UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Multilevel Logistic Regression Model for Client Churn Classification: A Generalized Linear Mixed Models Approach

Eduardo Soares Zanutti

Projeto de Pesquisa apresentado como parte dos requisitos para ingresso no Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC) modalidade Doutorado Direto

ICMC USP
SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO (USP/SC)
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO (ICMC)
Programa de Pós-Graduação em Ciência Computacional e Matemática Computacional (PPG-CCMC)

Eduardo Soares Zanutti

MODELO DE REGRESSÃO LOGÍSTICA MULTINÍVEL PARA A CLASSIFICAÇÃO DA ROTATIVIDADE DOS CLIENTES: UMA ABORDAGEM DE MODELOS LINEARES MISTOS GENERALIZADOS

Projeto de Pesquisa apresentado como parte dos requisitos para ingresso no Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC) modalidade Doutorado Direto

USP - SÃO CARLOS
Novembro 2021

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

BN – *Bayesian Network*

CART – *Classification And Regression Trees*

CRM – *Customer Relationship Management*

CRISP-DM – *Cross Industry Standard Process for Data Mining*

GLM – *Generalized Linear Models*

GLMM – *Generalized Linear Mixed-Effects Models*

GLM – *Generalized Linear Models*

HLR – *Hierarchical Logistic Regression*

HGLM – *Hierarchical Generalized Linear Mixed-Effects Models*

MLP – *Multi Layer Perceptron*

MLR – *Multilevel Logistic Regression*

OLS – *Ordinary Least Squares*

ANN – *Artificial Neural Network*

ROC – *Receiver Operating Characteristic*

SVM – *Support Vector Machines*

XGBOOST – *Extreme Gradient Boosting*

1 INTRODUÇÃO

1.1 Caracterização do tema da pesquisa

Durante a última década, o mercado passou por mudanças tornando os clientes muito mais exigentes, pois passaram a utilizar diversas outras plataformas tecnológicas como *web sites* e dispositivos *wireless* para interagir com as empresas que por sua vez começaram a investir em plataformas multicanais e em estratégias para retê-los, uma vez que um cliente consegue rapidamente absorver informações de diversos sites diferentes (RANGASWAMY; BRUGGEN, 2005).

Essa retenção tornou-se fundamental e um direcionador no gerenciamento de relacionamento com o consumidor, *Customer Relationship Managment* (*CRM*) por parte das instituições, pois a deserção de clientes impacta diretamente no lucro das instituições. Portanto torna-se crucial desenvolver um modelo de rotatividade eficaz e preciso para gerenciar com eficiência o relacionamento com o cliente (PFEIFER; FARRIS, 2004; PRASHANTH; DEEPAK; MEHER, 2017).

Dentre os modelos *estado-da-arte* da literatura para previsão de rotatividade de clientes (*client churn*) o *XGBoost* demonstra grande superioridade aos demais na predição correta de um grande número de possíveis desertores (*turners)* em comparação com a regressão logística, *Support Vector Machine (SVM),* árvore de decisão e *Random Forests* (SHARMA; GUPTA; MOHIT GOEL, 2020).

Entretanto modelos baseados em árvores não são interpretativos como modelos de abordagem tradicionais e a regressão logística é o mais balanceado tanto em precisão quanto poder explicativo, pois é possível identificar a relação causal das variáveis *X* na rotatividade dos clientes *Y*, sendo mais interpretativo gerencialmente (HILLS *et al.*, 2020).

Contudo, os modelos de regressão logística simples não são capazes de lidar com dados longitudinais que estão amplamente disponíveis nos bancos de dados das empresas devido à sua natureza assíncrona, bem como a comum existência de *missing values*, a forma encontrada por pesquisadores para adequá-los ao modelo é por meio de agregação ou retangularização transformando-os em dados estáticos. Já modelos multiniveis são capazes de lidar com os problemas de dados longitudinais (CHEN; FAN; SUN, 2012; JESKE; LI; WONG, 2012).

1.2 Problema e objetivo central da pesquisa

Na busca de encontrar respostas aos problemas apresentados e debruçada na literatura acadêmica dos últimos anos, este projeto terá como objetivo central comparar a acuracidade da regressão logística em perspectiva multinível que tem sido pouco explorada na predição da rotatividade de clientes com o *Extreme Gradient Boosting (XGboost)*, algoritmo recente com alto poder preditivo e que vem demostrando superioridade aos métodos tradicionais neste tipo de classificação (rotatividade de clientes).

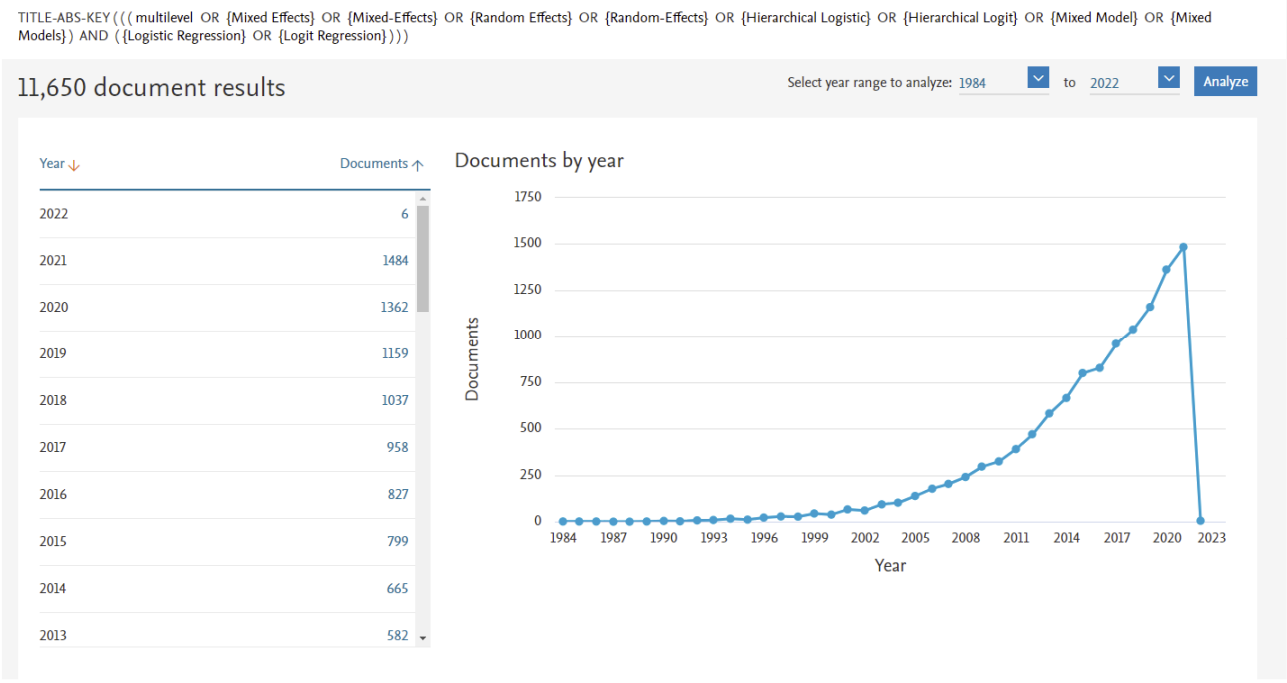
Serão objetivos específicos desta pesquisa: 1-Analisar a viabilidade de um modelo de regressão logística multinível para este problema. 2- Comprovar ou não a superioridade do *XGboost*. 3-Caso a regressão logística multinível demonstre superioridade, analisar a relação das variáveis escolhidas com a probabilidade de haver ou não a rotatividade de clientes de forma à auxiliar a gerência no entendimento do negócio. 4-Caso não haja

diferenças significativas entre um modelo de regressão logística simples, com um modelo multinível de acordo com as variáveis hierárquicas visíveis escolhidas. Será proposto um modelo misto com *Kmeans* para analisar a acurácia de uma possível regressão multinível logística de varáveis latentes com o *XGboost*. 5-Por fim espera-se contribuir com a academia em vista a pouca exploração de modelos de regressão multinível em problemas de negócios.

1.3 Justificativa da pesquisa

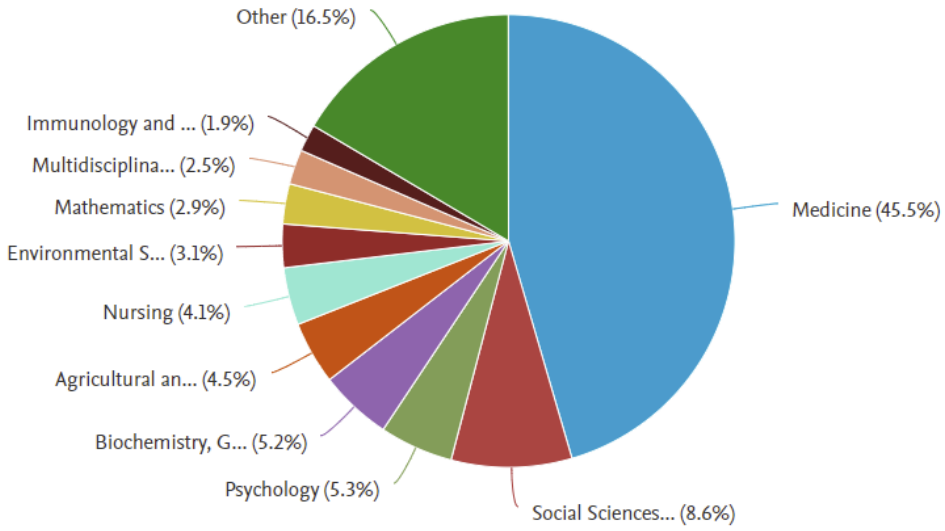
Partindo desse pressuposto o grande potencial de uma modelagem de regressão logística multinível é evidente. E em uma busca sistemática nas bases de dados relevantes da *SCOPUS*, *Engineer Village*, *IEEX* e *Web of Science* demonstraram o crescimento vertiginoso desta técnica (Figura 1), contudo é bastante utilizada nas áreas da saúde e ciências sociais (Figura 2), mas muito insipiente na área de negócios (Figura 3) demonstrando um grande potencial a ser explorado.

Figura 1 – Estudos de Regressão Logística Multinível ao longo dos anos.



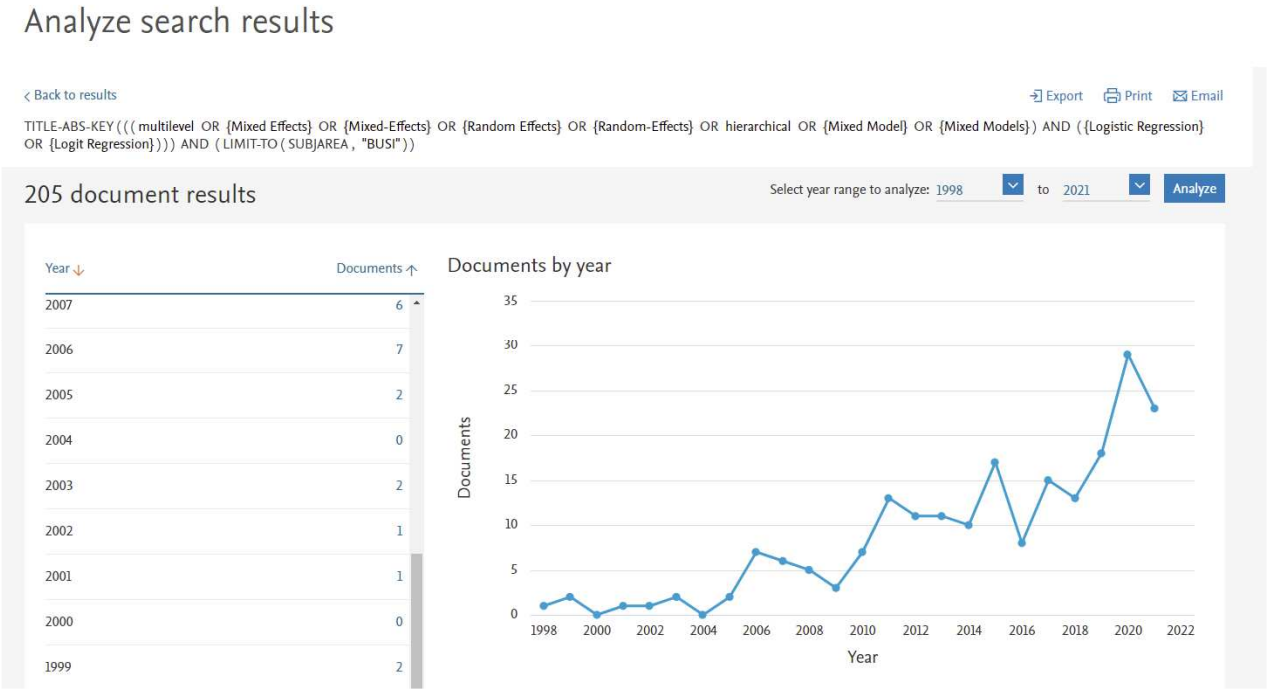
Fonte: Scopus, 2021.

Figura 2 – Estudos de Regressão Logística Multinível por categoria de estudo.



Fonte: Scopus (2021).

Figura 3 – Estudos de Regressão Logística Multinível focado na área de negócios.



Fone: Scopus (2021).

Quando partimos para a problemática de *client churn* um único artigo apresentado na 11ª conferência internacional de Computação Social e Media Social aborda o problema utilizando uma técnica multinível. Contudo, o modelo utilizado por Iwata, Otake e Namatame (2019) era uma regressão logística hierárquica baeysiana e não uma modelagem logística baseada em Modelos Lineares Generalizados Mistos (*Generalized Linear Mixed Models - GLMM)* como o proposto por Jeske, Li e Wong (2012).

Dado este panorama e identificado o *GAP* literário, torna-se relevante a investigação de uma modelagem de regressão logística multinível baseada em *GLMM* voltada à predição de *client churn* e confrontá-la com outros modelos de classificação em *estado-da-arte*, uma vez que a regressão logística linear já é bem performática e pode ser aprimorada com efeitos aleatórios nos interceptos e nas suas inclinações, bem como a inclusão de dados longitudinais, assíncronos ou *missing values* em sua modelagem.

2 METODOLOGIA

A pesquisa será realizada em uma das maiores varejistas de moda do Brasil a qual apresenta uma base de dados vasta e consolidada, devido à proximidade entre o pesquisador, empresa e a relevância da empresa perante o mercado brasileiro.

A priori, os dados a serem coletados serão extraídos diretamente do banco de dados no formato *.csv* e tabulados e operacionalizados no R-Studio. A tabela a ser analisada estará no formato *wide*, onde as observações estarão em linhas e as variáveis em colunas. A quantidade de observações ainda será definida, contudo as observações serão a nível de clientes em todo o território nacional, caso a base fique muito pesada será processada diretamente no *Apache Spark* em linguagem R.

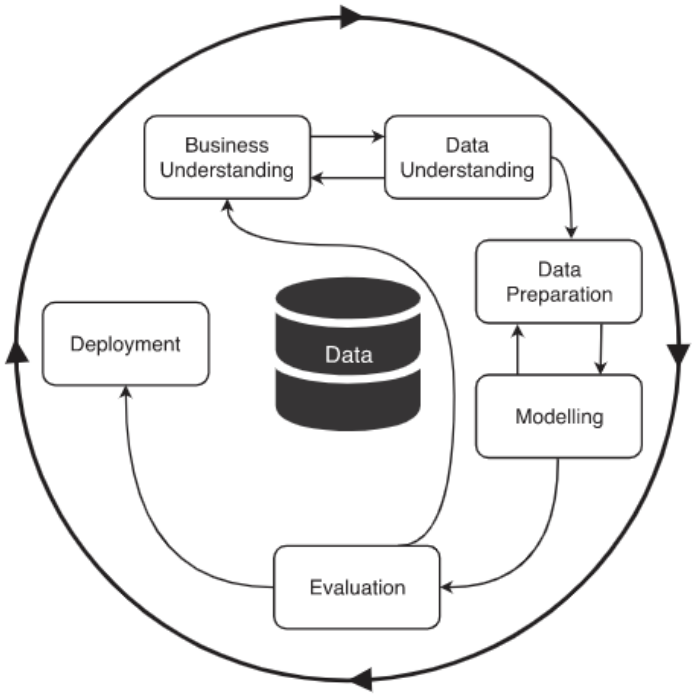
Pensando em uma abordagem metodológica completa, pesquisa estará estruturada de acordo com o modelo de referência *Cross Industry Standard Process for Data Mining* (CRISP-DM) que promove uma revisão de todo o ciclo de vida de um projeto de data mining, onde contempla suas fases, tarefas e relacionamentos entre elas. (CHAPMAN *et. al*, 1999).

Mesmo sendo a metodologia mais utilizada em projetos de mineração de dados, nem sempre seus resultados são positivos, pois não inclui algumas atividades de gerenciamento de projeto (MARISCAL *et. al*, 2010).

Contudo, ainda hoje é o modelo padrão *de facto* para mineração de dados e projetos exploratórios em ciência de dados e mesmo com suas limitações não deve ser descartado para projetos que vão dos dados ao conhecimento, quando se tem um claro objetivo de negócios que se traduz dentro de um objetivo de mineração de dados. (PLUMED *et. al*, 2021).

O modelo de referência CRISP-DM consiste em 5 fases de acordo com a Figura 4. São elas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e produção.

Figura 4 – Modelo de referência CRISP-DM.



Fonte: Plumed *et. al* (2021).

Primeiramente para que sejam extraídos dados relevantes para a criação do modelo, torna-se necessário o **entendimento do negócio**. Uma entrevista prévia com a gestão e a equipe técnica para nortear quais características das observações podem influenciar ou não na rotatividade dos clientes.

Posteriormente um segundo contato para esclarecer o devido **entendimento dos dados**, verificar os tipos de variáveis, granularidade, estrutura hierárquica, como são extraídas, a relação ao longo do tempo e a relação entre elas.

O próximo passo é todo o processo de limpeza e **preparação dos dados**, a verificação de *missing values* e a retirada ou não caso haja necessidade. Também faz parte do *setup* identificar a necessidade de alteração nos tipos de dados de *object* para *float* e *vise e versa*. Posteriormente as colunas do tipo *object* passarão por um processo de dummização.
Objetificando uma maior confiabilidade nos dados coletados, uma análise de *outliers* deverá ser feita repartindo as observações das colunas numéricas em quartis. E as que apresentarem distância maior que 1,6 dos quartis superior e inferior, serão excluídas.

Em busca de um melhor aproveitamento do número variáveis a serem utilizadas no modelo, será feito um procedimento *Stepwise* retirando as que não demonstrarem ser estatisticamente significante em um modelo de regressão linear simples.
Por fim os dados serão padronizados pela fórmula *z-score,* para que fiquem na mesma unidade de medida.

Depois de definidas as características relevantes, vem o processo de **modelagem,** nele o procedimento *Stepup* definirá a significância estatística do modelo multinível. Primeiramente, deve-se verificar os *p-values* do modelo nulo, desconsiderando a aleatoriedade nos interceptos e nas inclinações. Passando a 95% de confiança o próximo passo é considerar a

aleatoriedade nos interceptos e verificar a significância do modelo. E por fim avaliar estatisticamente uma equação multinível com interceptos e inclinações aleatórias.

No processo de **avaliação,** o modelo de regressão simples refinado pelo processo de *Stepwise* será confrontado com o hierárquico e na busca de uma validação mais relevante academicamente, em vista não só a avaliação de modelos estocásticos, bem como determinísticos, também será proposto um confronto com modelos em estado da arte na literatura como o *XGboost* e o *Decision Tree algorithm.*

Por fim, espera-se avaliar a curva ROC de ambos os modelos e os desempenhos de sensibilidade, especificidade e acurácia para que o melhor seja escolhido.

3 REVISÃO DA LITERATURA

3.1 Classificação

A Classificação é um dos problemas mais comuns em análise preditiva, é parecida com a análise de *cluster*, contudo enquanto *cluster* determina grupos com base na similaridade dos dados, sua tarefa consiste em atribuir rótulos novos a um objeto não rotulado de acordo com seus atributos preditivos, ou seja, busca no conjunto de dados já particionados esta informação, atribuindo-lhe classe, categoria ou um valor qualitativo. (MOREIRA; CARVALHO; HORVATH, 2019 AGGARWAL, 2015)

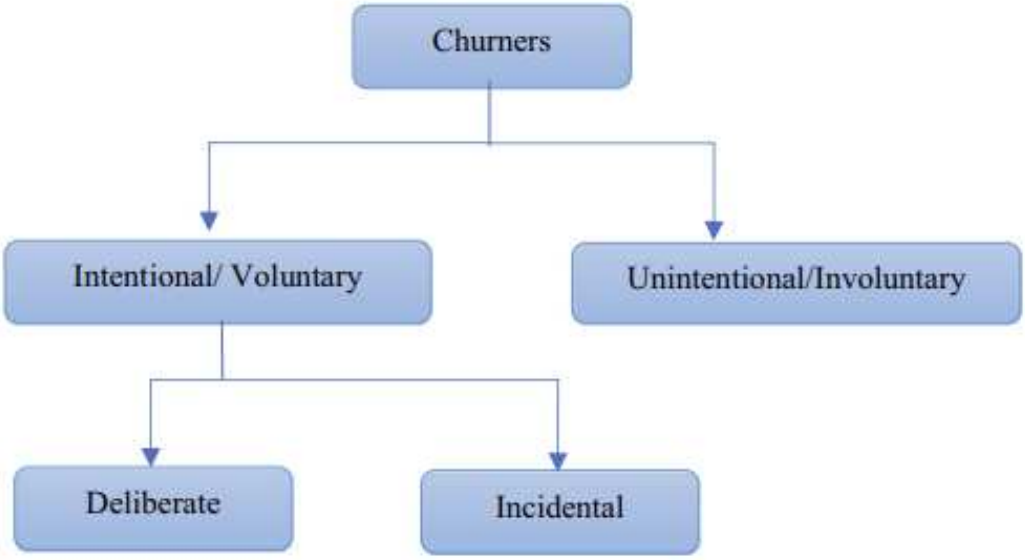
Essa categoria também chamada de variável categórica ou qualitativa (em vista a diferenciação das quantitativas e numéricas) possui escalas em um conjunto de categorias podendo estas serem binárias, para a existência ou não do evento (sim ou não, a favor ou contra, etc.), bem como nominal para três ou mais categorias em que não exista uma ordem/hierarquia nos dados entre as classes ou ordinal, quando exista esta relação entre elas. (AGRESTI, 2019)

3.1.1 Rotatividade de clientes

A predição da rotatividade normalmente pode ser considerada como um problema de classificação binária onde avalia-se entre *churn* ou *not churn*, em outras palavras a existência ou não do evento *churn*. (ZHAO; LI B.; LI X.; LIU; REN, 2005)

Churn vem da palavra *attrition* (atrito) ou *turnover* (rotatitividade), quando um cliente deixa de utilizar um serviço ou produto de uma determinada empresa chama-se de *customer churn* (rotatividade de clientes). Empresas de telefonia dividem estes clientes em *churners* voluntários os quais acharam serviços ou produtos mais atrativos ou involuntários quando não pagam suas contas ou quando são encontradas irregularidades nos contratos. *Churners* voluntários podem ser divididos em deliberados quando não estão satisfeitos, ou incidentais quando são causas externas como a troca de emprego ou mudança de localidade. Esta subdivisão pode ser pode ser visualizada na Figura 5. (JAIN; KHUNTETA; SRIVASTAVA, 2021)

Figura 5 – Tipos de rotatividade de clientes.

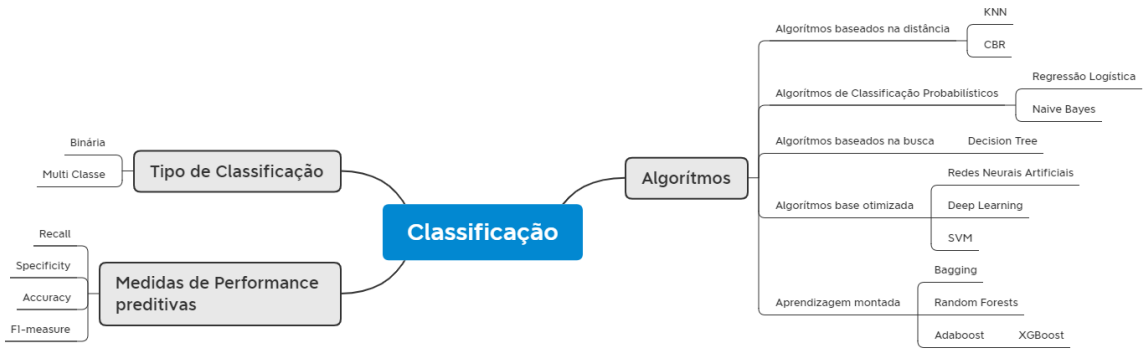


Fonte: Jain, Khunteta e Srivastava (2020).

3.1.2 Mapas Mentais

Seguindo um mapa mental de acordo com a Figura 6 adaptado do livro de Moreira, Carvalho e Horvath (2019), pode-se dividir a modelagem de problemas de classificação em: Tipos de classificação, de acordo com a escala das varáveis; Algoritmos, utilizados para a modelagem e resolução de problemas de categorização; e medidas de performance preditivas, estatística de controle e melhoria do modelo.

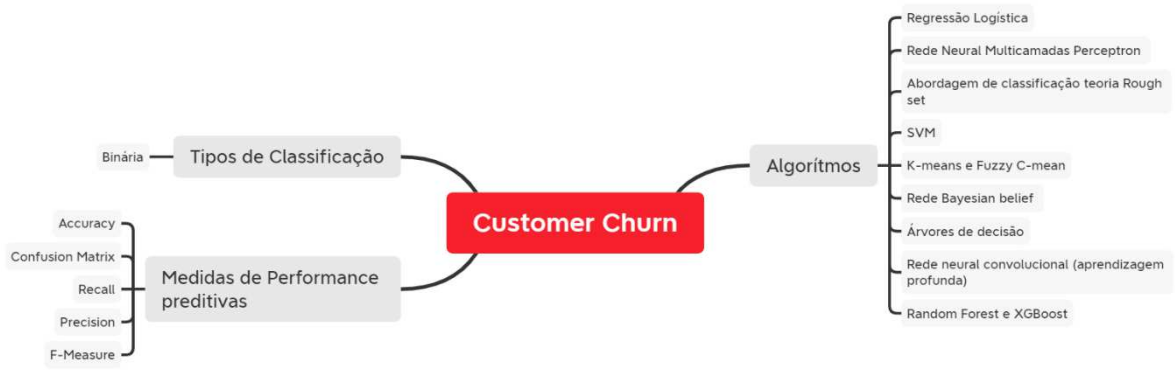
Figura 6 – Mapa Mental do livro *An Introduction to Data Analytics* sobre classificação.



Fonte: Moreira, Carvalho e Horvath (2019).

Quando o problema de classificação é *client churn* o mapa pode ser adaptado seguindo a revisão sistemática de Jain, Khunteta e Srivastava (2021) de acordo com a figura 7 onde serão utilizados algoritmos para a modelagem binária.

Figura 7 – Mapa Mental sobre *client churn.*



Fonte: Jain, Carvalho e Horvath (2021).

3.1.3 Algoritmos

3.1.3.1 Regressão Logística

O modelo de regressão logística binária tem como objetivo principal a estudar a probabilidade de ocorrência ou não de um evento de interesse *Y*, sendo $Y = 1$ a probabilidade de o evento ocorrer e $Y = 0$ a probabilidade de um não evento ocorrer com base no comportamento de variáveis explicativas. (FAVERO; BELFIORE, 2017)

Onde:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (1)$$

O maior problema com a probabilidade linear é que sua probabilidade está limitada a 0 e 1, mas funções lineares inerentemente não são limitadas. Por isso necessita-se retirar esses limites superiores e inferiores por meio do logarítimo das chances. (ALLISON, 2012)

Esse modelo foi proposto por Walker e Duncan (1967) foi a primeira transformação logística binária assumindo que o a probabilidade de ocorrência ou não do evento fosse uma curva sigmoide simétrica.

O modelo proposto era:

$$P = (1 + e^{-x'\beta})^{-1} \quad (2)$$

De acordo com Harrell (2001) a solução da equação para *x* resulta em:

$$1 - P = \exp(-x) / [1 + \exp(-x)] \quad (3)$$

E a sua inversa em:

$$x = \log\left[\frac{P}{1-P}\right] = Log[odds\ that\ Y = 1\ occurs] = \text{logit}\{Y=1\} \quad (4)$$

Favero e Belfiore (2017) demonstram que para adquirir a probabilidade de ocorrência ou não do evento em estudo em função do logito, isola-se *P* a partir da expressão:

$$\left[\frac{P}{1-P}\right] = e^{Z} \tag{5}$$

$$P = (1-P).e^{Z} \tag{6}$$

$$P.(1+e^{Z}) = e^{Z} \tag{7}$$

Com isso, obtém-se:
Probabilidade de ocorrência do evento:

$$\mathrm{P} = \left[\frac{e^{Z}}{1+e^{Z}}\right] = \left[\frac{1}{1+e^{-Z}}\right] \tag{8}$$

Probabilidade de ocorrência do não evento:

$$1 - \ \mathrm{P} = 1 - \left[\frac{e^{Z}}{1+e^{Z}}\right] = \left[\frac{1}{1+e^{Z}}\right] \tag{9}$$

3.1.3.1.1 Procedimento Stepwise

Stepwise é o método usual para selecionar um modelo que seja estatisticamente significante em modelos de regressão, onde são feitas adições e/ou subtrações sistemáticas dos termos de acordo com a sua abordagem e critério de corte. Na *abordagem backward simplification* é feito um *fit* do modelo com todas as variáveis originais e *a posteriori* é feita a busca e retirada uma a uma das que não demonstraram ser estatisticamente significantes. Já a abordagem *foward selection* assimetricamente começa de um modelo simples onde são adicionados sucessivamente termos que se demonstrarem serem mais significativos. O procedimento pode ser repetido até que todos os efeitos na fórmula sejam considerados significativos. (CALCAGNO; MAZANCOURT, 2010)

Determinar a remoção ou adição de um determinado termo em cada teste pode ser feito de várias maneiras uma delas e bastante utilizada é o critério de informação de Akaike (AIC) proposto por Venables e Ripley (2002) que faz parte da função *step()* do pacote *stats* (*version 3.6.2*) ou *stepAIC()* do pacote MASS *(version 7.3-54)* do R.

Estas funções podem ser usadas para automatizar o processo de seleção dos termos e contempla as duas abordagens bastando informar apenas: um modelo ajustado para iniciar o processo, quanto mais próximo do modelo final mais vantajoso; uma lista de duas fórmulas definindo os superiores e mais complexos ou inferiores e mais simples; e a escala de estimação. Se o ponto de partida for um modelo mais complexo o algoritmo irá performar em *backward simplification,* caso contrário em *foward selection.* (VENABLES; RIPLEY, 2002)

3.1.3.2 Rede Bayesiana (*Bayesina Belief Network*)

Uma rede bayesiana (BN) é um modelo gráfico para relacionamentos probabilísticos entre um conjunto de variáveis. A probabilidade bayesiana de um evento x é o grau de crença de uma pessoa nesse evento. Diferentemente da probabilidade clássica que é uma propriedade física do mundo, onde ela é calculada com base nas repetições de um ou mais eventos, a probabilidade bayesiana é uma propriedade da pessoa que atribui a probabilidade, ou seja, é uma probabilidade com base na crença, por isso também são conhecidas como redes de opinião. (HECKERMAN, 1996)

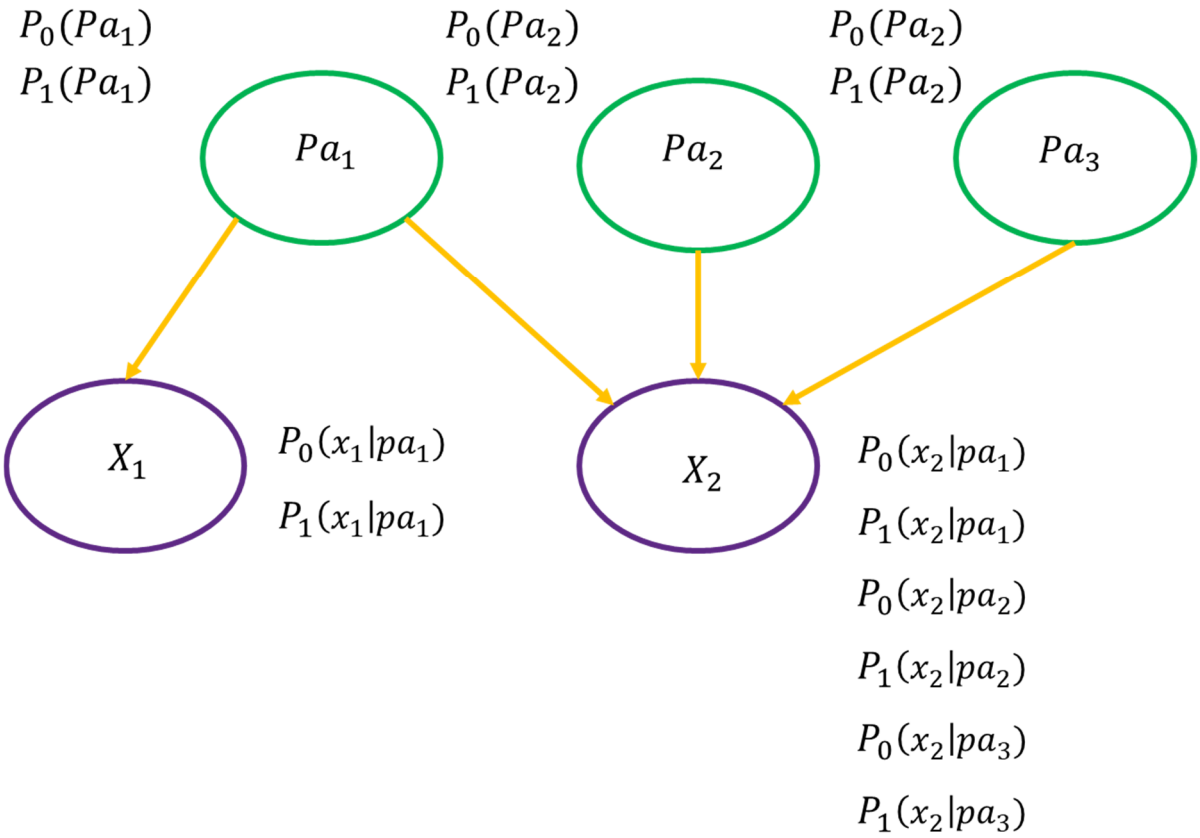
BN utiliza o Teorema de *Bayes* usado para calcular a probabilidade de um objeto pertencer para uma classe particular. Uma rede bayesiana para um conjunto de variáveis X = $\{X_1,..., X_n\}$ consiste em uma estrutura de rede S que codifica um conjunto de afirmações de independência condicional sobre as variáveis em X, e um conjunto P de distribuições de probabilidade local associadas a cada variável. Juntos, esses componentes definem a distribuição de probabilidade conjunta para X de acordo com a equação 10. (MOREIRA, CARVALHO; HORBATH, 2019; HECKERMAN, 1996)

$$p(x) = \prod_{i=1}^{a} p\,(x_i|pa_i) \tag{10}$$

Onde,
P(x) a distribuição de probabilidade conjunta para *X*
Xi para denotar as variáveis e seu nó correspondente
P_{ai}, Nó de hierarquia superior ao nó X_i em S

Em particular, para construir uma rede bayesiana para um determinado conjunto de variáveis, simplesmente desenhamos arcos das variáveis de causa para seus efeitos imediatos. Em quase todos os casos, isso resulta em uma estrutura de rede que satisfaz a definição da Equação 10, obtendo assim, a estrutura de rede na Figura 8.

Figura 8 – Estrutura de uma BN com 5 variáveis binárias sendo 3 pais.



Fonte: Adaptado de Heckerman (1996).

3.1.3.3 Teoria dos Conjuntos Aproximados (Rough Set Theory)

A Teoria dos Conjuntos Aproximados ou *Rough Set Theory* foi elaborada por Pawlak (1987), matemático polonês que elaborou uma filosofia que parte do pressuposto em que cada objeto do universo do discurso está associado à alguma informação.

$$A = (U, R) \tag{11}$$

U = Universo Finito
R = Relação equivalente indiscernível

Objetos que detêm as mesmas informações são indiscerníveis (similares) em uma visão da informação disponível sobre eles. Por sua vez o conjunto dos objetos semelhantes é chamado de conjunto elementar e forma uma grânula (átomo) básica de conhecimento sobre o universo. Qualquer união de alguns conjuntos elementares é referida como um conjunto preciso (*crisp*) ou aproximado (*rought*). Conjuntos aproximados não podem ser caracterizados em termos de informações sobre seus elementos, contudo podem ser aproximados dos conjuntos precisos, essas aproximações são chamadas de aproximações superiores e inferiores. (PAWLAK, 2002)

De acordo com ZIARKO (1993) as elas podem ser representadas pelas Equações 12 e 13 e as regiões de fronteira e negativa (intercessão e união) pelas equações 14 e 15, abaixo:

$$\underline{\mathrm{R}}_{\beta}X = \cup \{E \in R^{*}: c(E,X) \leq \beta\} \quad (12)$$

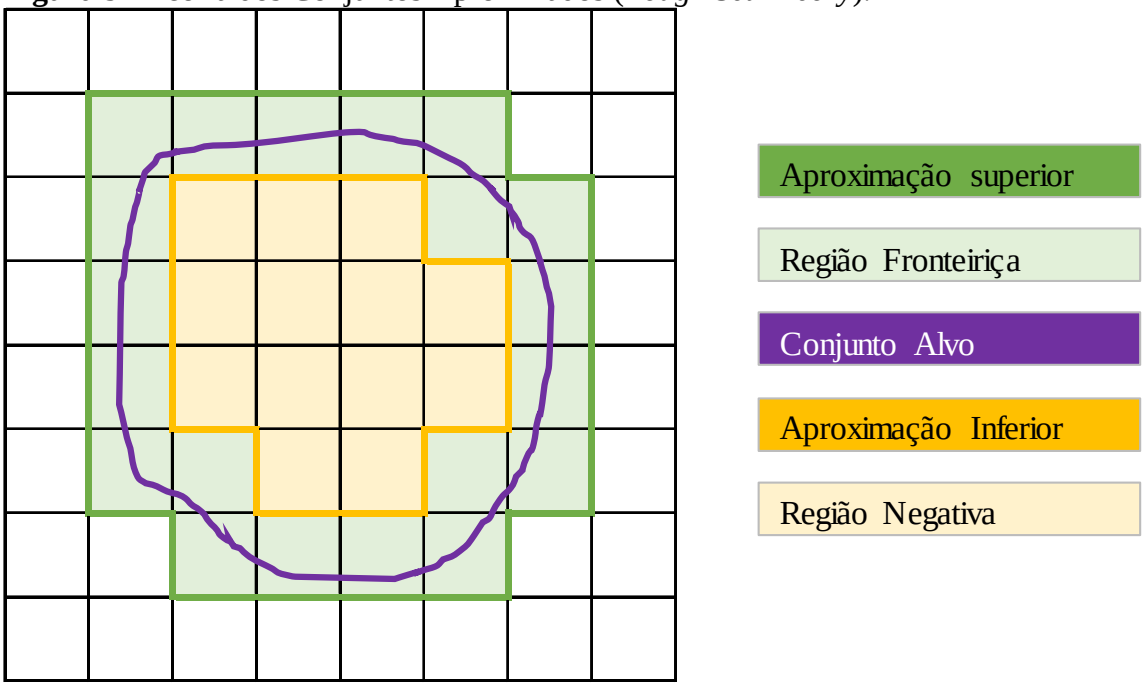
$$\overline{\mathrm{R}}_{\beta}X = \cup \{E \in R^{*}: c(E,X) \leq 1-\beta\} \quad (13)$$

$$\mathrm{BNR}_{\beta}X = \cup \{E \in R^{*}: \beta < c(E,X) \geq 1-\beta\} \quad (14)$$

$$\mathrm{NEGR}_{\beta}X = \cup \{E \in R^{*}: c(E,X) \geq 1-\beta\} \quad (15)$$

A Figura 9 consegue sintetizar bem essa ideia de universo, aproximações, limites:

Figura 9 - Teoria dos Conjuntos Aproximados (*Rough Set Theory*).



Fonte: Adaptado de Pawlak (1982).

Nos últimos anos as pesquisas e aplicações da teoria do conjunto aproximado vem atraindo a atenção dos pesquisadores. Combinações com outras tecnologias e métodos teóricos emergiram, como: Conjuntos aproximados probabilísticos, Modelo de decisão teórica por conjuntos aproximados, Modelo variável precisa por conjuntos aproximados, Modelo multigranulado por conjuntos aproximados e modelo de teoria dos jogos por conjuntos aproximados bem como aplicações variadas como conjuntos aproximados focados na redução de atributos, aquisição de regras, algoritmos inteligentes e classificação, sendo uma ferramenta muito útil em tarefas data mining. (ZHANG; XIE; WANG, 2016)

3.1.3.4 Support Vector Machines (SVM)

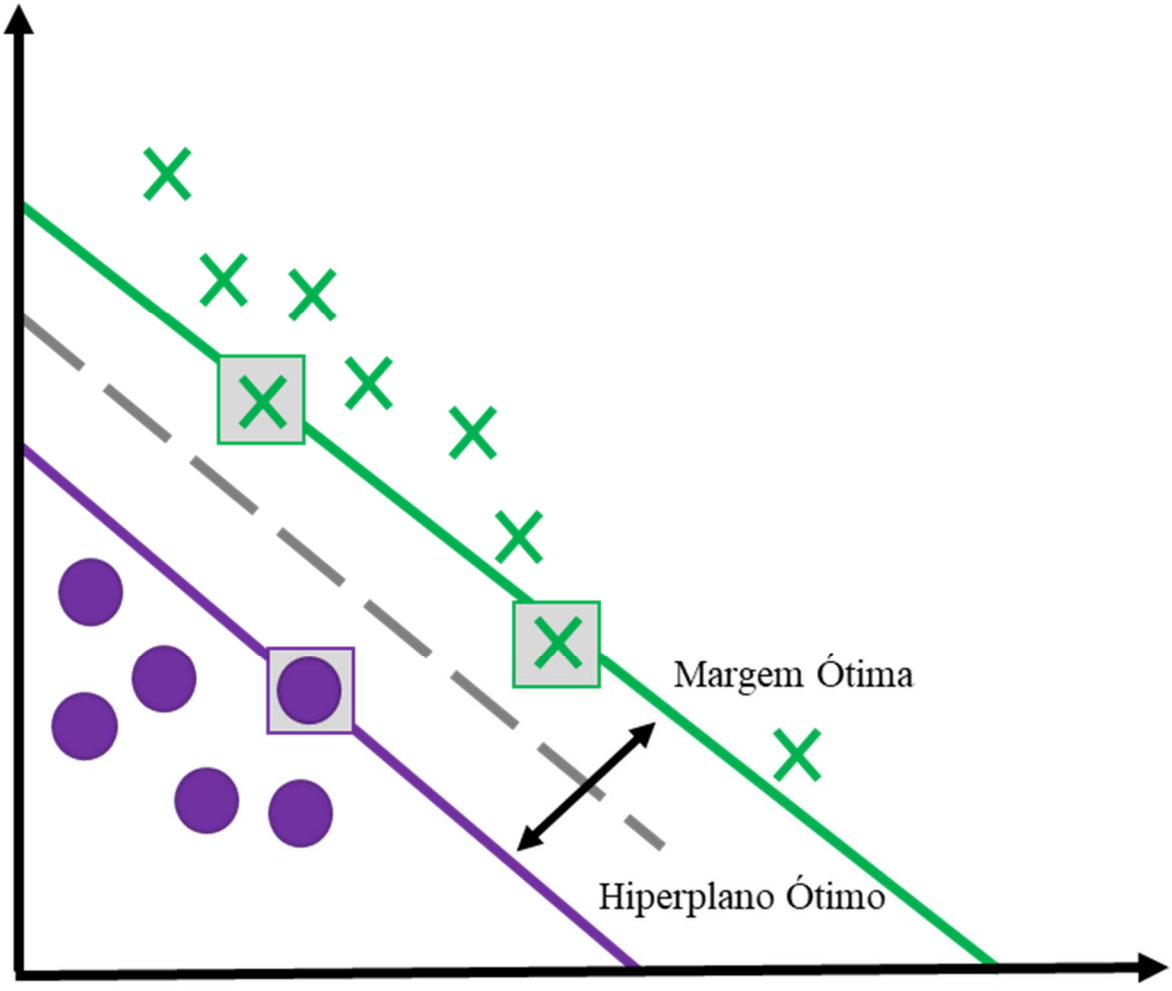
Support Vector Machines é uma máquina de aprendizagem desenvolvida com base na arquitetura das redes neurais e na teoria de aprendizagem estatística que mapeia a

entrada vetores em algum espaço de característica de alta dimensão por meio de algum mapeamento não linear escolhido *a priori*. Neste espaço, uma superfície de decisão linear é construída com propriedades especiais que garantem alta capacidade de generalização da rede. (CORTES; VAPNIK, 1996)

Este algoritmo veio a suprir dois dos problemas em aprendizagem de RNAs que são as gerações de diferentes modelos para diferentes funções e a falta de uma base matemática mais sólida. Ele foi originalmente projetado para tarefas de regressão e classificação binária. Nesse contexto os SVMs tentam encontrar um hiperplano linear ótimo de modo que a margem de separação entre os exemplos positivos e negativos seja maximizada. Para este fim, objetos de treinamento de diferentes classes são selecionados para serem vetores de suporte. Esses vetores definem uma fronteira de decisão capaz de maximizar a margem de separação entre a fronteira e os objetos das duas classes, contudo a margem a maximização reduz o número de modelos possíveis e aumenta a capacidade do modelo de generalização e, portanto, a ocorrência de *overfitting*. (COUSSEMENT; VAN DEN POEL, 2008; MOREIRA; CARVALHO; HORVATH, 2019)

Um exemplo de um problema separável em um espaço bidimensional. Os SVMs, marcados com quadrados cinzentos, definem a margem de maior separação entre as duas classes de acordo com a Figura 9.

Figura 10 – SVM e a margem de separação.



Fonte: Adaptado de Cortes e Vapnik (1996).

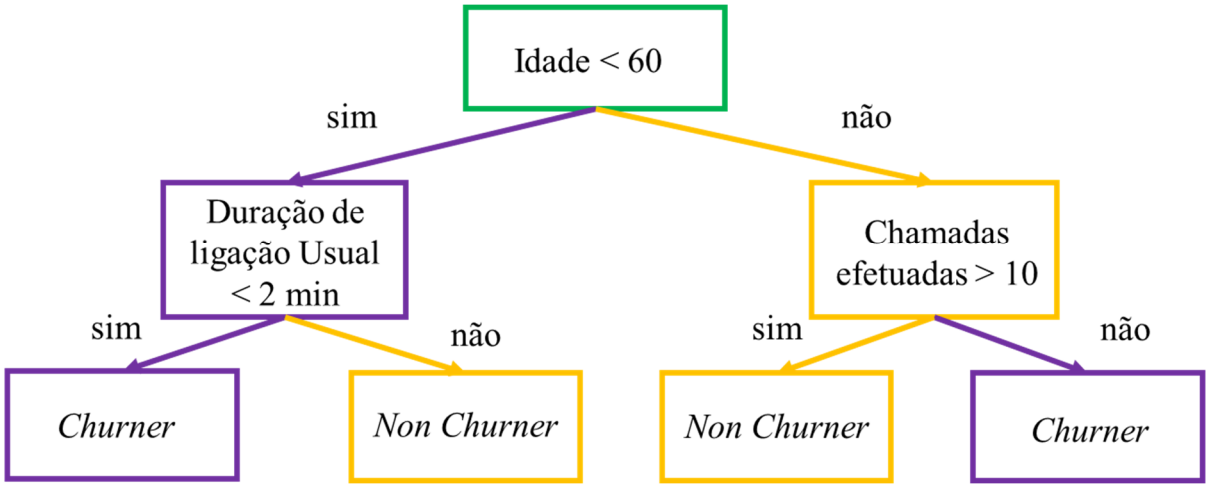
3.1.3.5 Árvore de decisão

Árvore de decisão é o modelo de predição mais popular voltado à problemática da rotatividade dos clientes segundo a revisão sistemática de Mahajan V., Misra e Mahajan R. (2015).

Esta vasta utilização, principalmente direcionada à área de negócios pode ser explicada pela alta interpretabilidade do modelo uma vez que pessoas são boas intérpretes de relatórios de fluxo onde o dado segue um ramo específico de acordo com o seu valor. Esta característica, torna-os adequados a problemas envolvendo extração de conhecimento. (MOREIRA; CARVALHO; HORVATH, 2019; STRECHT;MOREIRA; SOARES, 2021)

CART, *a Classification And Regression Tree* descrito *a priori* no livro de Breiman, Friedman, Olshen e Stone (1984), demonstra a estrutura dos classificadores binários por árvores de decisões que são criadas por repetidos cortes nos *subsets* formando subgrupos dado um determinado critério. A árvore cresce até que a diminuição da impureza fique abaixo de um limiar definido pelo usuário. Cada nó em uma árvore de decisão é uma condição de teste e a ramificação é baseada no valor do atributo que está sendo testado. Na Figura 10 é possível visualizar como funciona a *Decision Tree* para a previsão da rotatividade de clientes nas empresas de telecomunicação. A árvore está representando uma coleção de múltiplos conjuntos de regras. Ao avaliar um conjunto de dados do cliente, a classificação é feita atravessando a árvore até que um nó de folha seja atingido. O rótulo deste nó de folha (*Churner* ou *Non Churner*) é atribuído ao registro do cliente sob avaliação. (LAZAROV; CAPOTA, 2007)

Figura 11 – Árvore de decisão.



Fonte: Adaptado de Lazarov e Capota (2007).

3.1.3.6 Random Forest

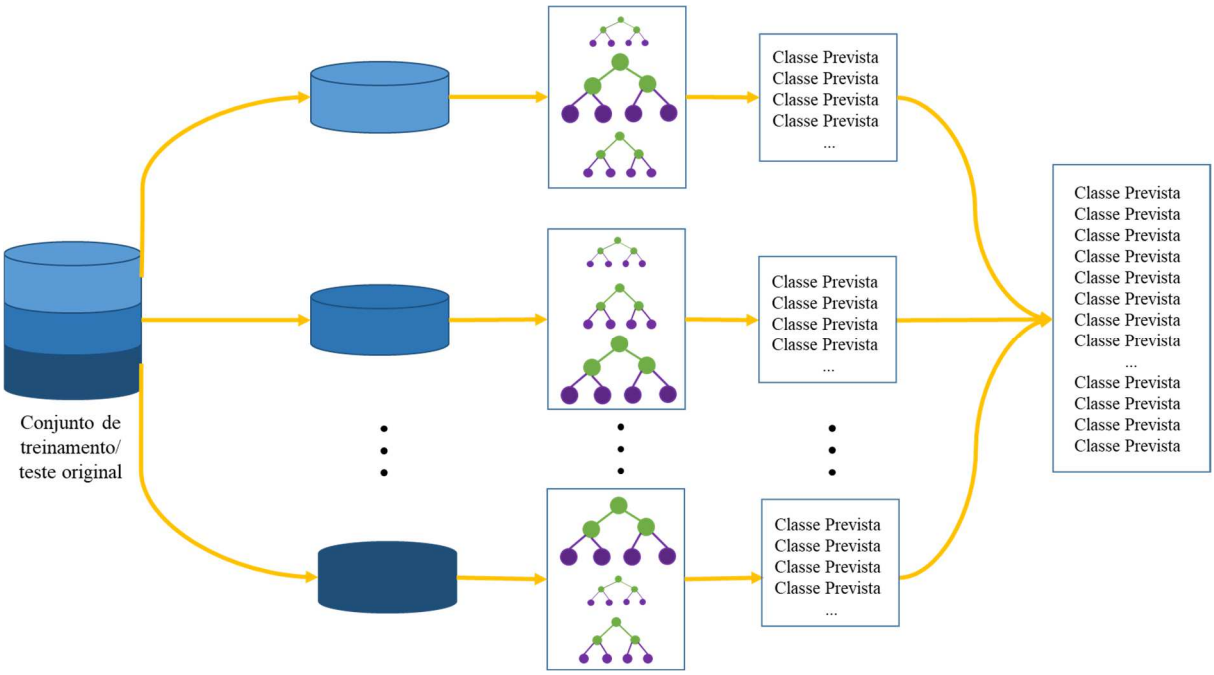
Random Forests foi desenvolvido por Breiman (2001) é uma combinação de arvores de decisões, onde cada uma é criada utilizando uma amostra *bootstrap* diferente. A seleção do classificador de base depende da variância e da natureza de polarização do classificador, em geral, os classificadores com alta variância e baixo viés são preferidos. Sendo assim, a árvore de decisão é usada como um classificador de base para florestas aleatórias. Em cada nó da árvore, apenas um número pré-definido de atributos selecionados aleatoriamente é utilizado. Arvores aleatórias crescem e usam uma floresta ou coleção de árvores de classificação

e cada árvore será treinado de forma independente e pode ser feito em paralelo. (MOREIRA; CARVALHO; HORVATH, 2019; JAYASWAL; PRASAD; TOMAR; AGARWAL, 2016)

Em seus experimentos Breiman (2001) utilizou *bagging* em conjunto com o recurso aleatório seleção. Cada novo conjunto de treinamento é desenhado, com substituição, do conjunto de treinamento original. Em seguida, uma árvore é cultivada no novo conjunto de treinamento usando a seleção aleatória de recursos. As árvores que cresceram não são podadas. Isso garante um aumento de precisão e reduz o erro de generalização do modelo bem como a força e as correlações entre as árvores.

Esse funcionamento do *random forests* é possível ser observado na Figura 12.

Figura 12 – *Random Forests.*



Fonte: Adaptado de Prasad, Tomar e Agarwal (2016).

3.1.3.7 XGBoost

XGBoost significa *Extreme Gradient Boosting*, que é uma combinação entre gradiente descendente e *boosting*. O *boosting* funciona aplicando sequencialmente um algoritmo de classificação a versões reponderadas dos dados de treinamento e, em seguida, obtendo uma votação por maioria ponderada da sequência dos classificadores produzidos, melhorando assim o desempenho consideravelmente da maioria dos algoritmos. (FRIEDMAN, HASTIE, TIBSHIRANI,2000; BHUSE; GANDHI; MESWANI; MUNI; KATRE, 2020)

XGBoost implementa o *CART* com aumento de gradiente e usa as informações da primeira derivada e da função de segunda derivada. O aumento do gradiente segue uma abordagem onde novos modelos são usados para calcular o erro ou resíduos do modelo anteriormente aplicado e, em seguida, ambos são combinados para fazer a previsão final. Ele também usa o gradiente descente para localizar *a minima* ou reduzir o valor da função de perda. O parâmetro de regularização e a função de complexidade são adicionados à função objetivo

para melhorar a precisão da previsão. (LALWANI; MISHRA; CHADHA; SETHI, 2021; TANG, 2020)

XGboost dá bons resultados no modelo de predição na maioria dos problemas devido à sua boa otimização de cache, mas leva mais tempo de treinamento para o processo de iteração e é de difícil interpretação. Contudo possui boa velocidade de execução e desempenho de modelo. (BHUSE; GANDHI; MESWANI; MUNI; KATRE, 2020)

A equação 16 demonstra como funciona a estatística de gradiente de segunda ordem utilizado pelo modelo.

$$L(\Phi) = \sum_i l(\hat{y}i, yi) + \sum_k \Omega(\text{fk}) \, , onde \; \Omega\,(f) = \gamma T + \frac{1}{2}\lambda||w||^2 \qquad (16)$$

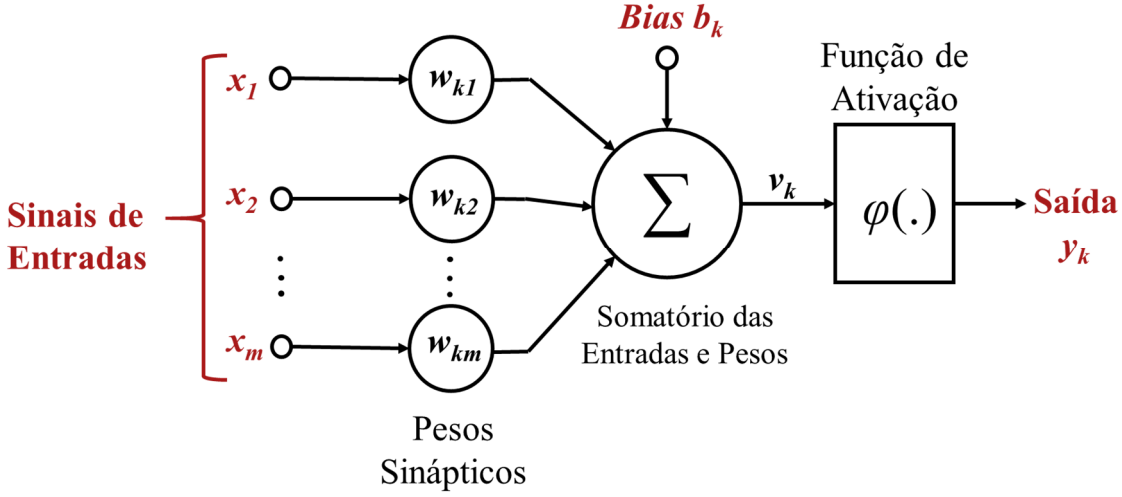
3.1.3.8 Rede Neural Multicamadas Perceptron

Redes Neurais Artificiais (ANN) são algoritmos computacionais modelados de forma estrutural e comportamental para serem semelhantes aos neurônios do cérebro humano, de forma com que ao serem treinados possam reconhecer e categorizar padrões complexos. (BISHOP, 1996)

O cérebro humano é um computador altamente complexo e não linear, que consegue realizar processamentos de informação de alta performance por meio da organização de seus componentes estruturais: os neurônios. Estas são células base, que processam informações e se comunicam com diversos outros em paralelo. (HAYKIN, 2008; BRAGA A. P.; CARVALHO A. P. L. F.; LURDEMIR T. B., 2007)

Cada neurônio artificial é formado por três elementos básicos representados no diagrama de blocos na figura 12, são eles:

Figura 13 Modelo não linear de um neurônio.



Fonte: Adaptado de Haykin (2008).

a) conjunto de sinapses ou elos de conexão, onde cada uma é representada por um peso ou força própria;
b) somador que irá realizar o somatório ponderado dos sinais de entrada;
c) função de ativação que irá restringir a amplitude de saída de um neurônio.

3.1.3.8.1 Função Ativação

A função ativação é representada por φ e define a saída de um neurônio em termos do campo local induzido. Existem alguns tipos básicos de função de ativação: função limiar, função limiar por partes, função Sigmoidal e função tangente hiperbólica. (HAYKIN, 2008)

A função Sigmoidal é preferível para a ativação de redes de dados binários, reduz a carga computacional, sendo assim a mais recomendada em casos de reconhecimento de padrão para classificação. (FAUSETT, 1994)

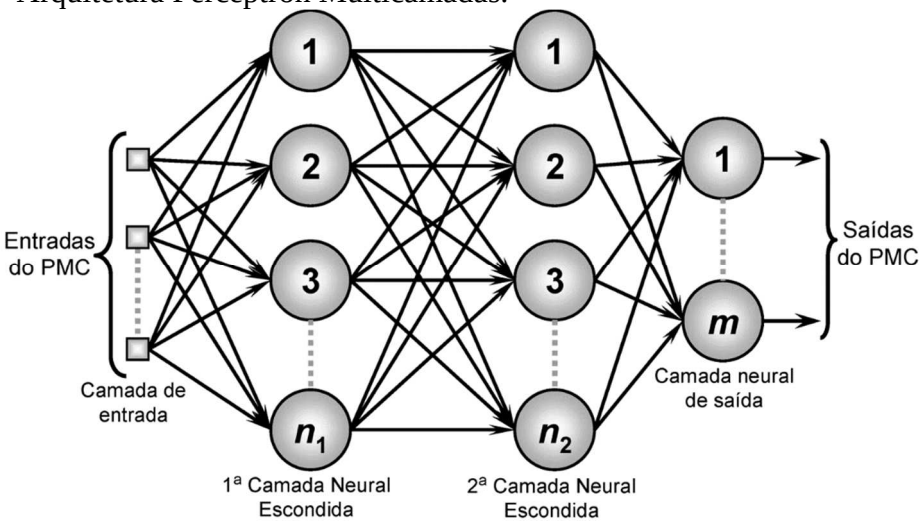
3.1.3.8.1.1 Função Sigmoidal

Tem sua função representada por um gráfico na fora de S, é a forma mais comum de uma função de ativação; assume sempre valores positivos e pode ser definida na equação 17.

$$\varphi(v) = \frac{1}{1 + \exp(-av)} \tag{17}$$

O MLP consiste em várias camadas (normalmente três camadas ou mais), onde cada camada está totalmente conectada com a próxima em um modo de alimentação direta. A primeira e a última camada representam as entradas e saídas do sistema, respectivamente. As conexões entre os nós são representadas como pesos. Arquiteturas mais complexas têm mais número de camadas. Um exemplo de ANN simples a arquitetura é mostrada na Figura 13. (ADWAN *et al.*, 2014)

Figura 14 - Arquitetura Perceptron Multicamadas.



Fonte: Adaptado de Haykin (2008).

3.2 Modelos Lineares Generalizados (GLM)

O termo GLM foi proposto por Nelder e Wedderbun (1972), que consolidou em seu estudo um conjunto de regressões lineares e não lineares como uma forma de padronizar variados estudos de regressão com base em peculiaridades que podem ser enquadradas dentro de componentes.

Os GLMs apresentam três componentes: 1- O componente aleatório, que identifica a variável de resposta *Y* e sua distribuição de probabilidade, entre elas, distribuição Normal ou Gamma para Mínimos Quadrados Ordinários (OLS), (onde a *Y* é variável contínua), Binomial para regressão logística, Poisson e Binomial negativa para dados de contagem e etc. 2- O Preditor linear ou a equação de predição, que especifica as variáveis exploratórias do modelo que entram como preditoras lineares do lado direito da equação de acordo com a equação 11. 3- A função de ligação canônica de cada modelo, sendo a mais simples a função $g(\eta_i) = \eta_i$ para OLS. (AGRESTI, 2019; FAVERO; BELFIORE, 2017)

$$\eta_i = \alpha + \beta_1.X_{1i} + \beta_2.X_{2i} + \cdots + \beta_k.X_{ki} \quad (18)$$

3.3 Dados Longitudinais

Dados Longitudinais, como o próprio nome presume, são dados provenientes de múltiplas observações de um mesmo objeto de estudo ordenadas ao longo do tempo. Economistas chamam isso de dados de painel ou séries temporais. A ordenação temporal das medições é importante porque as medições mais próximas no tempo dentro de um assunto são provavelmente mais semelhantes do que observações mais distantes. Os dados longitudinais têm uma estrutura natural hierárquica ou de vários níveis, com observações no nível inferior aninhadas ou agrupadas dentro de assuntos no nível superior. Levantamentos longitudinais usam o tempo do calendário em dias, meses ou anos, como a dimensão que separa as observações sobre o mesmo assunto. (WEISS, 2005)

Existem inúmeras vantagens em utilizar essas dimensões. Primeiro, na medida em que medições repetidas do mesmo sujeito não estão perfeitamente correlacionadas, estudos longitudinais são mais poderosos para um número fixo de assuntos. Ou seja, para atingir um nível semelhante de poder estatístico, menos assuntos são necessários em um estudo longitudinal, pois repetidas observações de o mesmo assunto, nunca, são perfeitamente correlacionadas. Assim, as medições repetidas de um único sujeito fornecem informações mais independentes do que uma única medida obtida de um único sujeito. Em segundo lugar, em um estudo longitudinal, cada sujeito pode servir como seu próprio controle, pois a variabilidade intrasujeito é menor do que a intersujeito, resultando em um teste mais sensível ou estatisticamente poderoso. Terceiro, estudos longitudinais permitem que um investigador separe os efeitos do envelhecimento (ou seja, mudanças ao longo do tempo dentro dos indivíduos), de efeitos de *coorte* (ou seja, diferenças entre os indivíduos na linha de base). Por fim, dados longitudinais podem fornecer informações sobre a mudança individual. (HEDEKER; GIBBONS, 2006)

3.4 Modelos lineares Mistos Generalizados (GLMM)

Baseado no estudo das GLMs proposto por Nelder e Wedderbun (1972) e no estudo de regressão separado de cada indivíduo para lidar com os problemas que vieram a ser chamados de efeitos aleatórios ou mistos proposto Korn e Wittemore (1979). Stirantelli, Laiard e Ware (1984) criaram os primeiros modelos de efeitos aleatórios para dados longitudinais de classificação, inclusive foi o primeiro modelo de regressão logístico a considerar efeitos mistos de dois níveis nos parâmetros da equação canônica. Posteriormente um modelo semelhante apresentado por Waclawiw e Liang (1993) aplicou um modelo também de efeitos mistos de 2 níveis, mas voltado para analisar respostas para dados discretos quando o número de repetidas observações são pequenas. Seu modelo conseguiu estimar a variância aleatória sem apelar para aproximações normais multivariadas.

Efeitos fixos são os parâmetros de uma equação canônica de uma GLM, descrevem o relacionamento entre uma variável dependente e suas preditoras. Já efeitos aleatórios, são valores randômicos associados com os níveis dos fatores aleatórios também chamados de variáveis categóricas latentes, em outras palavras, são os efeitos de uma variável de agrupamento (*cluster*) que incidem nos outros parâmetros da equação da GLM, podendo estes entrarem como interceptos ou coeficientes aleatórios. (AGRESTI, 2019; WEST; WELCH; GALECKI, 2014)

O termo GLMM foi cunhado a primeira vez por Breslow e Clayton (1993) e Wolfinger e O'Connell (1993), que sintetizaram em seus estudos parecidos, modelos GLM de efeitos fixos com uma extensão adicional que permitiu afeitos aleatórios também, estendendo a regressão ordinária a respostas não-normais e uma função de ligação da média de acordo com a equação 19.

$$g(\mu_{it}) = u_i + \alpha + \beta_1.X_{1it} + \beta_2.X_{2it} + \cdots + \beta_{kit}.X_{kit}\,, i = 1, \ldots, T. \qquad (19)$$

Onde,
Y_{it}, Observação *t* no cluster *i.*
u_i, Efeito aleatório para o cluster i
$\mu_i = E(Y_{it} \mid u_i)$, A média de uma variável de resposta dado um valor de efeito aleatório.
X_{kit}, Valor de uma variável exploratória T para a observação em que: k=1,2,3...T).

Este GLMM, tendo efeito aleatório como parte de um termo de interceptação, é chamado de interceptação aleatória modelo. Os efeitos aleatórios u_i não são observados, tratados como variáveis aleatórias independentes assumido como tendo uma distribuição normal N (0, σ^2). A variância σ^2 também é um parâmetro, chamado um componente de variância. (AGRESTI, 2019)

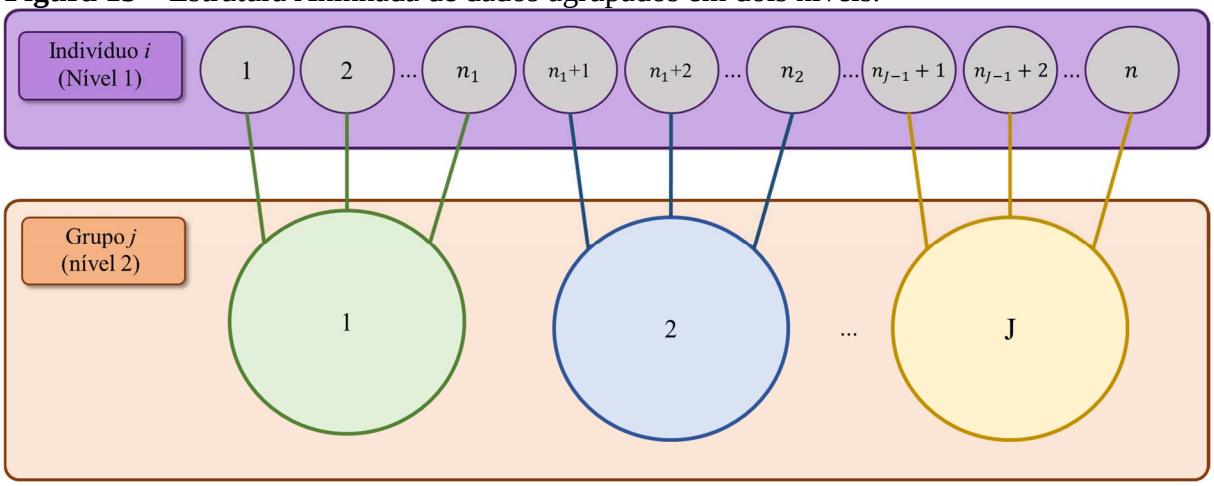
A estrutura linear de modelos de regressão combina muito bem com a modelagem hierárquica no contexto de modelos lineares normais, pois a estrutura hierárquica é utilizada como premissa adicional de linearidade e normalidade, preservando a linearidade do modelo como um todo e permitindo a conjugação em qualquer estágio do modelo. (MIGON; GAMERMAN; LOUZADA, 2015)

3.4.1 Modelos Hierárquicos Lineares Generalizados (HGLM)

Quando a hierarquia de um GLMM é explicitamente definida de modelos mais simples, que correspondem aos níveis de um conjunto de dados agrupados ou longitudinais, estes passam a serem chamados de HGLM ou modelos multiníveis (MLM), ou seja, modelos hierárquicos descrevem observações que possuem uma natureza aninhada: as unidades de um nível estão visíveis e contidas nas unidades de outro nível. Esses modelos geralmente tratam os termos das unidades como efeitos aleatórios, em vez de efeitos fixos, especialmente quando essas unidades são consideradas uma amostra de uma população de interesse. O modelo multinível contém termos de efeitos aleatórios para os diferentes níveis. (AGRESTI, 2019; WEST; WELCH; GALECKI, 2014)

HGLM consegue captar alteração nos dados das variáveis explicativas entre indivíduos de um determinado nível que permanecem inalteradas para certos grupos de indivíduos que representam um grupo superior. Isso fica fácil de identificar na Figura 15 onde é possível identificar indivíduos de 1 a *n*, aninhados em grupos de 1 a *J*. Já a Tabela 1 é a estrutura da base dados com dados aninhados em dois níveis. Onde, Y_{ij} = Observação *j* no grupo *i*, X_{Qij} = Variáveis explicativas de nível 1 para cada indivíduo *i* em cada grupo *j* e W_{ij} = Variáveis explicativas de nível 2 para cada grupo *j*. (FAVERO; BELFIORE, 2015)

Figura 15 – Estrutura Aninhada de dados agrupados em dois níveis.



Fonte: Adaptado de Fávero e Belfiore (2015).

Tabela 1 – Modelo geral de uma base de dados com estrutura aninhada em 2 níveis.

Observação (indivíduo i) Nível 1	Grupo j nível 2	Y_{ij}	X_{1ij}	X_{2ij}	...	X_{Qij}	W_{1j}	W_{2j}	...	W_{Sj}
1	1	Y_{11}	X_{111}	X_{211}		X_{Q11}	W_{11}	W_{21}		W_{S1}
2	1	Y_{21}	X_{121}	X_{221}		X_{Q21}	W_{11}	W_{21}		W_{S1}
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮		⋮
n_1	2	$Y_{n_1 1}$	$X_{1n_1 1}$	$X_{2n_1 1}$		$X_{Qn_1 1}$	W_{11}	W_{21}		W_{S1}
$n_1 + 1$	2	$Y_{n_1+1,2}$	$X_{1n_1+1,2}$	$X_{2n_1+1,2}$		$X_{Qn_1+1,2}$	W_{12}	W_{22}		W_{S2}
$n_1 + 2$	2	$Y_{n_1+2,2}$	$X_{1n_1+2,2}$	$X_{2n_1+2,2}$		$X_{Qn_1+2,2}$	W_{12}	W_{22}		W_{S2}
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮		⋮
n_2	2	$Y_{n_2 2}$	$X_{1n_2 2}$	$X_{2n_2 2}$		$X_{Qn_2 2}$	W_{12}	W_{22}		W_{S2}
⋮	⋮	⋮	⋮	⋮	...	⋮	⋮	⋮	...	⋮
$n_{J-1} + 1$	J	$Y_{n_{J-1}+1J}$	$X_{1n_{J-1}+1J}$	$X_{2n_{J-1}+1J}$		$X_{Qn_{j-1}+1J}$	W_{1J}	W_{2J}		W_{SJ}
$n_{J-1} + 2$	J	$Y_{n_{J-1}+2J}$	$X_{1n_{J-1}+2J}$	$X_{2n_{J-1}+2J}$		$X_{Qn_{j-1}+2J}$	W_{1J}	W_{2J}		W_{SJ}
⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮		⋮
n	J	Y_{nJ}	X_{1nJ}	X_{2nJ}		X_{QnJ}	W_{1J}	W_{2J}		W_{SJ}

Fonte: Fávero e Belfiore (2015).

3.4.1.1 Regressão Logística Multinível

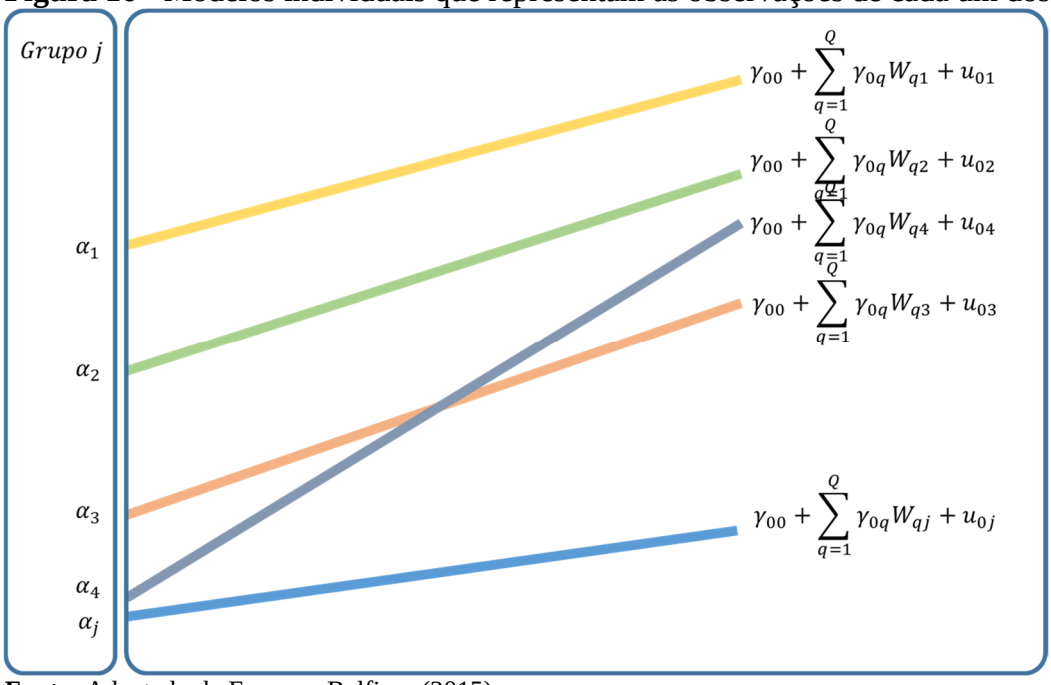
Regressão Logística Multinível (MLR) ou Regressão Logística Hierárquica (HLR) nada mais é do que um HGLM em que a função de ligação canônica $g(\mu)$ é não linear e expressa por $g(\mu)$ =log[*odds*], onde *odds* = μ/1 – μ (probabilidade de ocorrência ou não do evento μ), também chamada de função *logit link* e seu componente aleatório segue uma distribuição de probabilidade do tipo binomial. Assim a equação 20 pode ser reescrita pela equação 13 para classificação binária (AGRESTI, 2019).

$$logit[P(Y_{ijt} = 1)] = u_{ij} + \alpha_j + \beta_1.X_{1it} + \beta_2.X_{2it} + \cdots + \beta_{kijt}.X_{kijt}, i = 1, \ldots, T \qquad (20)$$

Onde,
Y_{ijt}, Resposta de ocorrência do evento (1 = sim, 0 = Não), para qualquer indivíduo *i* no grupo *j*.
u_{ij}, Efeito aleatório de um indivíduo *i* em um grupo *j*.
X_{kijt}, Valor de uma variável exploratória T para a observação em que: k=1,2,3...T).
α_j, Intercepto de cada grupo *j*.

Esse modelo consegue explicar a aleatoriedade de indivíduos *i* em cada grupo *j* representado por u_{ij}, mas não a influência dos grupos *j* nas aleatoriedades dos indivíduos *i*, em outras palavras, não conseguem explicar a interferência desses efeitos aleatórios de segundo nível nos parâmetros da equação canônica do modelo. Análogo a HGLM mais simples de regressão com componente aleatório normal é possível observar na Figura 16 diferentes interceptos e inclinações para cada grupo de um modelo HGLM. (FAVERO; BELFIORE, 2015)

Figura 16 - Modelos individuais que representam as observações de cada um dos J grupos.



Fonte: Adaptado de Favero e Belfiore (2015).

Por tanto, presume-se que cada intercepto de grupo de uma MLR tenha sua própria equação canônica, com diferentes parâmetros que devem ser incorporados no modelo para considerar os efeitos aleatórios que os grupos *j* (nível 2) exercem sobre os indivíduos *i* (nível 1), ou seja tanto os efeitos aleatórios nos interceptos quanto nas inclinações do modelo. Com isso, chega-se a um modelo padrão de MLR de acordo com a equação 21. (FAVERO; BELFIORE, 2015; AGRESTI, 2019)

$$logit[P(Y_{ijt} = 1)] = u_{ij} + s_j + \alpha + \beta_1.X_{1it} + \beta_2.X_{2it} + \cdots + \beta_{kijt}.X_{kijt}, i = 1, \ldots, T. \qquad (21)$$

Esse é um modelo multinível com interceptos aleatórios u_{ij} no nível de indivíduos e sj no nível dos grupos.

4 PLANO DE TRABALHO

4.1 Atividades Propostas

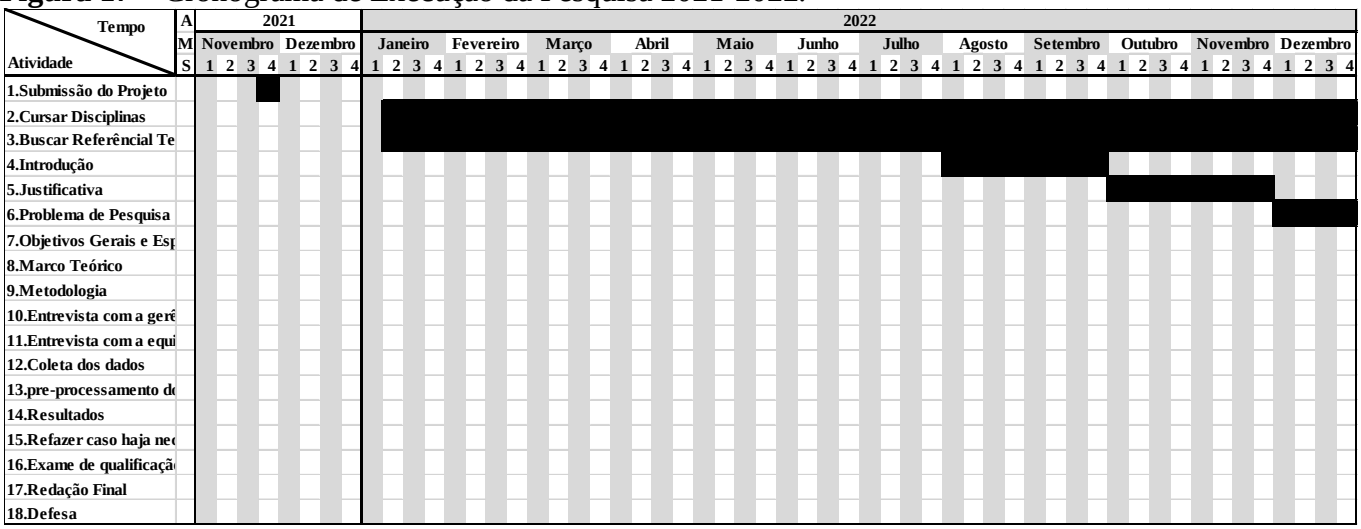
As seguintes atividades foram definidas no projeto para a conclusão da pesquisa:

1. Submissão do Projeto
2. Cursar Disciplinas
3. Buscar Referêncial Teórico
4. Introdução
5. Justificativa
6. Problema de Pesquisa
7. Objetivos Gerais e Específicos
8. Marco Teórico
9. Metodologia
10. Entrevista com a gerência
11. Entrevista com a equipe técnica
12. Coleta dos dados
13. Desenvolvimento
14. Resultados
15. Refazer caso haja necessidade
16. Exame de qualificação
17. Redação Final
18. Defesa

4.2 Cronograma das Atividades Propostas

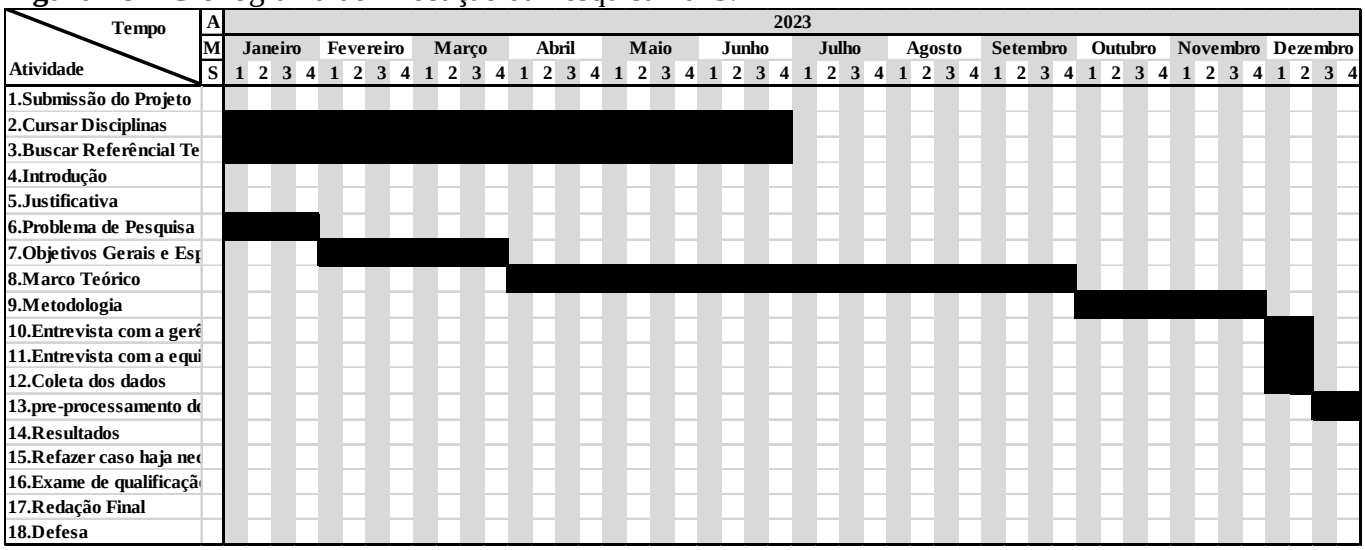
Com base nas atividades descritas anteriormente foi elaborado um cronograma programando a duração de cada atividade no tempo disposto nas figuras 16, 17,18 e 19, onde “A” Significa “Ano” “M”, ”Mês” e “S”, “Semana”.

Figura 17 – Cronograma de Execução da Pesquisa 2021-2022.



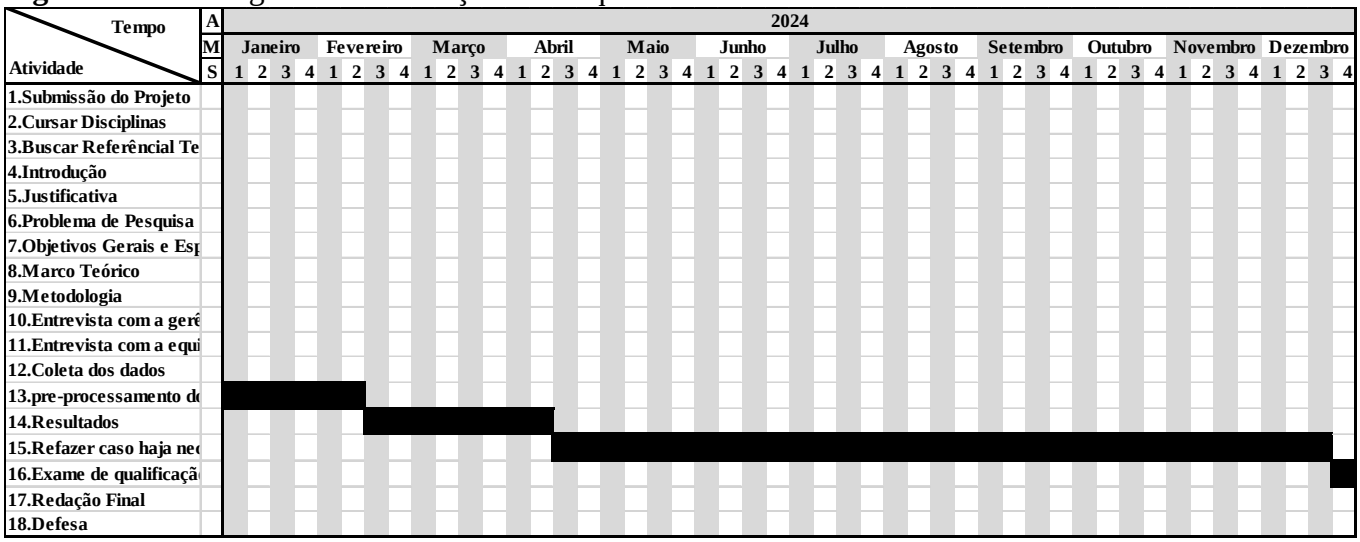
Fonte: O autor (2021).

Figura 18 – Cronograma de Execução da Pesquisa 2023.



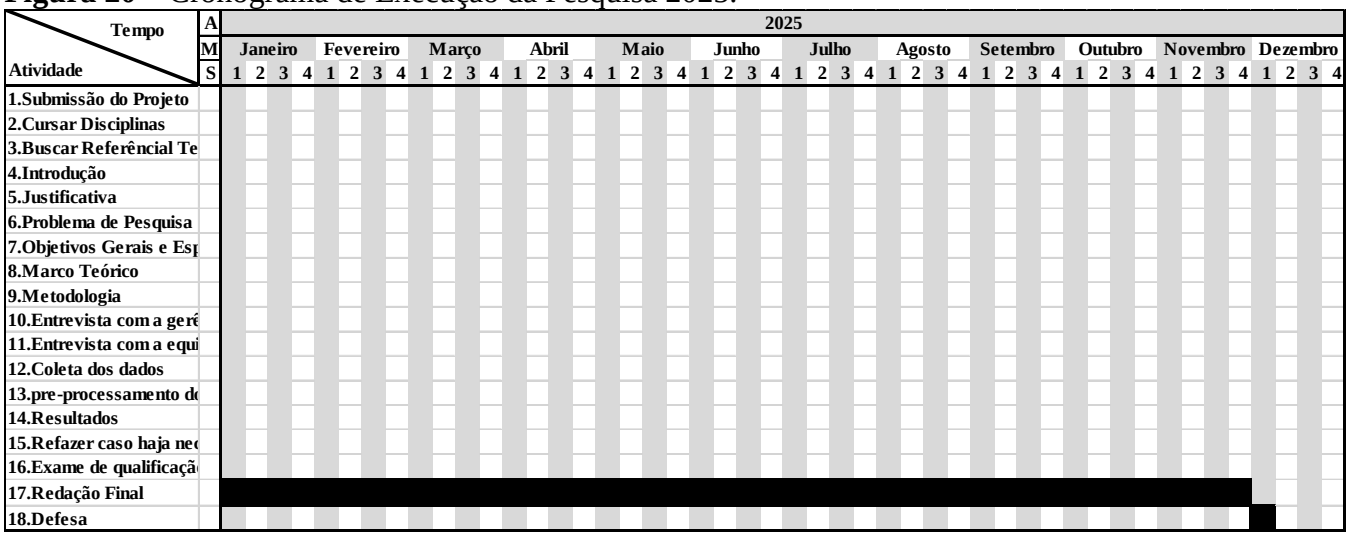
Fonte: O autor (2021).

Figura 19 – Cronograma de Execução da Pesquisa 2024.



Fonte: O autor (2021).

Figura 20 – Cronograma de Execução da Pesquisa 2025.



Fonte: O autor (2021).

Primeiramente é possível observar que na última semana de novembro do ano de 2021 será entregue este projeto de pesquisa para avaliação. Sendo aprovado as disciplinas serão cursadas em até um ano e meio iniciando em janeiro de 2021 e concluídas até junho de 2022. A busca de referencial teórico será feita em paralelo.

Em seguida uma repaginação nas partes introdutórias bem como o problema de pesquisa, objetivos e marco teórico serão concluídos até o final de outubro de 2023, para que em novembro inicie-se a metodologia e toda a parte coleta de dados e desenvolvimento. Para a coleta, as entrevistas serão marcadas nas primeiras semanas de dezembro, assim a análise e processamento dos dados será feita na metade de dezembro até final de janeiro de 2024. E com isso sejam processados os resultados dos modelos entres os meses de fevereiro, março e abril de 2024.

Os meses de maio até dezembro de 2024 ficarão em aberto caso haja necessidade de algum reprocessamento ou reanálise dos modelos, deixando aberto uma possível nova entrevista e contato novamente com a empresa. E em dezembro de 2024 estará marcado o exame de qualificação.

Por fim o ano de 2025 será reservado para a redação final de acordo com as correções e direcionamentos apontados pela banca examinadora do exame de qualificação. E somente em dezembro com tudo ajustado será apresentado e enviado a banca examinadora a versão final da pesquisa será a defesa dela.

5 REFERÊNCIAS

ADWAN O.; FARIS H.; JARADAT K.; HARFOUSHI O.; GHATASHEH N. **Predicting Customer Churn in Telecom Industry using Multilayer Preceptron Neural Networks: Modeling and Analysis**, Life Science Journal, v. 11, p. 75-81, 2014.

ALLISON P. D. **Logistic Regression Using SAS Theory and Application**, SAS publishing, 2012.

AGGARWAL C. C. **Data Mining**, The Textbook. Springer, 2015.

AGRESTI A. **An Introduction to Categorical Data Analysis**, 3 ed. Winley, 2019.

BHUSE P.; GANDHI A.; MESWANI P.; MUNI R.; KATRE N. **Machine Learning Based Telecom-Customer Churn Prediction**. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020.

BISHOP C. M. **Neural Networks for Pattern Recognition**, Oxford Press, 1996.

BREIMAN L. **Random Forests,** Machine Learning, v. 45, n. 1, 2001.

BRESLOW N. E.; CLAYTON D. G. **Approximate Inference in Generalized Linear Mixed Models.** Journal of the American Statistical Association, v. 88, n. 421, p. 9–25, 1993.

BRAGA A. P.; CARVALHO A. P. L. F.; LURDEMIR T. B. **Redes Neurais Artificiais - Teoria e Aplicações**, LTC, 2007.

CALCAGNO V.; MAZANCOURT C. **glmulti: an R package for easy automated model selection with (generalized) linear models**. Journal of Statistical Software, v. 34, n. 12 p. 1-29, 2010.

CHAPMAN P.; CLINTON J.; KHABAZA T.; REINARTZ T.; WIRTH R. **The CRISP-DM process model.** The CRIP–DM Consortium. 1999.

CHEN Z. Y.; FAN Z. P.; SUN M. A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. **European Journal of Operational Research,** v. 223, n. 2, p. 461-472, 2012.

CORTES C.; VAPNIK V. **Support-vector networks**, Machine Learning, v. 20, n. 3, 1996.

COUSSEMENT K.; VAN DEN POEL D. **Churn prediction in subscription services:** An application of support vector machines while comparing two parameter-selection techniques, Expert Systems with Applications, v. 34, n. 1, p. 313-327, 2008.

FAUSETT L. Fundamentals of Neural Networks: Architetures, Algorithms, and Applications. Pearson, 1994.

FAVERO L. P. L.; BELFIORE P. P. **Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata**. Elsevier, 2017.

FRIEDMAN J.; HASTIE T.; TIBSHIRANI R. **Additive logistic regression:** a statistical view of boosting. The Annals of Statistics, v.28, n. 2, p. 337-407, 2000.

HARRELL, F. E. **Regression modeling strategies:** with applications to linear models, logistic regression, and survival analysis, v. 608, Springer, 2001.

HAYKING S. H. **Neural Networks and Learning Machines**, Prentice Hall, 2008.

HEDEKER D.; GIBBONS R. D. **Longitudinal data analysis**, v. 451. John Wiley & Sons, 2006.

HILLS W.; DANIEL W.; LU M. Y.; SCHAER O.; ADAMS S. Modeling Client Churn for Small Business-to-Business Firms. **Systems and Information Engineering Design Symposium.** p. 1-7, 2020.

HECKERMAN D. **Bayesian Networks for Data Mining**. Data Mining and Knowledge Discovery v. 1 n. 1, 1997.

IWATA M.; OTAKE K.; NAMATAME T. Analysis of the Characteristics of Customer Defection on a Hair Salon Considering Individual Differences. **Social Computing and Social Media. Communication and Social Communities,** v.2, p. 378-391, 2019.

JAIN H.; KHUNTETA A.; SRIVASTAVA S. **Telecom churn prediction and used techniques, datasets and performance measures: a review**, Telecommunication Systems, vol. 78, n. 4, p. 613-630, 2021.

JAYASWAL P.; PRASAD B. R.; TOMAR D.; AGARWAL S. **An Ensemble Approach for Efficient Churn Prediction in Telecom Industry.** International Journal of Database Theory and Application, v. 9, n. 8, p. 211-232, 2016.

JESKE D. R.; LI, J.; WONG V. On the effectiveness of Mixed Model Based Logistic Classifiers for Longitudional Data. **Integration: Mathematical Theory and Applications**, v.3, n.3, p. 233, 2012.

KORN E. L.; WHITTEMORE A. S. **Methods for analyzing panel studies of acute health effects of air pollution.** Biometrics, v. 35, n.4, p.795-802, 1979.

LALWANI P.; MISHRA M. K.; CHADHA J. S.; SETHI P. **Customer churn prediction system:** a machine learning approach. Computing, 2021.

LAZAROV V.; CAPOTA M. **Churn prediction**. Bus. Anal. Course. TUM Comput. Sci, v. 33, p. 34, 2007.

LEE Y.; NELDER J. A.; PAWITAN Y. **Generalized Linear Models with Random Effects**, Chapman e Hall/CRC, 2006.

MARISCAL G.; MARBAN O.; FERNANDEZ C. **A survey of data mining and knowledge discovery process models and methodologies.** The Knowledge Engineering Review, v. 25, n. 2, p. 137-166, 2010.

MAHAJAN V.; MISRA R.; MAHAJAN R. **Review of Data Mining Techniques for Churn Prediction in Telecom**, Journal of Information and Organizational Sciences, v. 39, n. 2, p. 183-197, 2015.

MIGON H. S.; GAMERMAN D.; LOUZADA F. **Statistical Inference.** An Integrated Approach. ed. 2, CRC PRESS, 2015.

MOREIRA J.; CARVALHO A. C. P. F.; HORVÁTH T. **A General Introduction to Data Analytics**. John Wiley & Sons, 2019.

NELDER J. A.; WEDDERBURN R. W. M. **Generalized Linear Models**. Journal of the Royal Statistical Society: Series A, 1972.

PAWLAK Z. **Rough Sets**. International Journal of Computer and Information Sciences, v. 11, n. 5, 1892.

PAWLAK Z., **Rough sets and intelligent data analysis**. Information Sciences, v. 147, n.1, 2002.

JAYASWAL P.; PRASAD B. R.; TOMAR D.; AGARWAL S. **An Ensemble Approach for Efficient Churn Prediction in Telecom Industry.** International Journal of Database Theory and Application, v. 9, n. 8, p. 211-232, 2016.

PRASHANTH R.; DEEPAK K.; MEHER A. K. High accuracy predictive modelling for customer churn prediction in telecom industry**. In International Conference on Machine Learning and Data Mining in Pattern Recognition,** p. 391-402, Springer, 2017.

PLUMED *et al.* **CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories**, IEEE Transactions on Knowledge and Data Engineering, v. 33, n. 8, p. 3048-3061, 2021.

PFEIFER P. E.; FARRIS P. W. The elasticity of customer value to retention: The duration of a customer relationship. **Journal of Interactive Marketing,** vol. 18, n. 2, p. 20-31, 2004.

RANGASWAMY A.; BRUGGEN G. H. V. Opportunities and challenges in multichannel marketing: An introduction to the special issue. **Journal of Interactive Marketing,** v. 19, n. 2, p. 5-11, 2005.

SHARMA T.; GUPTA P., NIGAM V., GOEL M. Customer Churn Prediction in Telecommunications Using Gradient Boosted Trees. **International Conference on Innovative Computing and Communications**. Advances in Intelligent Systems and Computing, v. 1059, Springer, Singapore, 2020.

STIRATELLI R.; LAIRD N.; WARE J. H. **Random-Effects Models for Serial Observations with Binary Response.** Biometrics, v.40, n.4, p.961–971, 1984.

STRECHT P.; MOREIRA J. M.; SOARES C. **INMPLODE:** A framework to interpret multiple related rule-based models. Expert Systems, v. 38, n. 6, 2021

TANG P. **Telecom Customer Churn Prediction Model Combining K-means and XGBoost Algorithm**, 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2020.

VENABLES W. N.; RIPLEY B. D., Modern Applied Statistics with S, **Statistics and Computing**, ed. 4. Springer, 2002.

WEISS, R. E. **Modeling longitudinal data, v**. 1, New York, Springer, 2005

WEST B.T.; WELCH K. B.; GALECKI A. T. **Linear Mixed Models:** A Practical Guide Using Statistical Software, 2ª ed. Chapman and Hall/CRC, 2014.

WACLAWIW M. A.; LIANG K. Y. **Prediction of Random Effects in the Generalized Linear Model.** Journal of the American Statistical Association, v. 88, n. 421, p.171–178, 1993.

WALKER S.H.; DUNCAN D. B. **Estimation of the probability of an event as a function of several independent variables**, Biometrika, v. 54, n. 1-2, p. 167–179, 1967.

WOLFINGER R.; O'CONNELL M. **Generalized linear mixed models a pseudo-likelihood approach.** Journal of Statistical Computation and Simulation, v. 48, n. 3-4, p. 233-243, 1993.

ZHANG Q.; XIE Q.; WANG G. **A survey on rough set theory and its applications**. CAAI Transactions on Intelligence Technology, v. 1, n. 4, p. 323-333, 2016.

ZHAO Y.; LI B.; LI X.; LIU W.; REN S. Customer Churn Prediction Using Improved One Class Support Vector Machine. **Advanced Data Mining and Applications,** p. 300-306**,** 2005.

ZIARKO W. **Variable Precision Rough Set Model**, Journal of Computer and System Sciences, v. 46, p. 39-59, 1993.