popular é retornada. (HAN; KAMBER, 2006).

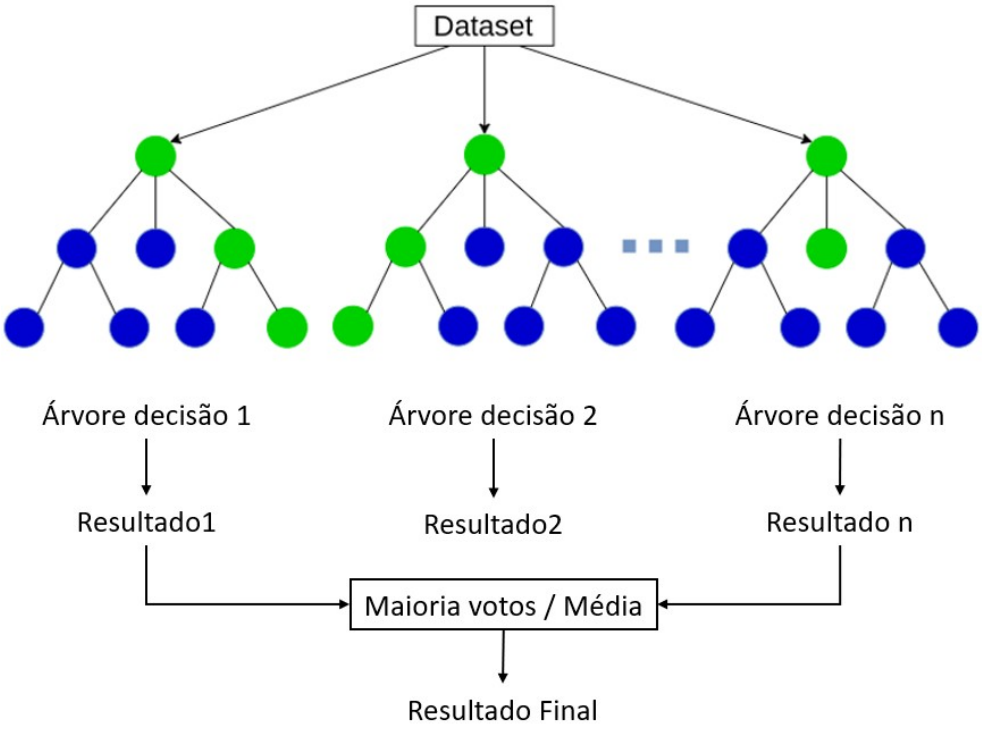


Figura 7: Árvores de decisão formando uma Random Forest. Fonte: Elaboração própria.

3 DEMONSTRAÇÃO DO PROCESSO DE KDD

Neste capítulo é realizado a utilização de algoritmos de aprendizado de máquina sobre a base de dados de empréstimos bancários descrita a seguir. Foi utilizado o modelo de etapas de KDD conforme definido por Fayyad et al no segundo capítulo para demonstrar o funcionamento de todo o processo, passo a passo, com objetivo de exemplificar a aplicabilidade do algoritmo de Random Forest e Decision Tree para obtenção de conhecimento.

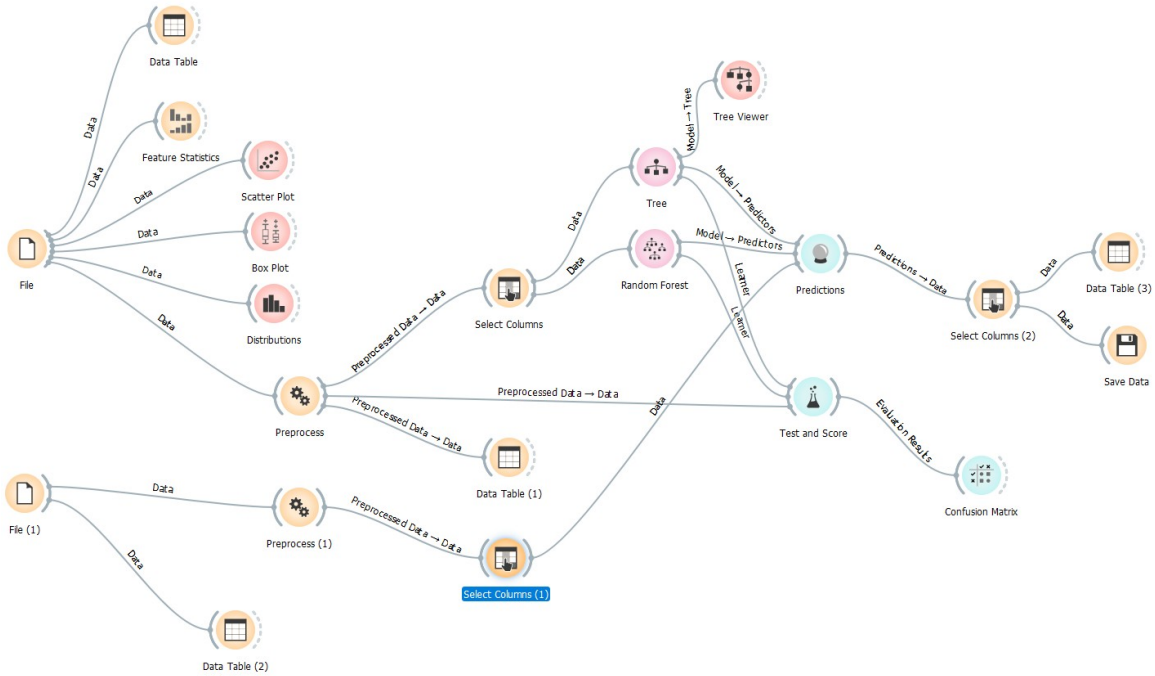


Figura 8: Todas as etapas do processo de KDD Mapeadas no Orange Data Mining. Fonte: Elaboração própria.

3.1 Orange Data Mining

Neste processo, é utilizado uma ferramenta open source de mineração de dados chamada de Orange Data Mining, ela é uma ferramenta que permite que seja criado todo o processo de mineração de dados dentro dela sem a necessidade de programar em alguma linguagem, porém os conhecimentos de estatística, negócio dentre outros citados continuam extrema importância para o sucesso da aplicação e que o processo seja bem sucedido.

Funcionalidades

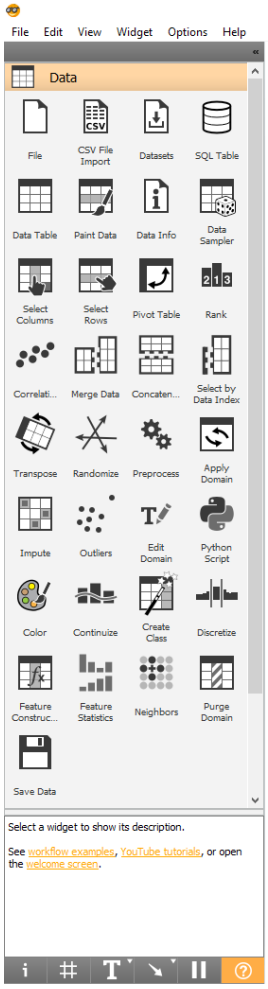


Figura 9: Funcionalidades do software Orange Data Mining. Fonte: Elaboração própria.

Esta é a tela principal, a qual ao lado esquerdo ficam todas as funcionalidades disponíveis na versão atual deste trabalho, a versão 3.26.0, e na parte direita ficam todas as etapas do processo que foi ou será criado. Na aba esquerda temos 5 categorias de funcionalidades, que são:

- ***Data***

  Cada widget representa uma ferramenta ou técnica para ser utilizado nos processos de carga e preparação dos dados.

- ***Visualize***

  Aqui se tem todas as ferramentas necessárias para uma análise exploratória de dados como gráficos de dispersão, gráficos de linha, dentre outros.

- ***Model***

  Nesta aba se tem os modelos de machine learning supervisionados.

- ***Evaluate***

  Nesta aba tem técnicas para avaliar a acurácia dos modelos.

- ***Unsupervised***

  Nesta aba tem se os modelos de machine learning não supervisionados.

3.2 Base de dados

A base de dados (*Predict Loan Eligibility for Dream Housing Finance company*) utilizada neste processo está em formato CSV, que é um formato amplamente utilizado para este tipo de aplicações. O arquivo contém as seguintes colunas e suas respectivas descrições:

Tabela 2: Informações das colunas da base de dados e a descrição de cada coluna. Fonte: Elaboração própria.

variáveis	Descrição
Loan_ID	Valor único ID
Gender	Gênero
Married	Se é casado (Y para Sim e N para não)
Dependents	Número de dependentes
Education	Graduação (se é ou não graduado)
Self_Employed	Se é autônomo
ApplicantIncome	Renda do requerente
CoapplicantIncome	Renda do co-requerente
LoanAmount	Montante do empréstimo em milhares
Loan_Amount_Term	Prazo do empréstimo em meses
Credit_History	Histórico de crédito atende às diretrizes
Property_Area	Local da Propriedade
Loan_Status	Variável alvo – Situação do empréstimo

Conforme é visto na tabela acima, já existe uma coluna target para se basear todo o processo, além disso, observa se que só é possível dois valores, sim ou não para empréstimo concedido. Com isso, é necessário utilizar-se de algoritmos que entrem nesta categoria de solução de problemas.

3.3 Ambiente de Trabalho

Para a execução dos softwares utilizados neste trabalho, foi utilizado um *Desktop* com processador AMD Ryzen 5 1600AF, 16Gb de memória RAM, SSD de 480Gb, placa de vídeo AMD RX 5600XT e o sistema operacional Windows em sua versão 10. O Software utilizado foi o Orange Data Mining em sua versão 3.26.0.

3.4 Seleção de dados

Como só é utilizado uma tabela de dados, não terá necessidade de aplicar técnicas de redução de dados ou mesmo agrupamento de bases de dados. Caso fossem dados oriundos de diversas localidades como banco de dados, internet, arquivos em excel, seria nesta etapa a qual seriam escolhidos os dados para serem fornecidos aos algoritmos de mineração de dados.

Carregando os dados no Orange, já se pode ver que o arquivo tem 614 linhas e 11 colunas (aqui chamadas *features*), a qual 2% delas tem valores *missing* (valores faltantes). Aqui também é possível há mudar o formato de dados, caso haja necessidade.

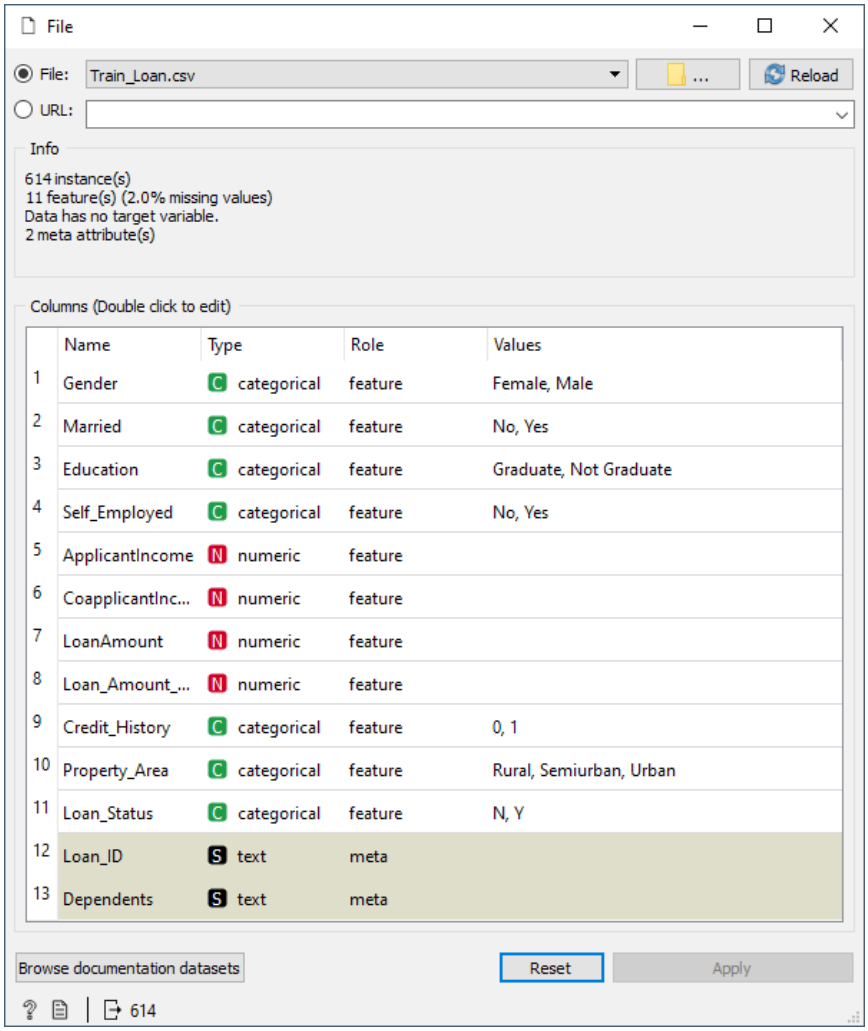
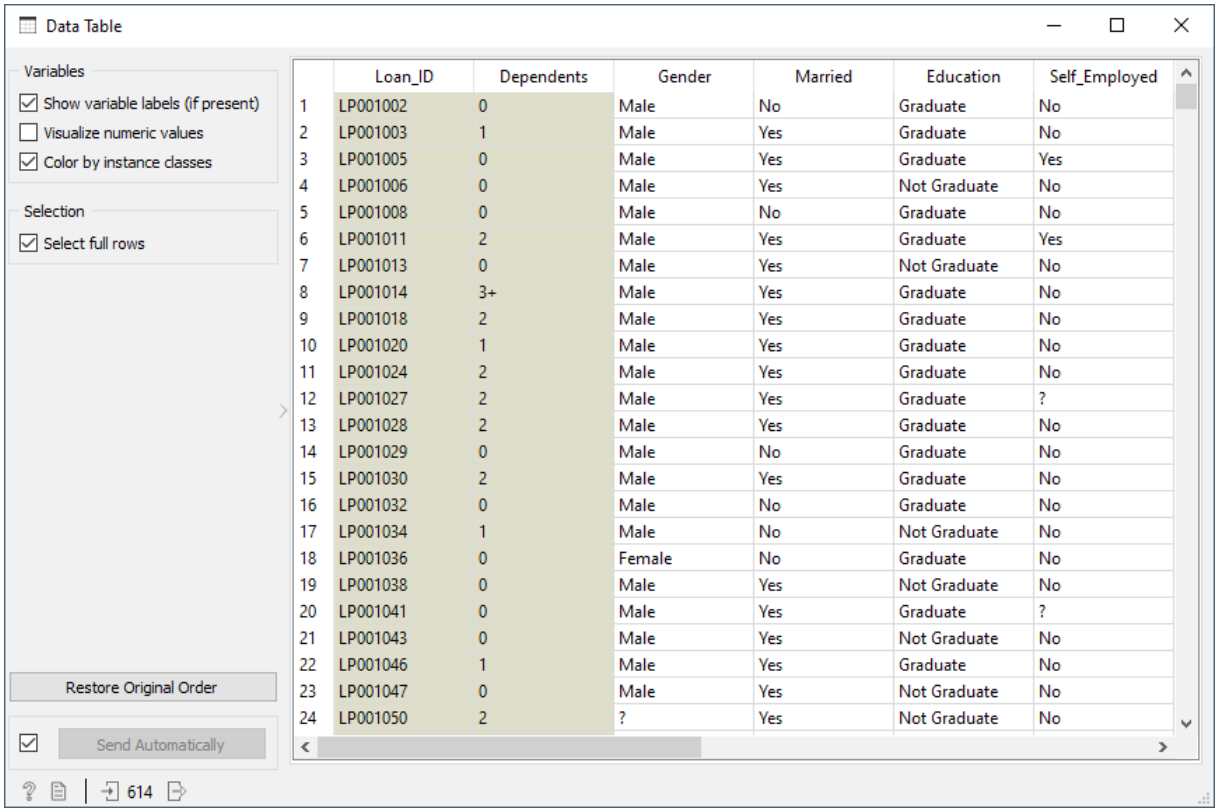


Figura 10: Tipos dos dados dataset. Fonte: Elaboração própria.

Na **figura 11** tem se a visualização conforme vista em banco de dados e planilhas, onde se pode ter uma ideia dos tipos de dados que estão em cada coluna.



	Loan_ID	Dependents	Gender	Married	Education	Self_Employed
1	LP001002	0	Male	No	Graduate	No
2	LP001003	1	Male	Yes	Graduate	No
3	LP001005	0	Male	Yes	Graduate	Yes
4	LP001006	0	Male	Yes	Not Graduate	No
5	LP001008	0	Male	No	Graduate	No
6	LP001011	2	Male	Yes	Graduate	Yes
7	LP001013	0	Male	Yes	Not Graduate	No
8	LP001014	3+	Male	Yes	Graduate	No
9	LP001018	2	Male	Yes	Graduate	No
10	LP001020	1	Male	Yes	Graduate	No
11	LP001024	2	Male	Yes	Graduate	No
12	LP001027	2	Male	Yes	Graduate	?
13	LP001028	2	Male	Yes	Graduate	No
14	LP001029	0	Male	No	Graduate	No
15	LP001030	2	Male	Yes	Graduate	No
16	LP001032	0	Male	No	Graduate	No
17	LP001034	1	Male	No	Not Graduate	No
18	LP001036	0	Female	No	Graduate	No
19	LP001038	0	Male	Yes	Not Graduate	No
20	LP001041	0	Male	Yes	Graduate	?
21	LP001043	0	Male	Yes	Not Graduate	No
22	LP001046	1	Male	Yes	Graduate	No
23	LP001047	0	Male	Yes	Not Graduate	No
24	LP001050	2	?	Yes	Not Graduate	No

Figura 11: Visualização dos dados contidos no dataset. Fonte: Elaboração própria.

Visualizando os dados de uma forma mais estatística na **figura 12**, pode se perceber a média, a dispersão, os valores máximo e mínimo e a moda das variáveis categóricas. Também é possível verificar o percentual de valores nulos em cada coluna/variável.

Analisando os dados conforme descrito na figura abaixo, percebe se que por exemplo crédito é negado (em azul) em grande parte das pessoas que não possuem histórico de crédito, que a média de empréstimos solicitados é de 146 mil dólares aproximadamente e que a média salarial é 5400 dólares, são geralmente homens, casados e com graduação.

Figura 12: Informações estatísticas dos dados. Fonte: Elaboração própria.

A **figura 13** mostra que a renda média é de 5412 dólares enquanto a mediana é de 3813 dólares, alertando que essa variável tem uma distribuição calda acentuada para a direita. A média provavelmente está sendo distorcida pela presença de outliers, no caso, há um solicitante com renda de mais de 80 mil dólares.

Figura 13: Média e mediana da renda dos requerentes. Fonte: Elaboração própria.

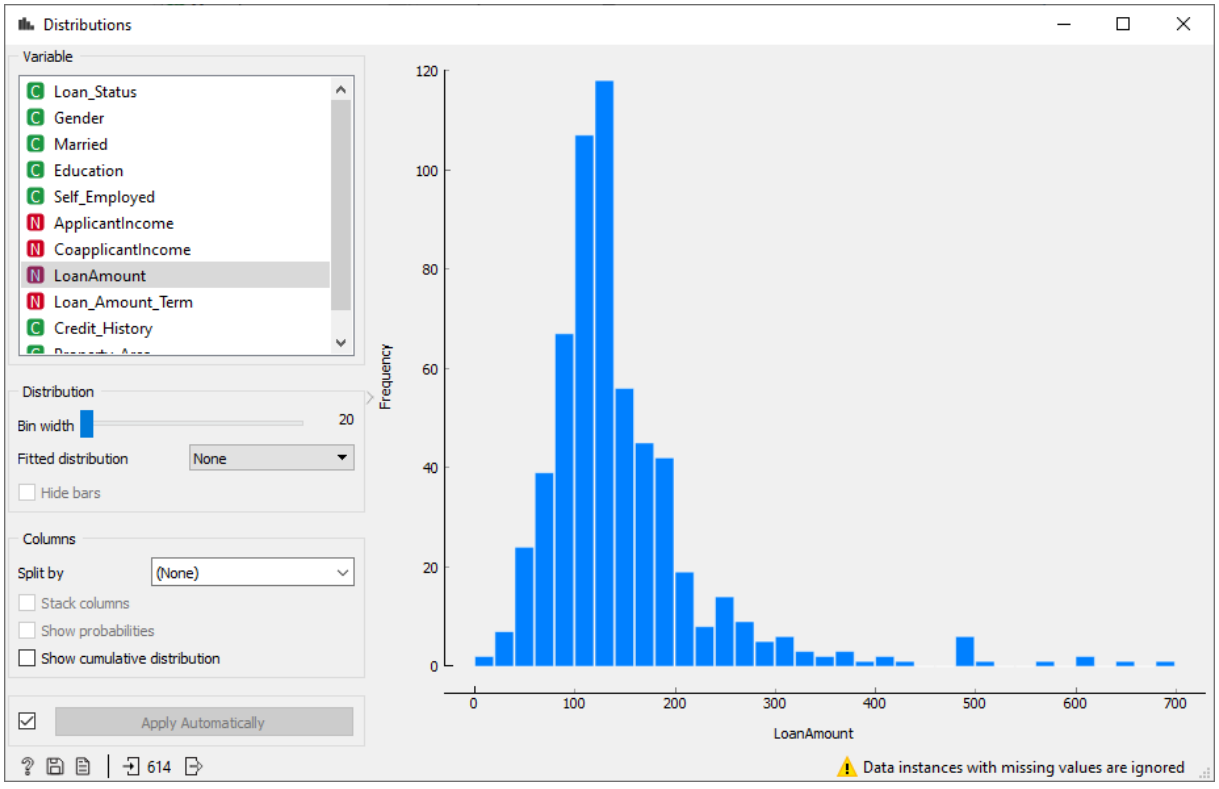


Figura 14: Histograma com a variável montante de empréstimo. Fonte: Elaboração própria.

Na **figura 15** Percebe se que a quantidade de empréstimo solicitada pelos homens é maior que a das mulheres, enquanto a média da solicitação de homens é de aproximadamente 130 mil dólares e a de mulheres fica em torno de 113 mil dólares.

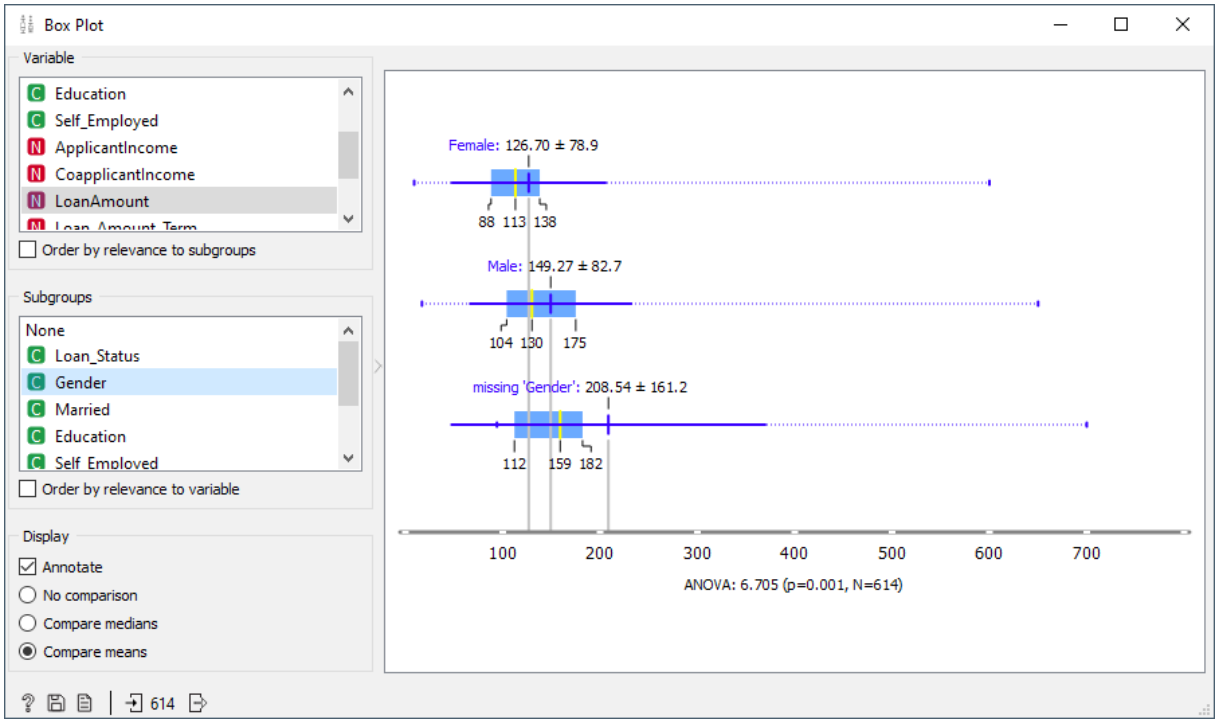


Figura 15: Boxplot com a média de solicitação empréstimo por homens e mulheres. Fonte: Elaboração própria.

Na **figura 16** percebe se que a quantidade de empréstimo solicitado está fortemente ligada a renda do solicitante, indicando que quanto maior a renda, maior o empréstimo solicitado.

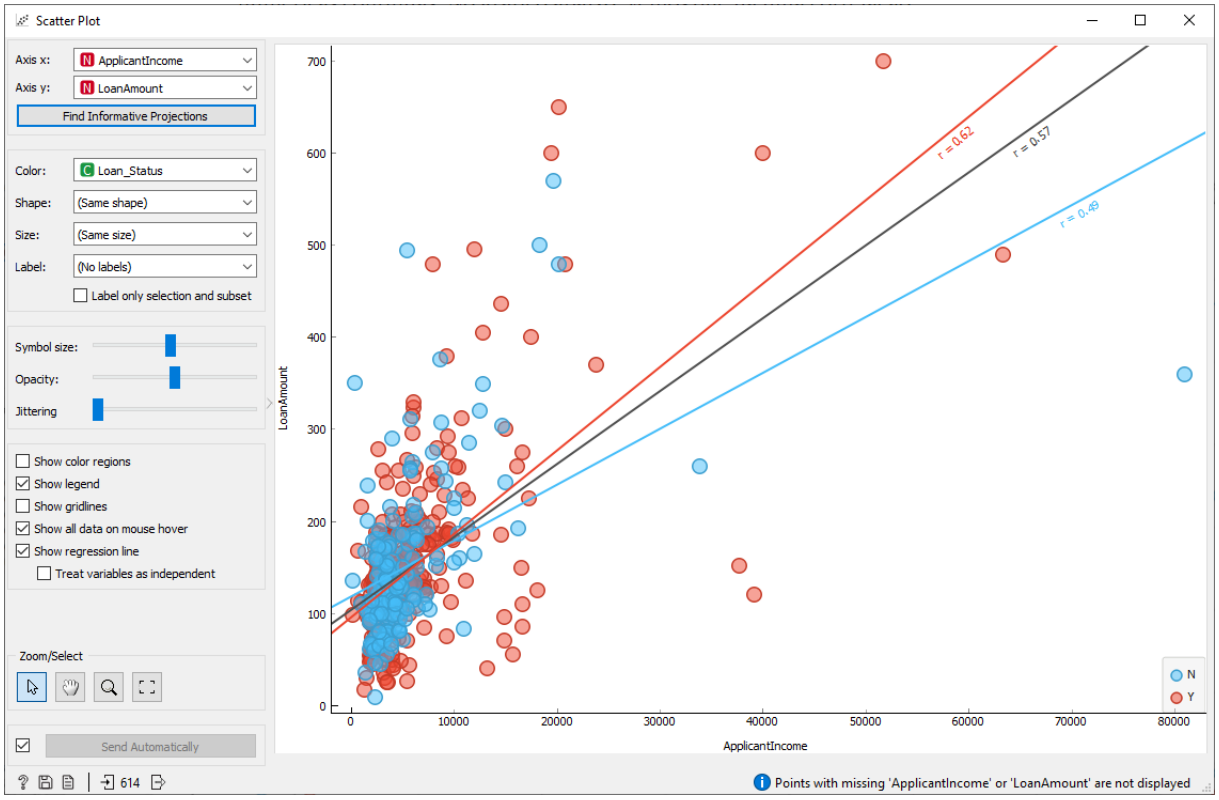


Figura 16: Gráfico de dispersão com a linha de regressão entre a renda e o montante de empréstimo solicitado pelo requerente. Fonte: Elaboração própria.

A **figura 17** mostra a quantidade de reprovações de carta de crédito por localidade. Percebe se que as maiores solicitações bem como as aprovações são as de áreas semiurbanas.

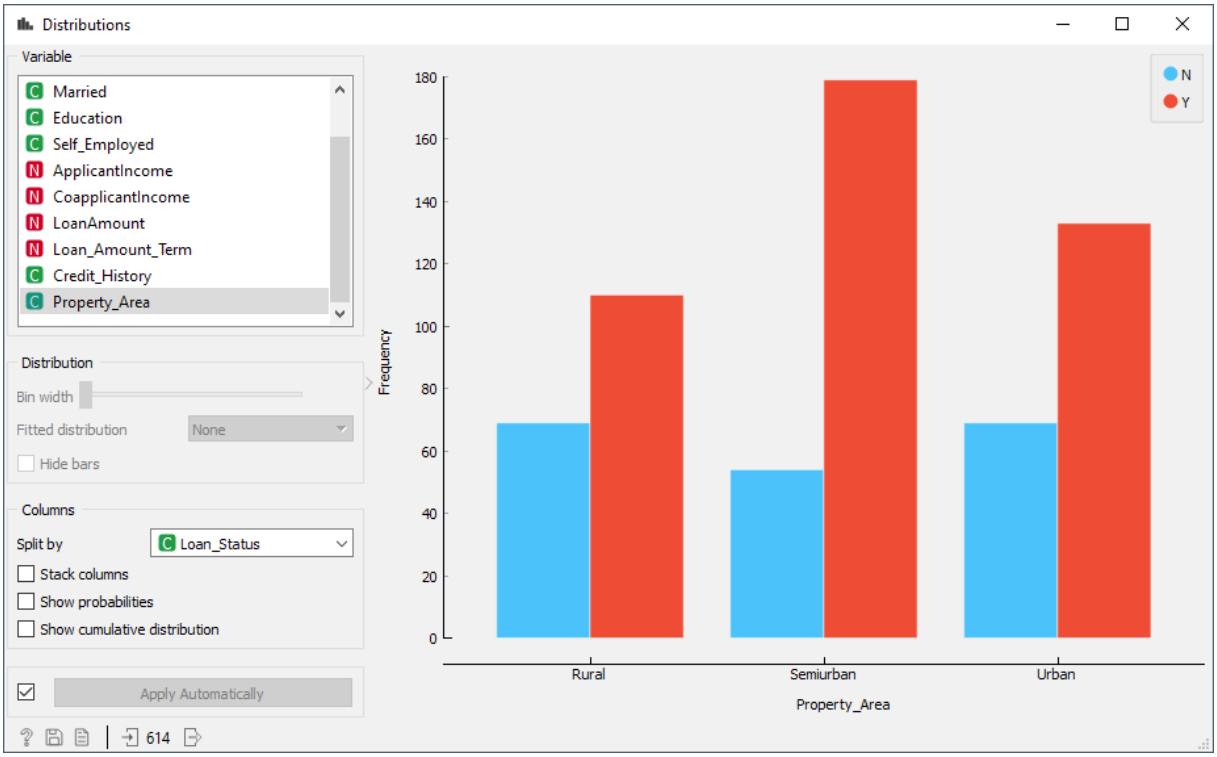


Figura 17: Distribuição das aprovações de crédito por localidade. Fonte: Elaboração própria.

3.5 Pré-Processamento

Nesta etapa, são tratados os dados faltantes que podem ocasionar erros na aplicação dos algoritmos e se pode fazer isso de diversas formas, as mais comuns para dados numéricos são: excluir a coluna por inteiro, acrescentar a média dos valores nos dados faltantes ou além de acrescentar essa média, criar uma coluna booleana que diga se aquela variável estava faltando ou não. Para dados categóricos, as definições viram números dentro de uma nova coluna ou pode se criar uma coluna para cada valor desta coluna original, geralmente chamados de cardinalização.

Nesta aplicação, será utilizado a moda para valores categóricos faltantes e a média dos valores para valores numéricos. Como se tem outliers bastante discrepantes, optou se por utilizar a mediana para centralizar o conjunto de dados. Também foi transformado as variáveis categóricas em valores numéricos de acordo com sua cardinalidade, neste caso, como só tem basicamente dois valores para cada colunas, pode se dizer que são binárias.

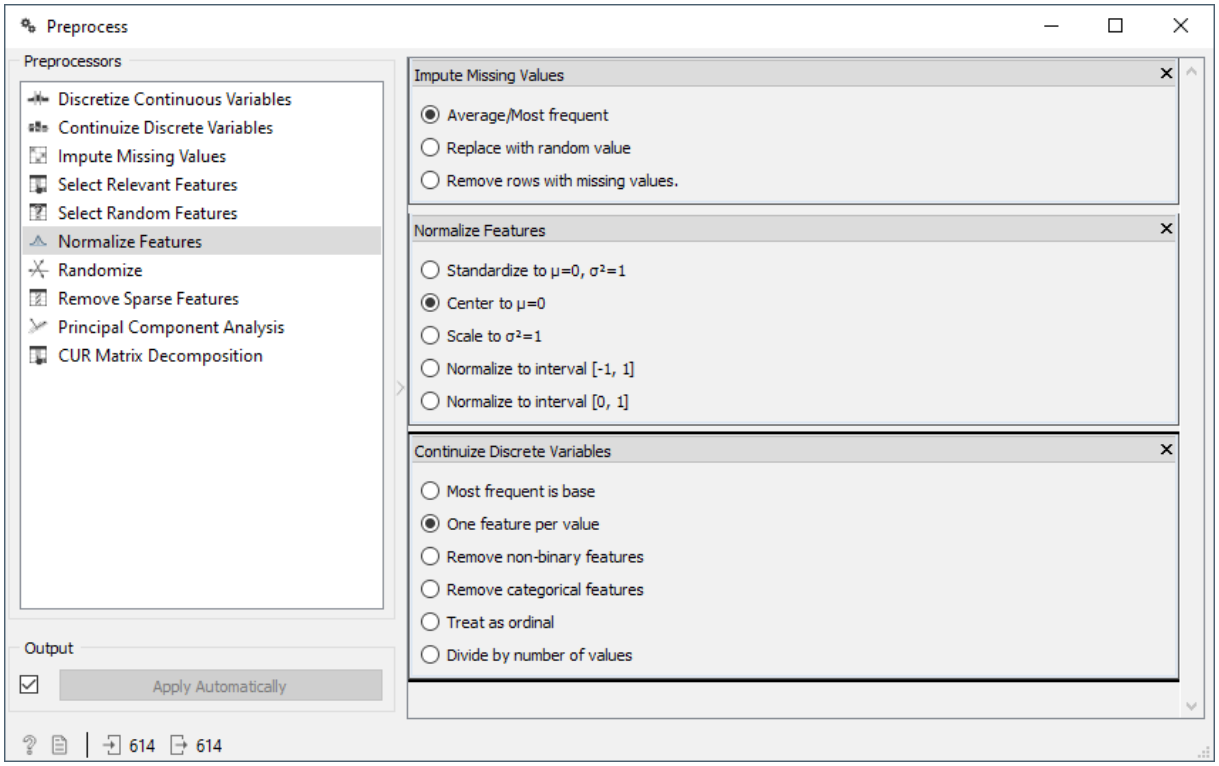
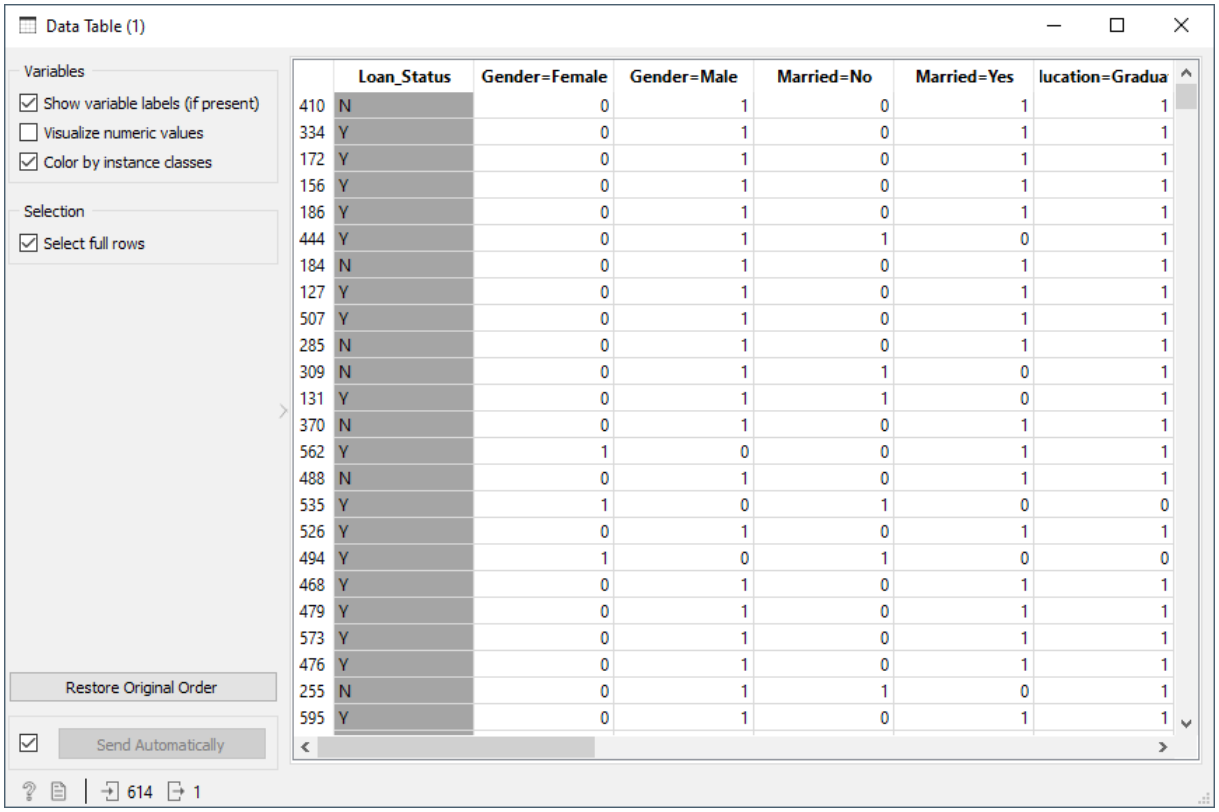


Figura 18: Pré-processamento dos dados. Fonte: Elaboração própria.

Aqui se verifica como estão os dados após a etapa de pré-processamento, como citado acima, a coluna que define o gênero, foi dividida em duas. Isso vale para outras colunas categóricas.



	Loan_Status	Gender=Female	Gender=Male	Married=No	Married=Yes	lucation=Gradua
410	N	0	1	0	1	1
334	Y	0	1	0	1	1
172	Y	0	1	0	1	1
156	Y	0	1	0	1	1
186	Y	0	1	0	1	1
444	Y	0	1	1	0	1
184	N	0	1	0	1	1
127	Y	0	1	0	1	1
507	Y	0	1	0	1	1
285	N	0	1	0	1	1
309	N	0	1	1	0	1
131	Y	0	1	1	0	1
370	N	0	1	0	1	1
562	Y	1	0	0	1	1
488	N	0	1	0	1	1
535	Y	1	0	1	0	0
526	Y	0	1	0	1	1
494	Y	1	0	1	0	0
468	Y	0	1	0	1	1
479	Y	0	1	0	1	1
573	Y	0	1	0	1	1
476	Y	0	1	0	1	1
255	N	0	1	1	0	1
595	Y	0	1	0	1	1

Figura 19: Visualizando os dados após pré-processamento. Fonte: Elaboração própria.

3.6 Formatação

Como algumas colunas foram divididas e transformadas em valores binários, não há necessidade de levá-las para o algoritmo, pois poderia até interferir na acurácia do modelo como mostrado nos exemplos acima. A coluna gênero foi dividida em Masculino com valores 0 para não e 1 para sim e o mesmo com feminino, elas contêm basicamente o mesmo tipo de informação e não é necessário continuar com elas para o algoritmo. Isso também aconteceu com as colunas *Education, Credit_History e Married*. Portanto, elas foram removidas do dataset que será usado pelos algoritmos.

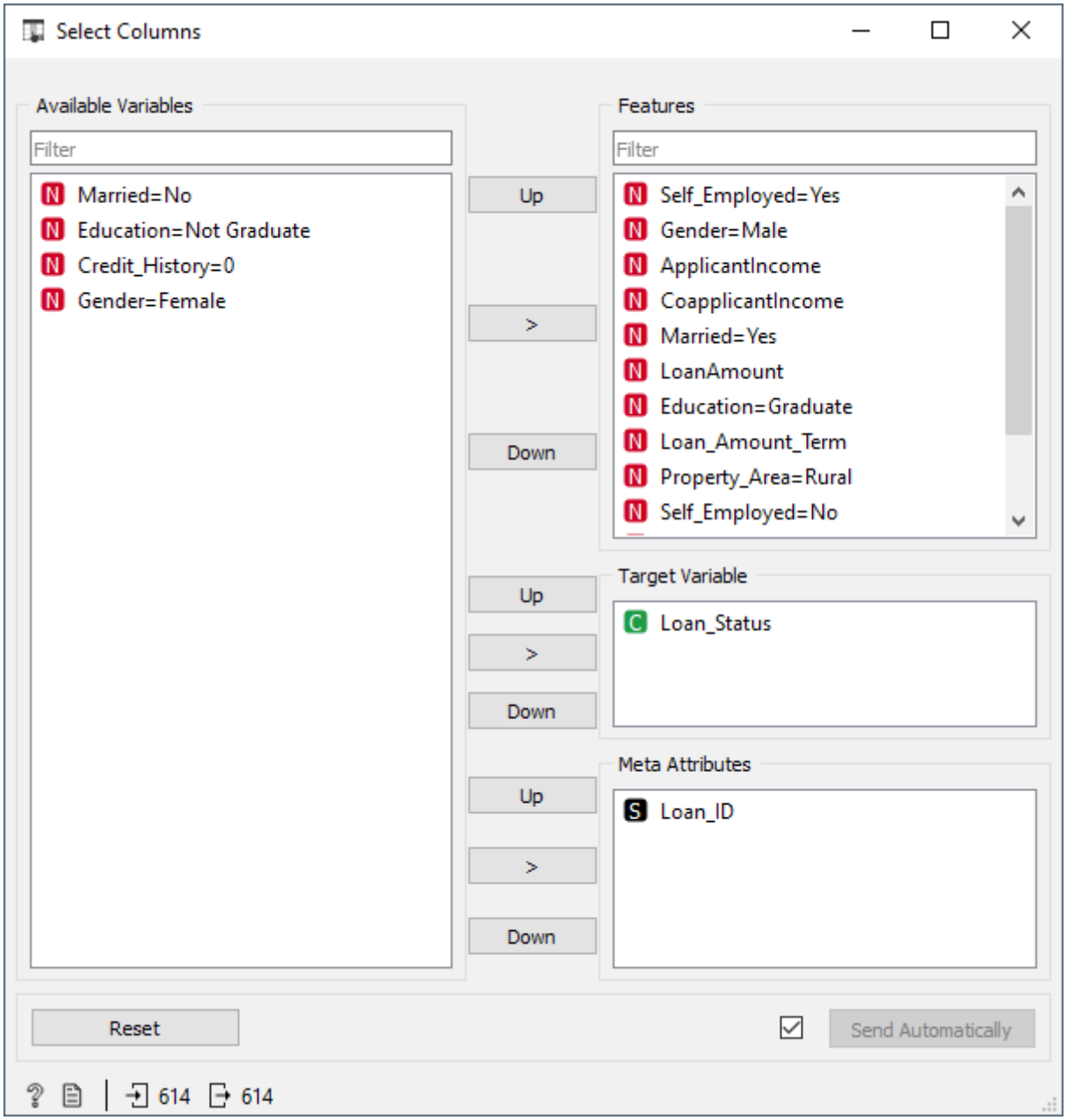


Figura 20: Seleção de colunas que não tem necessidade de uso após a etapa de pré-processamento.

Fonte: Elaboração própria.

3.7 Mineração

Aqui tem se a etapa mais crucial de todo processo, logo foram utilizados dois tipos de algoritmos que se enquadram nos modelos de classificação citados no referencial teórico. Serão avaliados dois modelos, o *Random Forest* e o *Decision Tree.*

Random Forest

A *Random Forest* constrói um conjunto de árvores de decisão. Cada árvore é desenvolvida a partir da reamostragem dos dados de treinamento. Ao desenvolver árvores individuais, um subconjunto arbitrário de atributos é desenhado (por isso o termo “Random”), a partir do qual o melhor atributo para a divisão é selecionado. O modelo final é baseado na maioria dos votos de árvores desenvolvidas individualmente na floresta.

Na aplicação tem se os seguintes parâmetros a serem definidos:

- Nome do Modelo;
- Quantas árvores terão no modelo;
- Quantos atributos para cada nó (caso esteja desmarcada, será igual a raiz quadrada do número de atributos nos dados);
- Reprodução dos resultados (há a possibilidade de corrigir a semente para geração de árvore, semente fixa para gerador aleatório, o que permite a reprodução dos resultados);
- Limite de profundidade individual de cada árvore;
- Limite para o menor subconjunto para que possa haver uma divisão.

Com as informações acima, foi criada a *Random Forest* da seguinte forma seguindo algumas características padrões e outras definidas pelo autor deste trabalho.

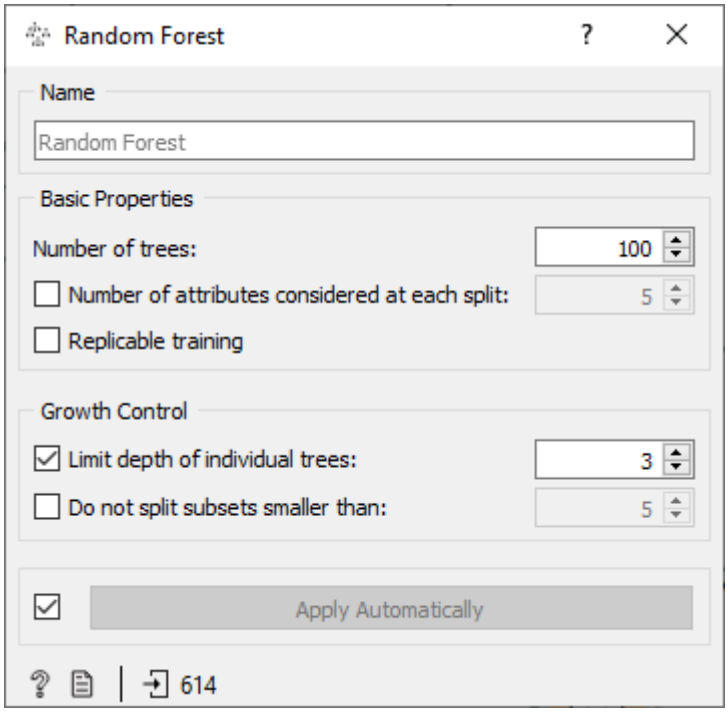


Figura 21: Parâmetros do algoritmo de Random Forest. Fonte: Elaboração própria.

Decision Tree

Decision Tree é um algoritmo simples que divide os dados em nós por pureza de classe. É um precursor da Random Forest.

Na aplicação tem se os seguintes parâmetros a serem definidos:

- Construir uma árvore binária (dividida em dois nós filhos);
- Número mínimo de instância nas folhas;
- Não dividir os subconjuntos menores que;
- Limitar a profundidade máxima da arvore.

Com as informações acima, foi criada a Decision tree da seguinte forma seguindo algumas características padrões e outras definidas pelo autor deste trabalho.

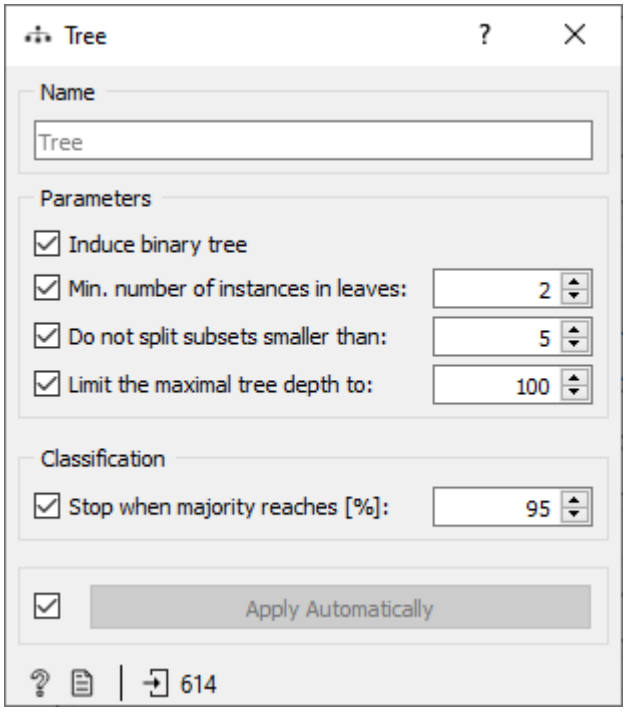


Figura 22: Parâmetros do algoritmo de Decision Tree. Fonte: Elaboração própria.

3.8 Avaliação

Ao avaliar o modelo, se obteve as seguintes pontuações na matriz de confusão. No *Random Forest*, ele acertou 92,1% dos dados que realmente não era para liberar crédito e 79% dos que era para liberar o crédito quanto ao algoritmo de *Decision Tree* obteve, 49,6% e 79,5% respectivamente.

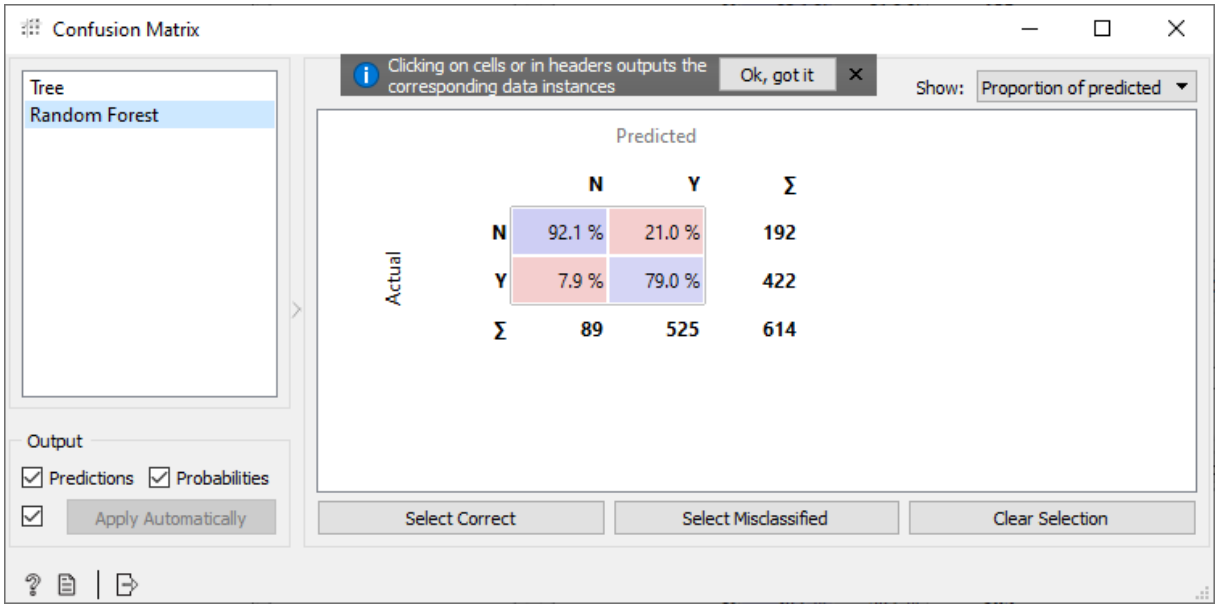


Figura 23: Matriz de confusão do algoritmo de Random Forest. Fonte: Elaboração própria.

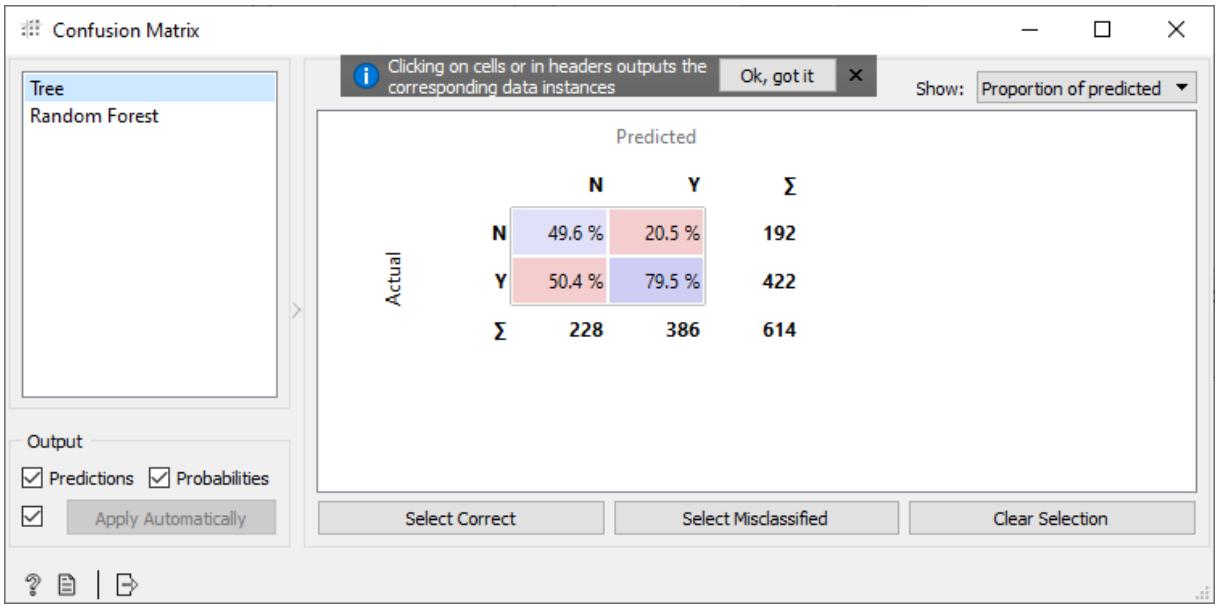


Figura 24: Matriz de confusão do algoritmo de Decision Tree. Fonte: Elaboração própria.

Agora verificando *widget* de *test and score*, que mostra outros dados temos as seguintes informações. Através do *Cross Validation* que consiste em retirar uma pequena parcela do *dataset* para validação e a cada iteração outra parte do *dataset* é utilizada para esta mesma validação, o algoritmo *Random Forest* obteve 80,9% de acurácia e 83,1% de precisão enquanto o *Decision Tree* obteve 68,4% de acurácia e 70,2% de precisão.

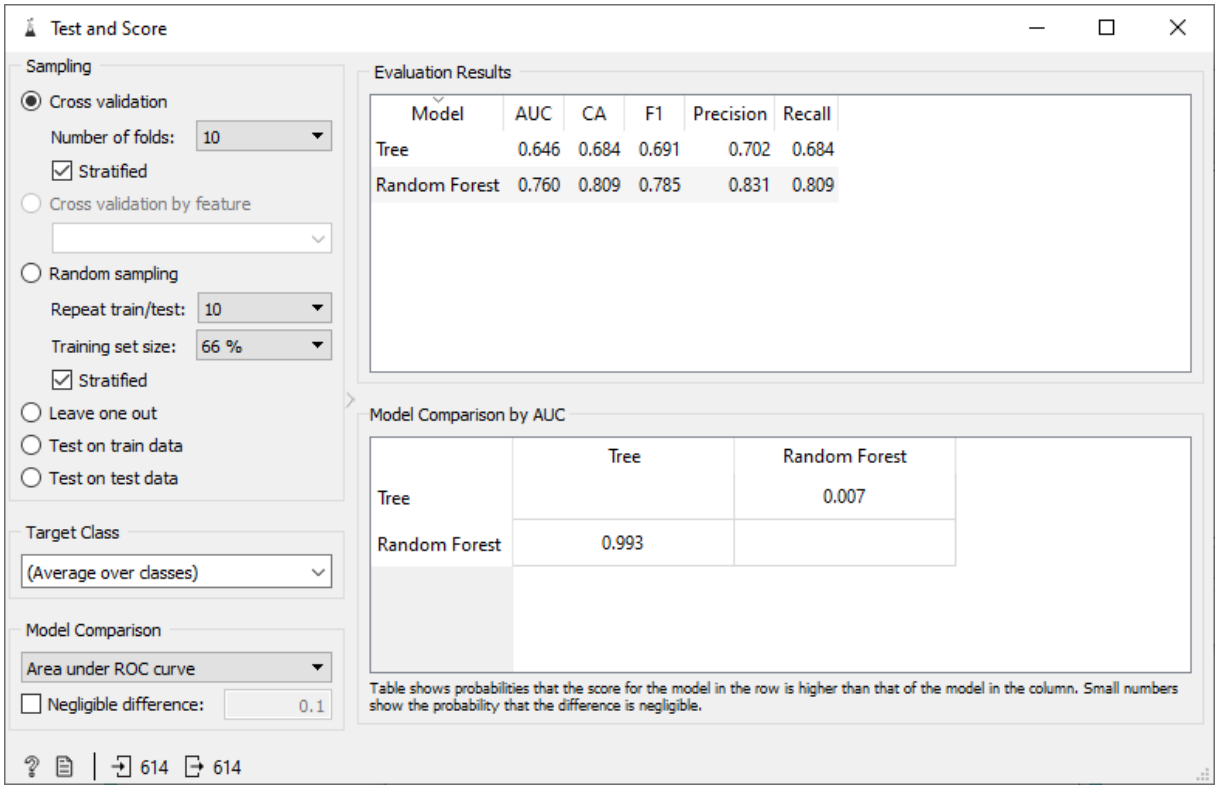


Figura 25: Testes utilizando validação cruzada. Fonte: Elaboração própria.

Ao validar os modelos nesta primeira iteração, percebe se que os dois modelos atingiram um nível satisfatório de precisão e acurácia. Caso isso fosse uma aplicação para uma empresa ou o projeto em si que precisa de uma precisão maior, outras formas deveriam ser avaliadas, como verificar quais dados no dataset poderiam ser limpos, pois não tinham utilidade para o algoritmo, mudar os parâmetros do próprio algoritmo, verificar se há algum tipo de vazamento nos dados, procurar por outros modelos de Machine Learning que fossem mais adequados a classificação dentre outras opções.

4 CONSIDERAÇÕES FINAIS

O processo de KDD definido no capítulo 2 mostrou-se eficiente, pois foi possível gerar um conhecimento sobre dados além de conseguir classificá-los. Suas etapas foram seguidas e o resultado gerado foi capaz de responder à pergunta levantada para o experimento.

Foi necessário realizar uma análise nos dados para identificar o melhor modelo a ser aplicado neste trabalho. Os algoritmos de *Random Forest* e *Decision Tree* em si não possuem alta complexidade computacional (são bastante utilizados e populares). Observou se que quanto maiores as folhas ou as árvores o processamento se tornava cada vez mais lento, porém aumentar deliberadamente essas funcionalidades não significa que o modelo se torne cada vez mais preciso, mas sim pode ocorrer o contrário, ele apenas gravar os dados e não conseguir replicar a novas entradas, o que na ciência de dados se chama de *overfiting*.

O experimento realizado neste trabalho representa apenas uma das enormes possibilidades de se seguir na área de Mineração de Dados. Seguindo a estruturação proposta é possível aplicar esse experimento em infinitos contextos, visto que a cada dia novos dados são produzidos, uma vez que novos sistemas de informação surgem todos os dias.

Observa-se que cada vez mais investimentos serão aplicados na área de MD pois a informação gerada pelo processo tem um enorme potencial econômico, se bem aplicadas (como no exemplo prático deste trabalho).

Este trabalho foi realizado em um software que embora não muito conhecido, mas tem uma capacidade enorme de realizar todas as etapas de mineração de dados, com grandes quantidades de dados ou não, dados em tempo real ou não. Com a prática deste trabalho foi possível aplicar apenas uma pequena capacidade desta ferramenta que apesar de não exigir conhecimentos de programação em si, exige bastante conhecimentos de estatística, de *Machine Learning* dentre outras habilidades.

REFERÊNCIAS BIBLIOGRÁFICAS

ANGELONI, M. T. **Elementos intervenientes na tomada de decisão**. Ci. Inf., Brasília, janeiro/abril 2003. 17-22. Acesso em: 03 de junho de 2020 Disponível em: < https://www.researchgate.net/publication/26349980_Elementos_intervenientes_na_ tomada_de_decisao >.

BATISTA, Gustavo Enrique A. P. A. B. **Pré-Processamento de Dados em Aprendizado de Máquina Supervisionado.** Tese de Doutorado (Doutorado em Ciência da Computação e Matemática Computação). USP – São Carlos, 2003. Acesso em: 05 de junho de 2020. Disponível em:<http://www.teses.usp.br/teses/disponiveis/55/55134/tde-06102003-160219/publico/TeseDoutorado.pdf>.

BERRY, M.J.A.; LINOFF, G. **Data Mining Techniques**. New York: John Wiley & Sons, Inc. 1997.

CABENA, P., HADJINIAN, P., STADLER, R., VERHEES, J., e ZANASI, **A. Discovering data mining: from concept to implementation.** Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998.

CHANG, H. C. & HSU, C.C. **Using Topic Keyword Clusters for automatic Document Clustering.** Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05), Kota Kinabalu, Sabah, 2005.

CHAPMAN, P. et al, 2000. **CRISP-DM 1.0 - Step-by-step data mining guide.**

DALLANORA, Julio Fernando. **Análise do Perfil de Clientes a partir de Técnicas de Mineração de Dados.** UNIJUI – Universidade Regional do Noroeste do Estado do Rio Grande do Sul, Ijuí, 2009.

DUNHAM, M. H, 2002. **Data Mining:** Introductory And Advanced Topics Pearson Education

FAYYAD, Usama; Piatetsky-Shapiro, G; Smyth, P. ***From Data mining* to *Knowledge Discovery in Databases*.** American Association for Artificial Intelligence, 1996.

GIL, C. A. **Como Elaborar Projetos de Pesquisa**, 2002.

HAN, J; KAMBER, M. **Data Mining: Concepts and Techniques**. Elsevier, 2006.

HAN, Jiawei; KAMBER, Micheline. **Data Mining: Concepts and Techniques**. San Diego: Academic Press, 2001

HOSKING, J.R.M., PEDNAULT, SUDAN, M. **A statistical perspective on data mining**. New York, NY, EUA,1997.

JAIN, Anil e DUBES, Richard. **Algorithms for clustering data.** Prentice-Hall, Inc. Upper Saddle River, NJ, USA ©1988

MCCUE, C. **Data Mining and Predictive Analysis** - Intelligence Gathering and Crime Analysis. Elsevier, 2007.

MICHIE, D.; SPIEGELHALTER, D.; TAYLOR, C**. Machine Learning, Neural and Statistical Classifications**. Ellis Horwood, 1994.

PASSOS, M. **Modelos de dispositivos de microondas e Ópticos através de redes neurais artificiais**. Natal,2006.

PRASS, F. S. KDD – **Uma Visão Geral do Processo**, 2007. Acesso em: 05 de junho de 2020 em: < http://fp2.com.br/blog/wpcontent/uploads/2012/07/kdd_uma_visao_geral_do_processo.pdf>.

REZENDE, Solange O. **Mineração de Dados.** XXV Congresso da Sociedade Brasileira de Computação. UNISINOS – São Leopoldo, RS, 2005. Acesso em 02 de junho de 2020. Disponível em:<http://www.lbd.dcc.ufmg.br/colecoes/enia/2005/0102.pdf>.

REZENDE, Solange O. **Sistemas Inteligentes: Fundamentos e Aplicações.** Barueri, SP: Ed. Manole, 2003.

ROKACH, L.; MAIMON, O. **Top-down induction of decision trees classifiers** - A survey. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, v. 35, n. 4, p. 476–487, 2005. ISSN 10946977.

RONALDO GOLDSCHMIDT, EMMANUEL PASSOS. **Data mining: Um guia prático**, 2003.

SILVA, Denilson R. **Análise e Triagem de Padrões em Processamento de Descobertade Conhecimento em Base de Dados.** Dissertação de Mestrado (Mestrado em Ciência da Computação) - Pontifícia Universidade Católica do Rio Grande do Sul – RS, 2000. Acessado em 05/06/2020. Disponível em: <http://www.pucrs.br/inf/pos/dissertacoes/arquivos/denilson.pdf>.

THIOLLENT, Michel. **Metodologia da pesquisa-ação**,1985.

WEISS, S.; INDURKHYA, N. **Predictive Data Mining: a practical guide**. San Francisco: Morgan Kaufmann Publishers, 1998