**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

Alexandre Fábio de L.S Sales

**Modelo Preditivo para Aprovação de Cartão de Crédito**

**Utilizando Algoritmos Classificadores.**

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Rio de Janeiro

2023

**SUMÁRIO**

1. **Introdução .......................................................................................................** 
   1. **Contextualização .....................................................................................**
   2. **O problema proposto**
2. **Coleta de Dados**
3. **Processamento/Tratamento de Dados**
   1. **Outliers**
4. **Análise e Exploração dos Dados**
5. **Criação de Modelos de Machine Learning**
6. **Apresentação dos Resultados**
7. **Links**

**REFERÊNCIAS**

**1. Introdução**

* 1. **Contextualização**

**Cartão de crédito** é um dos meios de pagamentos mais populares no Brasil e no Mundo. Segundo a ABECS (Associação Brasileira das Empresas de Cartões de Crédito e Serviços), no Brasil, **as transações em cartões foram em torno de R$ 3,31 trilhões no ano de** 2022**.**

Os primeiros cartões de crédito surgiram nos Estados Unidos na década de 1920 por donos de empreendimentos - hotéis, postos de gasolina, lojas de departamentos - que os davam aos clientes mais fiéis e abonados. Uma espécie de cartão de fidelidade.

Somente em 1949, o esquecimento da carteira para pagar a conta de um jantar de negócios, inspirou a ideia de cartão para restaurantes por um executivo de uma corporação de crédito - Frank Macnamara. Em 1950 foi lançado “The Diners Club” (Clube da Janta), considerado o primeiro cartão de crédito universal. Assim, surgiu um novo conceito em vez das empresas oferecerem crédito, haveria um intermediário, sendo aceito em vários estabelecimentos. No início o cartão era de papel, aceito apenas em restaurantes e pessoas influentes. As empresas que aceitavam aderir ao modelo pagavam um percentual por transação, e os assinantes uma taxa anual.

O desenvolvimento foi rápido, no 3º aniversário existiam 42 mil membros e 330 estabelecimentos no EUA cadastrados. Os associados pagavam US$3 ano, e os estabelecimentos 7%. No Brasil chegou em 1954.

Várias outras empresas adotaram o novo modelo, e assim foi criada a indústria de cartões de crédito. Algumas bandeiras mais conhecidas do mercado, se originaram a partir do final da década de 50, por exemplo:

1958 – American Express Card

1958 – BankAmericard que se tornou VISA (1976-77)

1966 – Intercard Bank que se tornou Mastercard (1979)

Com o tempo os cartões se popularizaram. Surgem novos formatos com a evolução tecnológica, por exemplo, incorporação de chips, senhas e aproximação. Interessante verificar que ainda tem muito espaço para o crescimento. Segundo o site Statista e Banco Mundial a penetração varia com o país, função da cultura ou necessidade de investimentos das organizações. A média mundial para a população acima de 15 anos é de apenas 10%.

* O Canadá foi um dos três países do mundo em 2021, onde a posse de cartão de crédito entre consumidores com 15 anos ou mais era superior a 70%.
* No Brasil a estimativa é que em torno de 40.43% da população com 15 anos ou mais, possuía um cartão de crédito.
  1. **O problema proposto**

No mundo de hoje, onde as transações online se tornaram a norma, a fraude com cartões de crédito tornou-se uma grande preocupação para bancos e instituições financeiras. É essencial que os bancos prevejam se um cliente apresenta um risco de crédito para minimizar as perdas financeiras por inadimplência ou fraude.

Um cartão de crédito é emitido por um banco ou empresa de serviços financeiros que permite aos titulares do cartão emprestar fundos para pagar bens e serviços com comerciantes que aceitam cartões como forma de pagamento. Os cartões de crédito impõem a condição de que os titulares do cartão devolvam o dinheiro emprestado, acrescido de quaisquer juros aplicáveis, bem como quaisquer encargos adicionais acordados, integralmente até a data de cobrança ou ao longo do tempo.

O problema proposto consiste na construção de um algoritmo para auxiliar na tomada de decisão do cálculo de novos clientes de cartão de crédito. O trabalho não deve ser visto como a construção de uma metodologia para definir a aquisição de novos clientes de cartão de crédito, mas sim uma ferramenta que apoia a decisão baseada em dados históricos.

Temos como objetivos dessa análise:

• Realizar a análise descritiva dos dados dos clientes de cartão de crédito;

• Verificar a correlação entre eles;

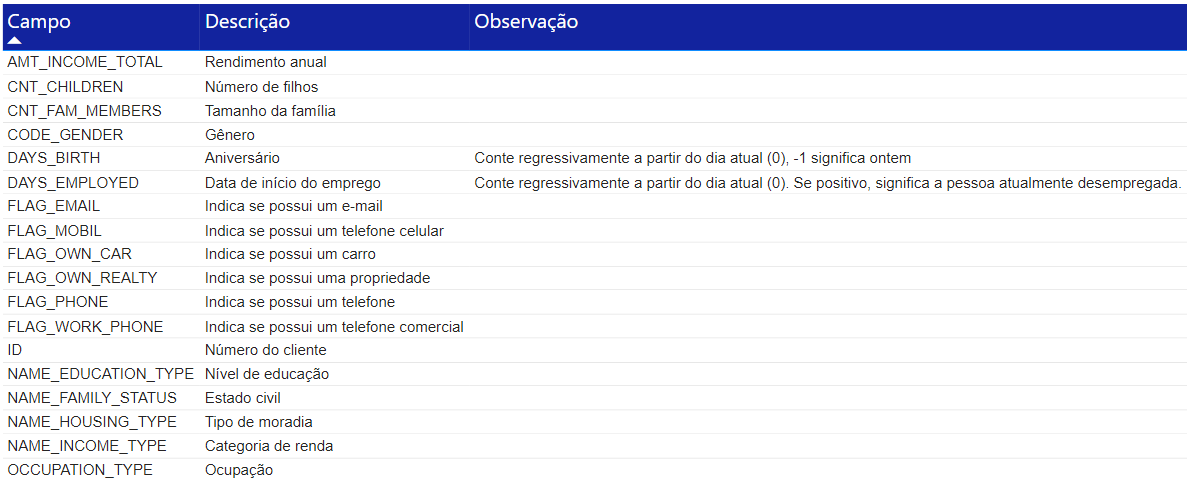
• Criar modelos preditivos para clientes que apresentem um risco de inadimplência ao adquirir um cartão de crédito, utilizando os algoritmos de classificação Árvore de Decisão, Regressão Logística, Naïve Bayes, Gradiente Descendente, KNN (K - Nearest Neighbors) e Randon Forest.

Os dados utilizados neste trabalho contém transações efetuadas com cartões de crédito, para análise exploratória, extraídos do site <https://www.kaggle.com/datasets/> .

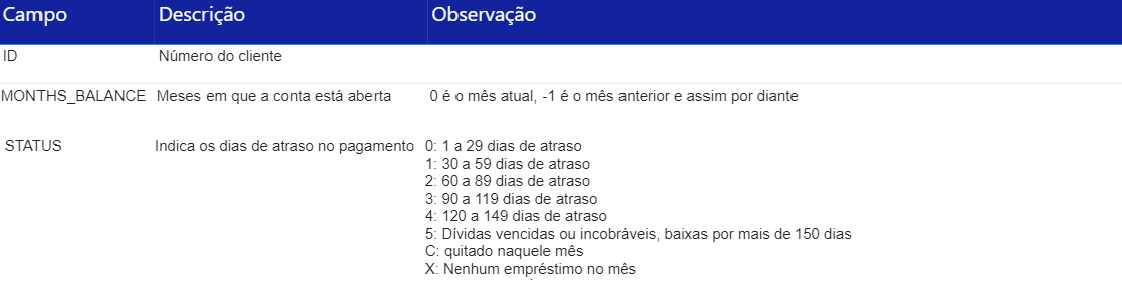
**2. Coleta de Dados**

Para o tratamento do problema proposto, foram utilizados dois *datasets* (conjunto de dados) relativos às transações de cartões de crédito, extraídos do site <https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction/data>.

O primeiro dataset, "application\_record.csv", traz as informações cadastrais dos clientes, que podem ser utilizadas como recursos para a previsão. Esse dataset possui os seguintes campos:



O segundo dataset, "credit\_record.csv", registra o comportamento dos clientes de cartão de crédito. Esse dataset possui os seguintes campos:



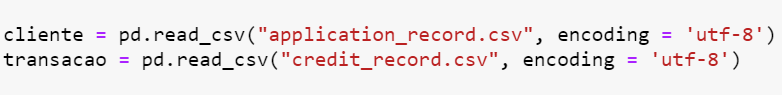
**3. Processamento/Tratamento de Dados**

O processamento e o tratamento dos dados foram feitos utilizando a linguagem Python, versão 3.8.17, no ambiente Jupyter Notebook, versão 6.5.4. Dentro da linguagem, utilizou-se a biblioteca “pandas” que é uma poderosa ferramenta para tratamento e análise de dados.

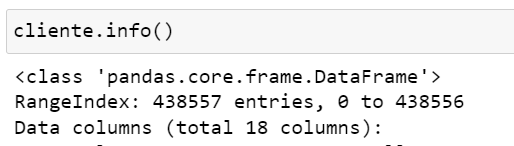
Inicialmente foi necessário importar a biblioteca “*Pandas*” (Figura 8), que é uma biblioteca de código aberto de licença BSD (em inglês, Berkeley Software Distribution) que fornece estruturas de dados de alto desempenho e fáceis de usar, além de ferramentas de análise de dados para a linguagem de programação Python.

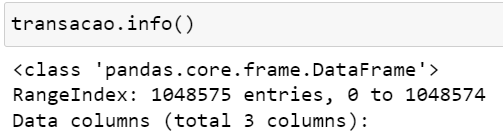


Em seguida, deve-se fazer a leitura e tratamento dos *datasets*, que, nesse caso, será feita individualmente para cada um deles. Os *datasets* “application\_record.csv” e “credit\_record.csv” serão processados e importados para as variáveis “cliente” e “transação” por meio do comando “pd.read.csv”, seguindo os parâmetros estabelecidos nos arquivos com as características dos *datasets*.



O uso do comando “pd.read.csv”, cria uma estrutura bidimensional de dados, como uma planilha, chamada daframe. A utilização da função “info ()”, apresenta informações sobre os dataframes criados.

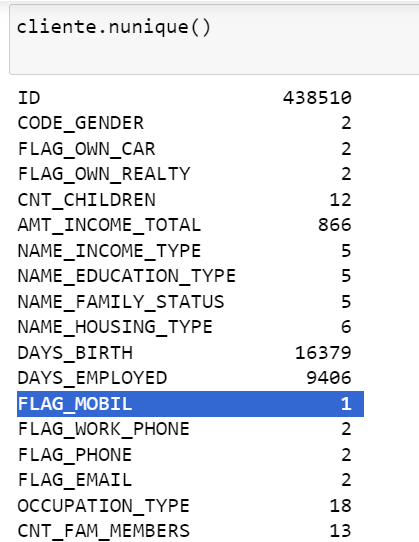


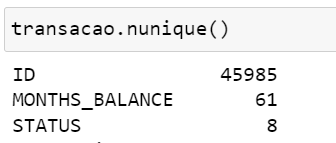


Os dataframes criados apresentam as seguintes dimensões:

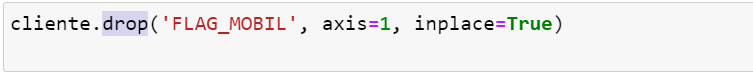
* Dataframe “cliente” - 438557 linhas, divididas em 18 colunas.
* Dataframe “transacao” - 1048575 linhas, divididas em 3 colunas.

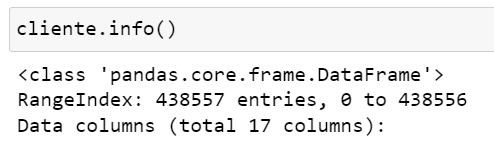
A coluna FLAG\_MOBIL não foi mantida nesse estudo, pelo fato de que essa coluna é uma constante e por isso foi desconsiderada. O dataframe “transacao” não apresenta colunas com valores contantes.



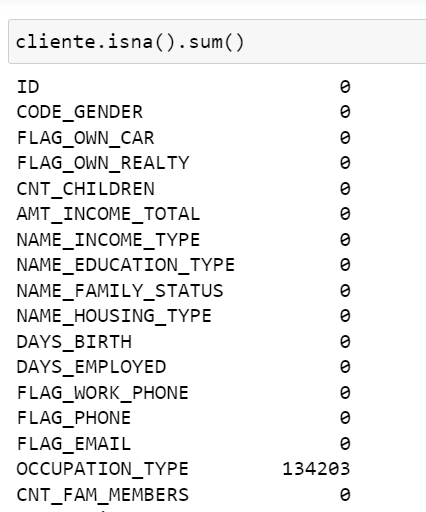


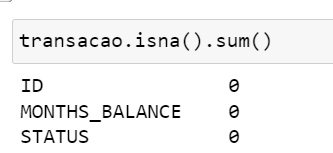
Para remover a coluna FLAG\_MOBIL, foi utilizado o seguinte código:





O dataframe “cliente” apresenta agora 17 colunas com as 438557 entradas filtradas anteriormente. Para verificar se há dados nulos nos dataframes, utilizou-se a função “isnull()”, que faz essa análise, em conjunto com a função “sum()”, que, nesse caso, soma os valores encontrados na função anterior.

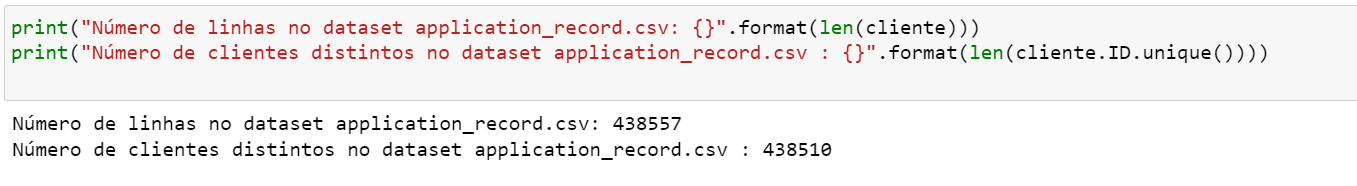




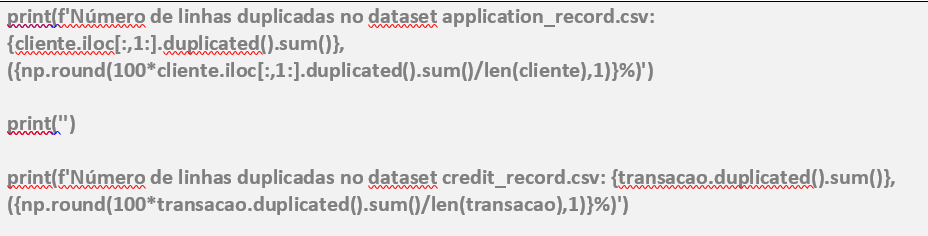
Após esse comando, verificou-se que o dataframe cliente, apresenta dados nulos na coluna OCCUPATION\_TYPE que necessitam de tratamento. Nesse caso, os dados nulos serão substituídos pelo valor 'Other'. Foi utulizada a função “fillna” para a substituição.

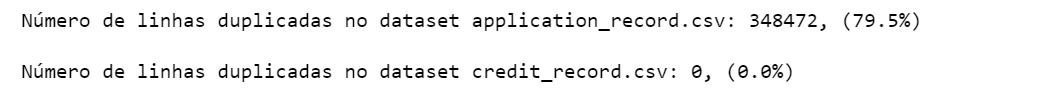


O número de clientes distintos não pode ser menor que o tamanho do dataframe application\_record.csv, e cada linha deve corresponder a um cliente único. A verificação de dados duplicados foi confirmada pelos método duplicated.



A função “cliente.iloc[:,1:].duplicated().sum()” irá determinar os ID’s duplicados no dataframe “cliente”. Iremos manter o primeiro ID encontrado, após a eliminação dos dados duplicados no dataframe “cliente”. O dataframe “transacao” não possui dados duplicados.





Para remoção dos registros duplicados foi utilizado o método “drop\_duplicates”.

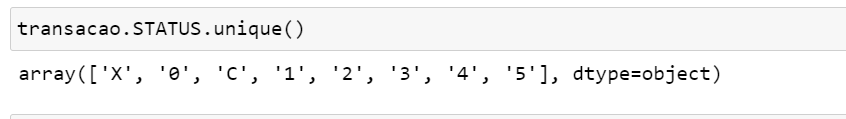


Para modelos de machine learning supervisionados é fundamental a etapa de criação de uma coluna target no dataset. Em problemas que isso é possível, faz com que os modelos e resultados obtidos possam ser explorados com maior facilidade.

Na maioria das situações, um algoritmo de aprendizado de máquina supervisionado é usado para derivar a variável target. Esse algoritmo usa dados históricos para aprender padrões e descobrir relacionamentos entre outras partes do seu conjunto de dados e a variável target.

O caso em análise busca avaliar a previsibilidade se o usuário é um cliente 'bom' ou 'ruim' com base em alguns parâmetros. Dessa forma, a informação que utilizaremos como resultado é a coluna STATUS, que traz os dias em atraso no pagamento.

As linhas que possuem target=1 representam os clientes que estão com no mínimo um mês de atraso, e será considerado um cliente "ruim".



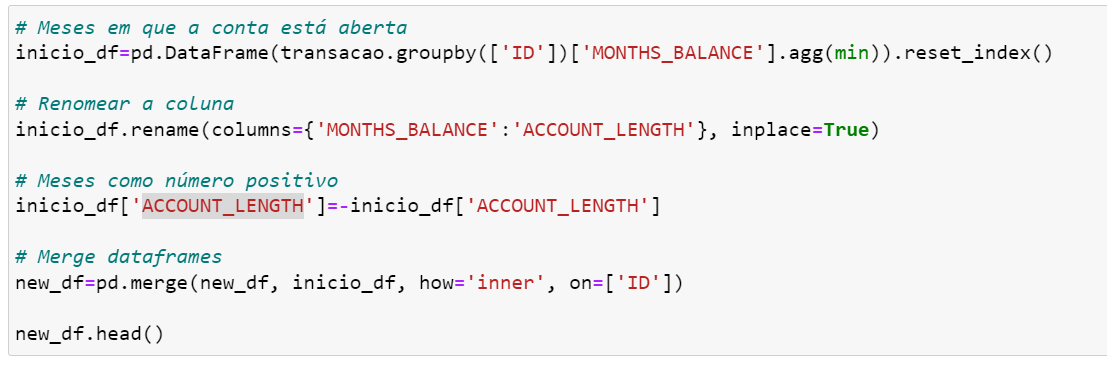


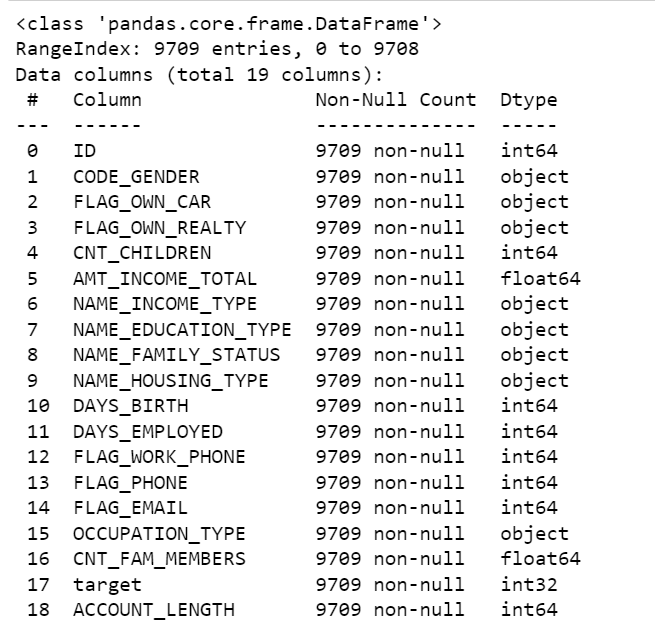
O objetivo agora é criar um único dataset consolidado (new\_df), gerado a partir da integração (join/merge) dos datasets gerados. Para isso, foi necessário agrupar o dataset “transacao”, por meio de seu ID de cliente, através dos comandos “groupby” e “agg(max)” que, em conjunto, selecionam o valor máximo de “target” e o ID para cada cliente.



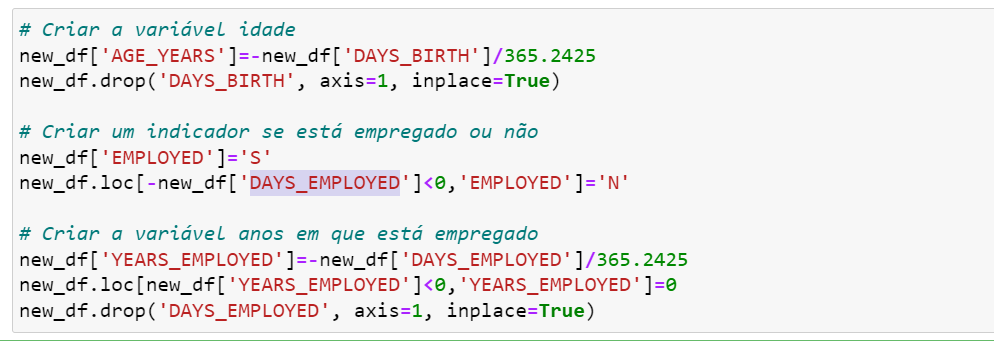
O próximo passo é incluir a coluna “MONTHS\_BALANCE”, que foi removida no merge anterior. Faz sentido que o número de meses em que a conta está aberta, seja correlacionado com o risco do cliente (uma vez que há mais oportunidades de perder pagamentos).

Como a coluna “MONTHS\_BALANCE” apresenta valores negativos, informando o número de meses em que a conta está aberta, foi selecionado o valor mínimo da coluna para cada ID. Após a seleção, a coluna será renomeada para “ACCOUNT\_LENGTH”, os valores transformados para positivo e incluídos no dataframe “new\_df”.

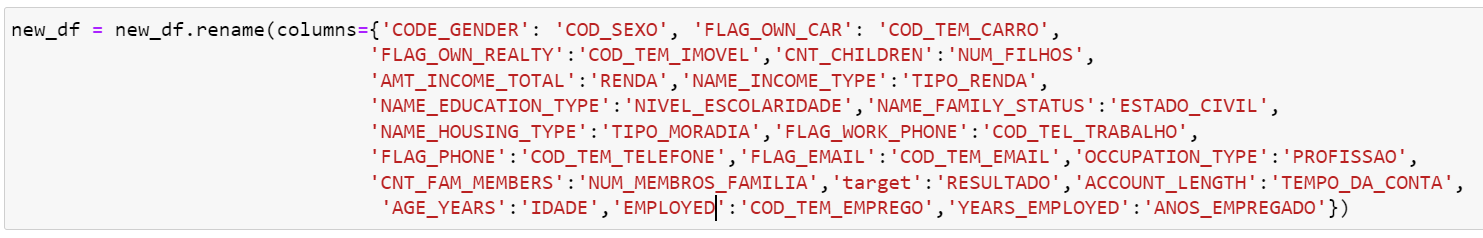




Percebe-se que o novo dataframe possui 9709 linhas. Agora iremos transformar as variáveis “DAYS\_BIRTH” e “DAYS\_EMPLOYED”, em variáveis que indiquem a idade, os anos de emprego e se o cliente está empregado. As antigas variáveis transformadas serão eliminadas do dataset.



Como o dataset apresenta os nomes das colunas na língua inglesa, a tradução para a língua portuguesa foi realizada para um melhor entendimento.

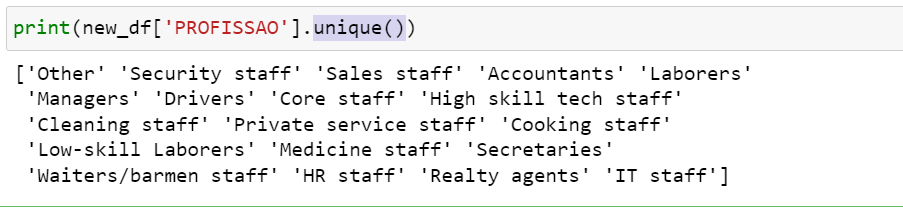


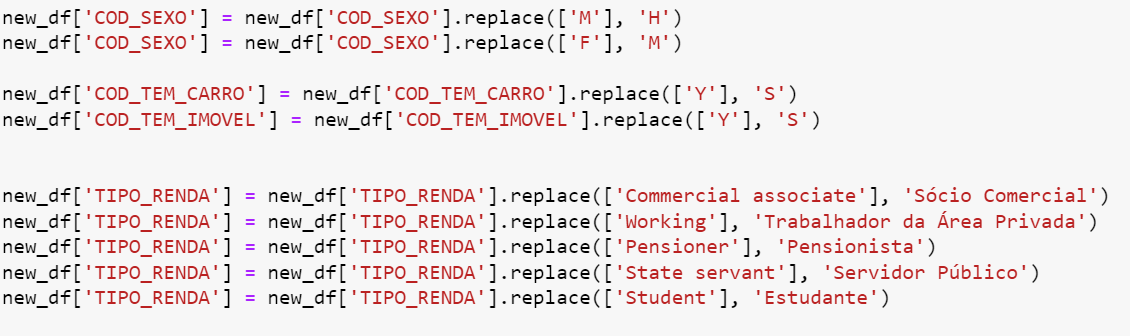
A transformação dos valores das colunas COD\_SEXO, COD\_TEM\_CARRO, COD\_TEM\_IMOVEL, TIPO\_RENDA, TIPO\_MORADIA, NIVEL\_ESCOLARIDA DE, ESTADO\_CIVIL e PROFISSAO do tipo “object” para a língua portu

guesa também foi realizada. A coluna NUM\_MEMBROS\_FAMILIA foi transfor

mada para o tipo int, correspondendo aos valores apresentados. O uso da fun

ção “unique()” apresenta os valores distintos das colunas que serão traduzidas para língua portuguesa.





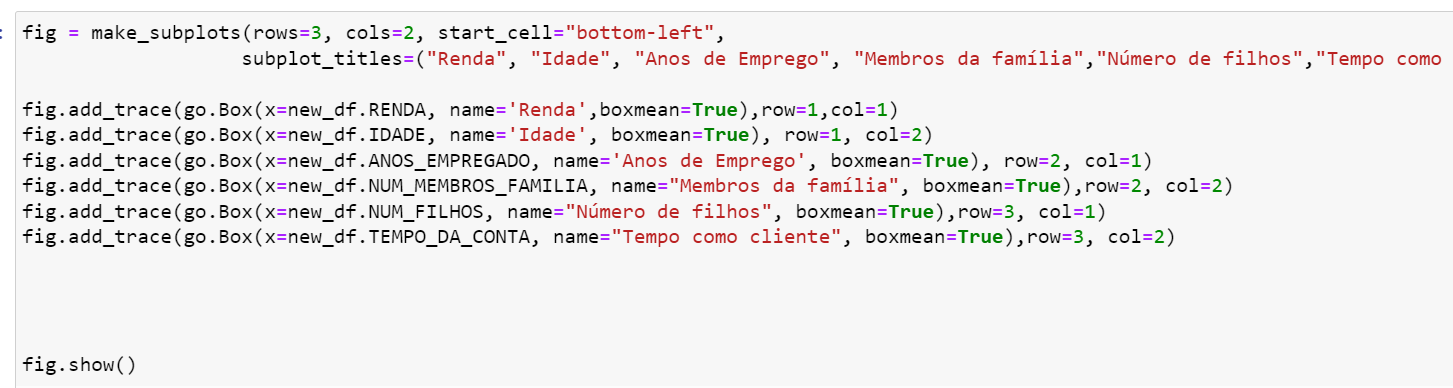


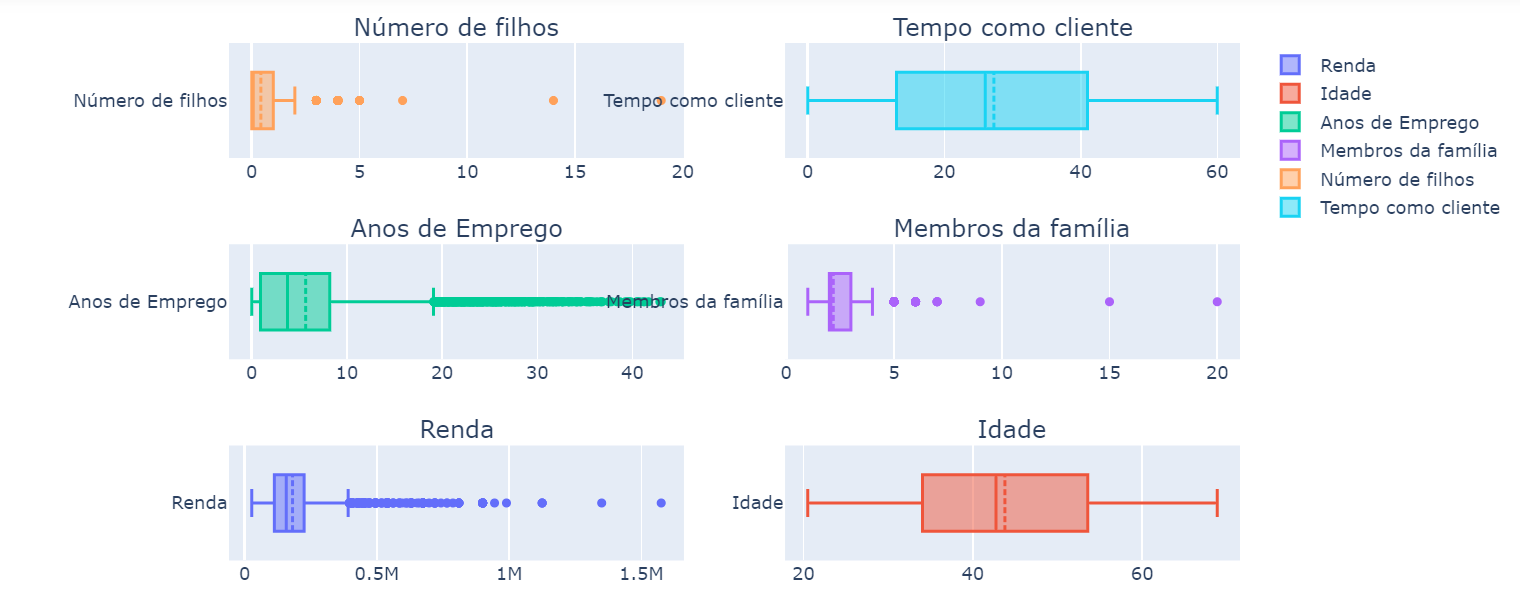
**3.1 Outliers**

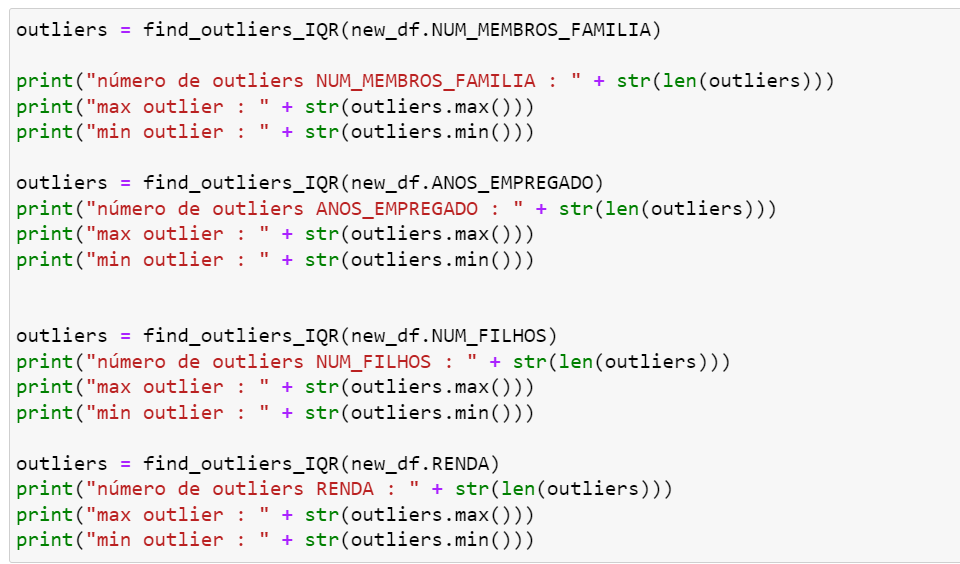
Um outlier é uma observação que está a uma distância anormal de outros valores em uma amostra aleatória de uma população. Eles podem aparecer devido a erros na entrada de dados ou medição, ou apenas porque há variação na população que você está observando. Seja qual for o motivo pelo qual aparecem, saber como identificar e lidar com valores discrepantes é uma parte importante da limpeza de dados.

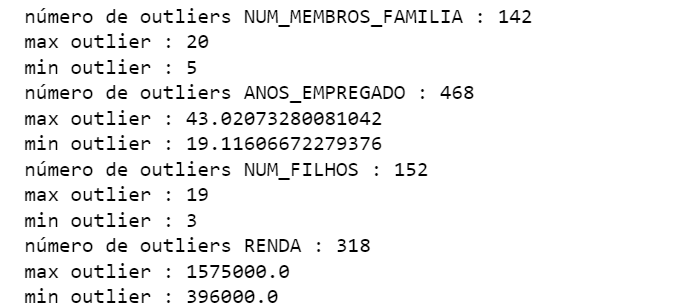
Os valores discrepantes serão removidos pelo método IQR, também chamado de "intervalo interquartil". Primeiro classificando os elementos em ordem crescente e depois encontrando Q1, Q3 e iqr pela diferença de Q3 e Q1. Finalmente, definindo o limite superior e inferior e removendo-os.

Uma das maneiras de ver se o conjunto de dados tem valores discrepantes é traçar os dados em gráfico. A função “go.Box” será utilizada para apresentar um gráfico de caixa que é uma representação estatística da distribuição de uma variável através de seus quartis. As extremidades da caixa representam os quartis inferior e superior, enquanto a mediana (segundo quartil) é marcada por uma linha dentro da caixa. As colunas que informam a renda, idade, anos de emprego, membros da família, número de filhos e tempo como cliente serão analisadas.

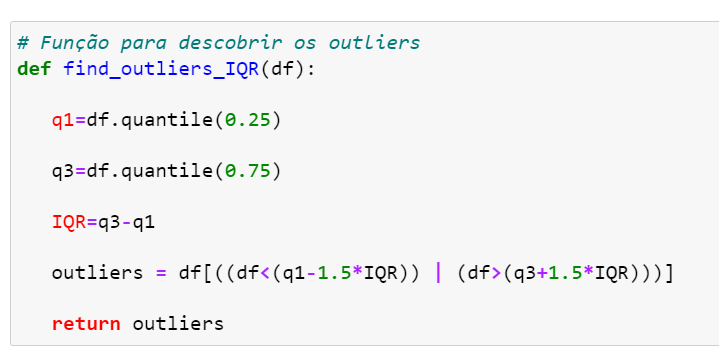


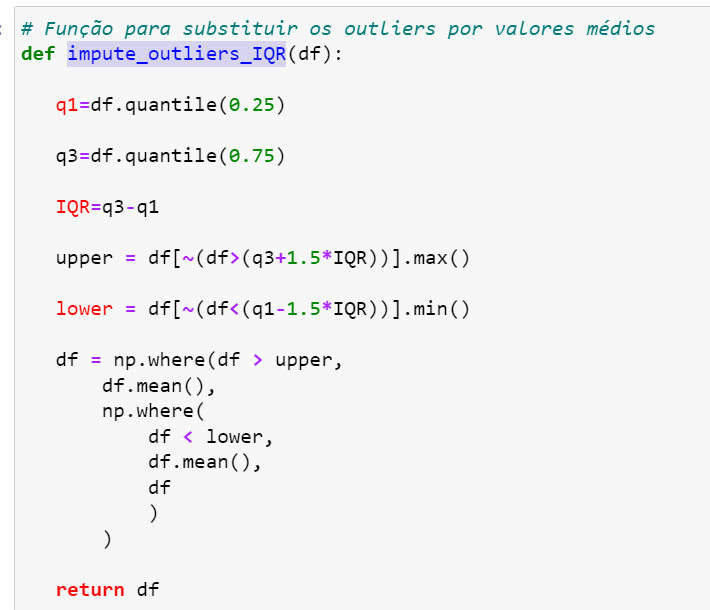


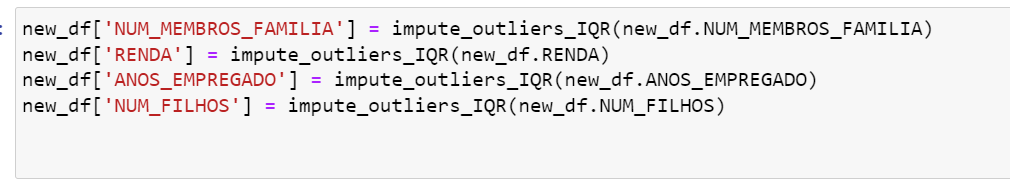




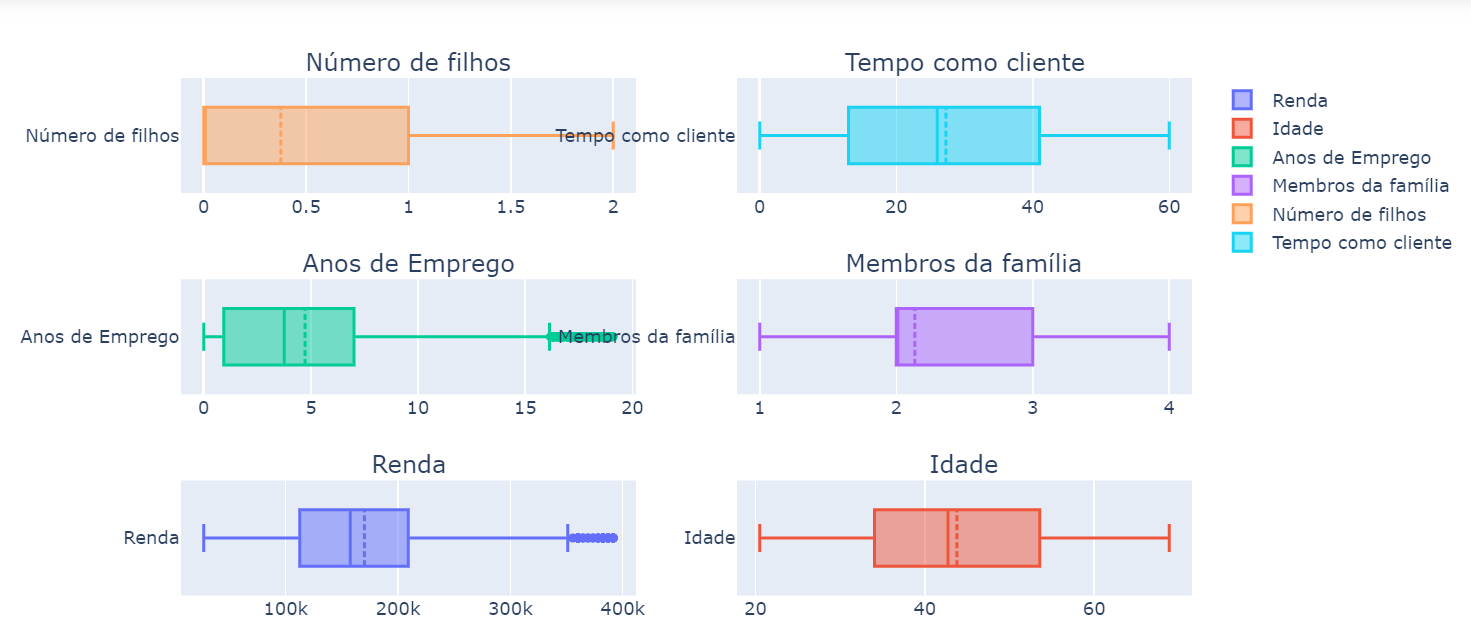
Pode-se observar que existem outliers para o número de filhos, anos de emprego, membros da família e renda. A função “find\_outliers\_IQR” foi utilizada para descobrir os outliers e a substituição dos outliers por valores médios, será feita pela função “impute\_outliers\_IQR”.





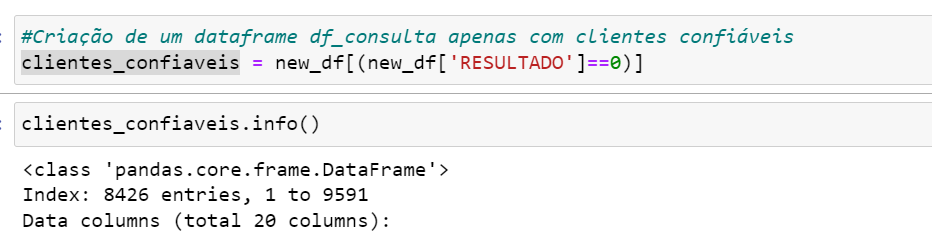


Após a substituição dos outliers, o gráfico de caixa não apresenta valores discrepantes.



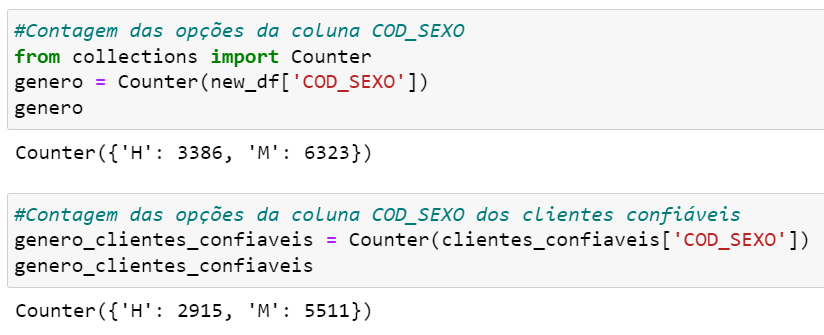
**4. Análise e Exploração de Dados**

O objetivo da análise é verificar a discrepância de característica entre os os clientes que possuem cartão de crédito e os que efetivamente foram considerados como confiáveis. Sempre que possível, as informações serão separadas nesses critérios e, para isso, a partir do dataframe “new\_df” será criado o dataframe “clientes\_confiaveis” apenas com os clientes confiáveis.



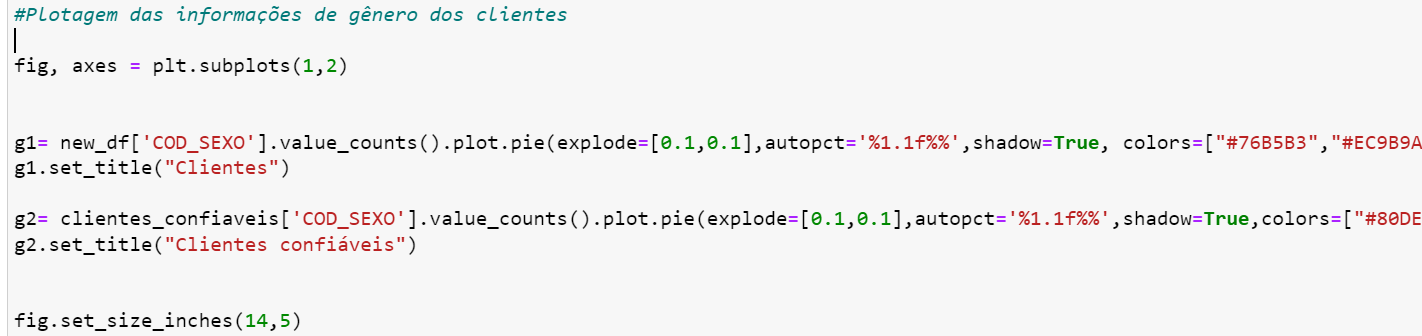
Observa-se que o dataframe “clientes\_confiaveis” possui 8426 linhas, o que corresponde a um grande número em relação ao total de linhas no dataframe original. Inicialmente a análise será feita pelo gênero dos clientes, fazendo a contagem do total de clientes por gênero, como também, a contagem por gênero dos clientes confiáveis.

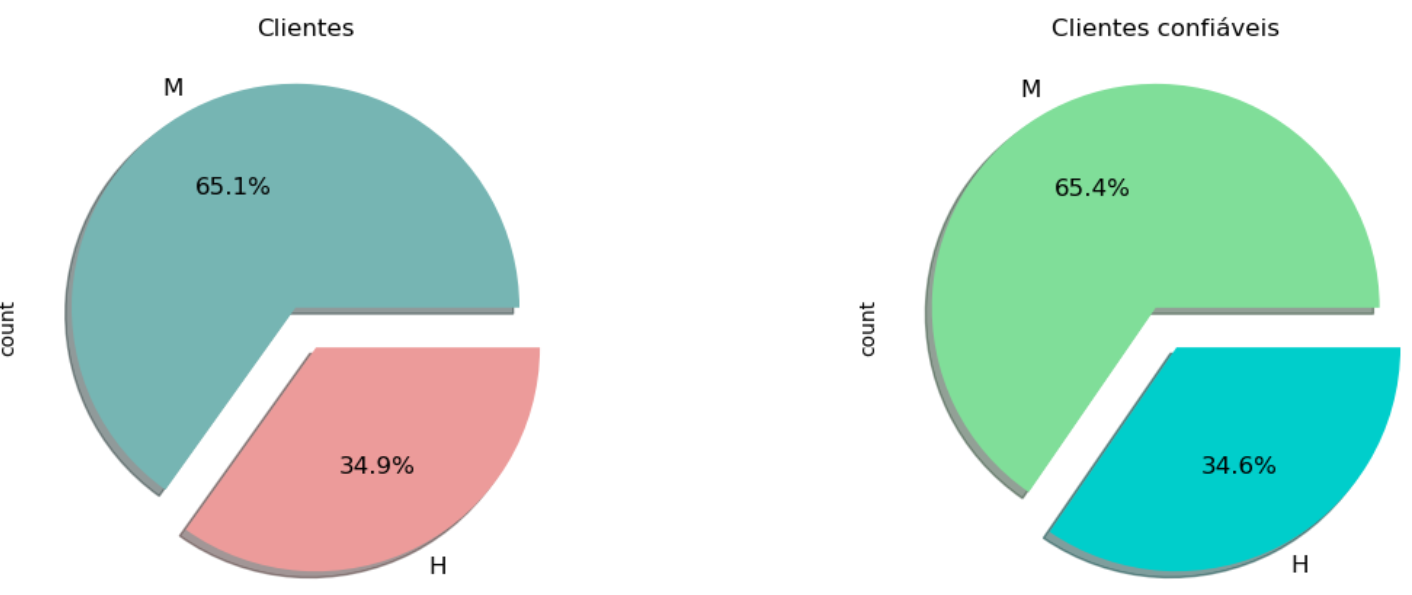
Faremos a importação da biblioteca “Collections”, com o uso da função "Counter", que conta quantas vezes uma determinada opção aparece em uma série de dados a armazena esses valores em um dicionário.



Para facilitar a visualização das informações sobre cada gênero, através biblioteca “Matplotlib” para a criação de gráficos, serão criados os gráficos de pizza com os percentuais e gêneros.

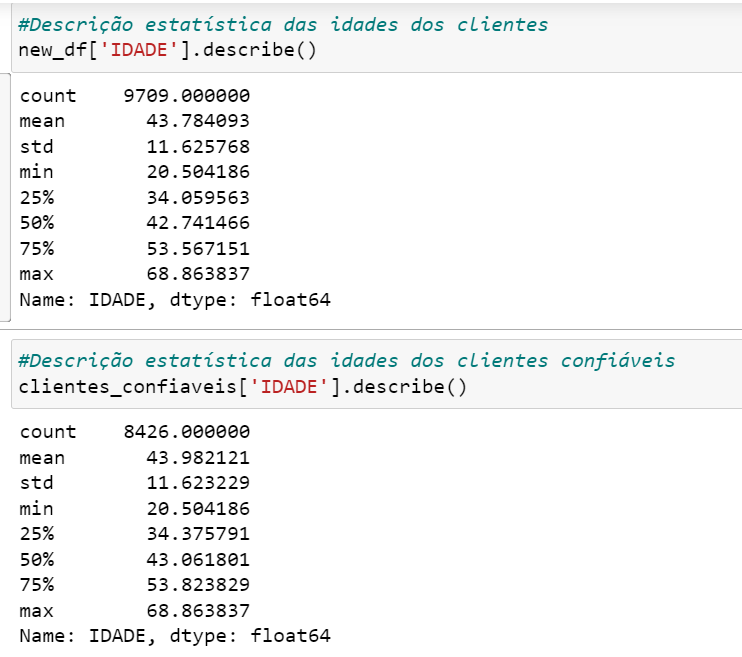


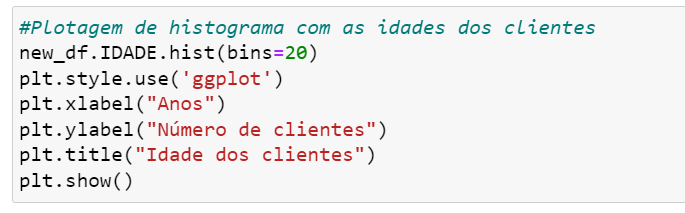


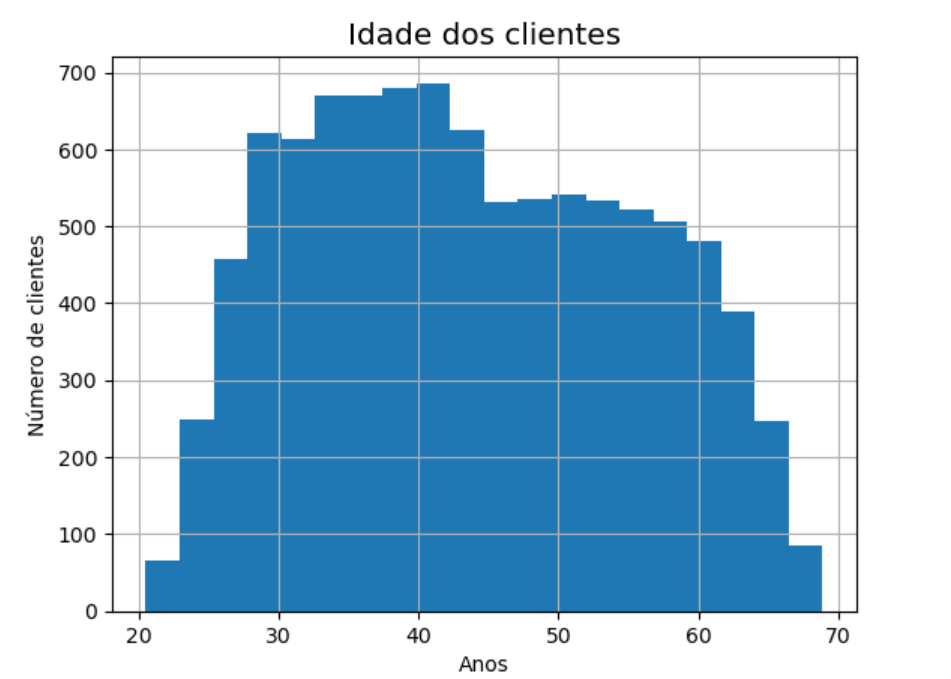


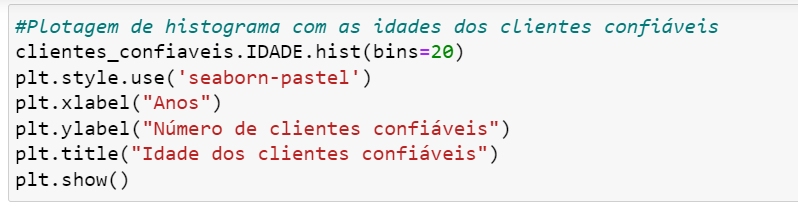
Analisando os gráficos pode-se perceber que o percentual de cada gênero é muito semelhante entre o total dos clientes e os clientes confiáveis. Em ambos os gráficos, a maioria do percentual é do sexo feminino com valores de 65.1% e 654%.

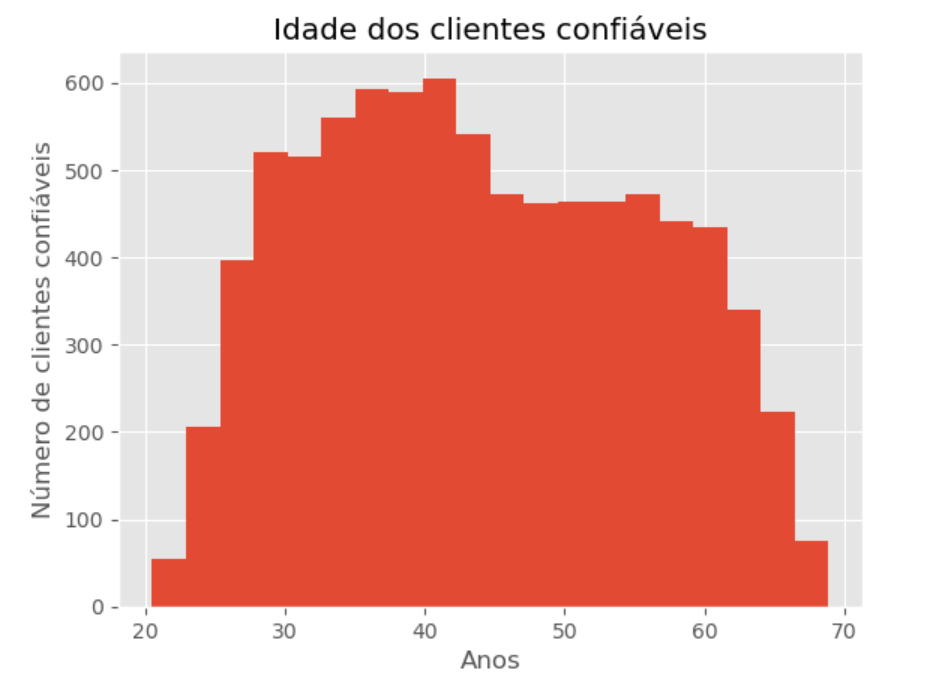
O próximo item a ser analisado é a idade dos clientes. Para essa análise será utilizada a função “describe” que apresenta os principais indicadores estatísticos de uma série de dados. Na visualização da distribuição dessas idades serão utilizados histogramas, seguindo os mesmos estilos descritos anteriormente.





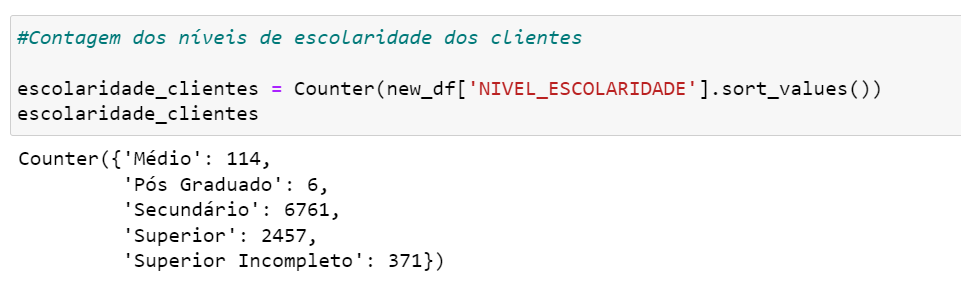


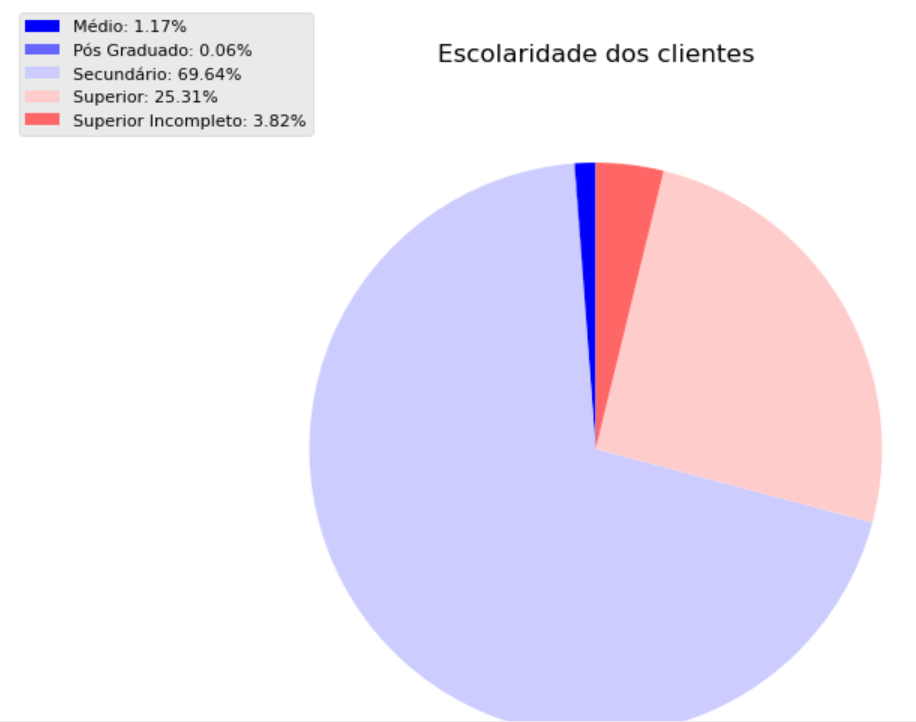


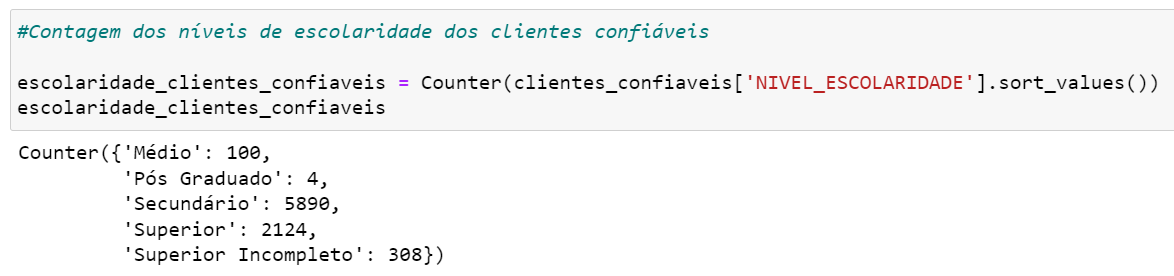


Analisando os histogramas, não se verifica nenhuma mudança significativa. Essa constatação pode ser confirmada por meio dos indicadores estatísticos das duas séries de dados que são muito similares, com variações mínimas na média de idade, por exemplo: 43,78 anos para todos os clientes e 43,98 anos para os clientes confiáveis.

A próxima análise será o nível de escolaridade dos clientes. Utilizaremos novamente a função “Counter”, para determinar quantas vezes um nível de escolaridade aparece no dataframe, e o gráfico de pizza com os valores e percentuais dos níveis de escolaridade.



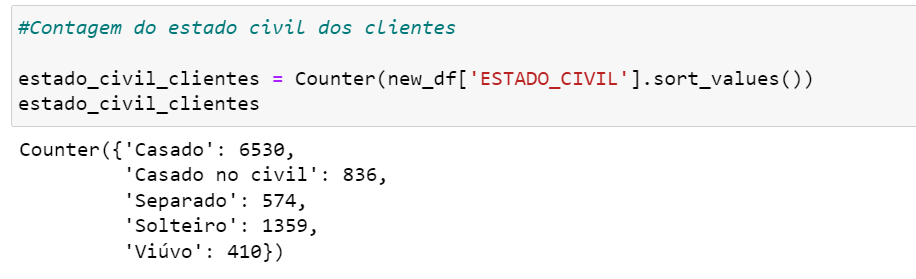


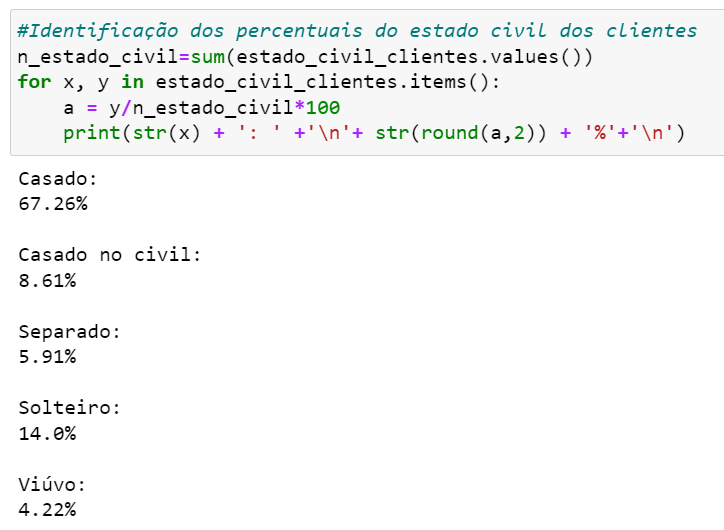


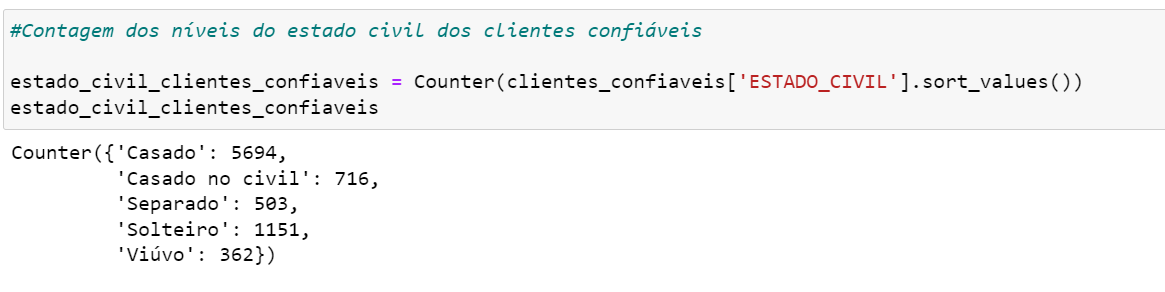


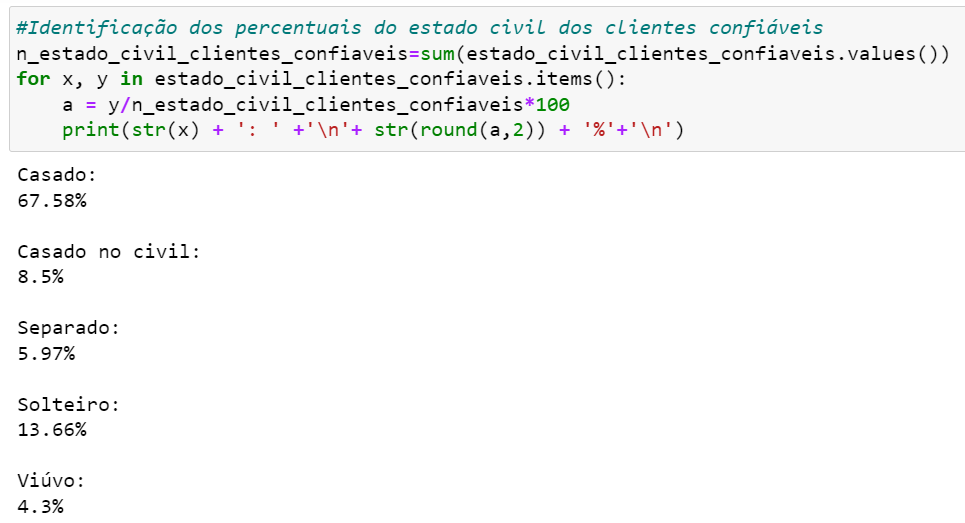
Percebe-se uma grande semelhança entre os níveis de escolaridade de todos os clientes e os clientes confiáveis. É importante destacar que a grande maioria dos clientes são do nível secundário.

A próxima análise será o estado civil. Os passos utilizados para a análise dos níveis de escolaridade serão semelhantes para o estado civil dos clientes, ou seja, a contagem dos valores nos dataframes e a apresentação dos gráficos de pizza. Porém, foi criado um laço de repetição “for” para apresentação dos percentuais de cada estado civil.



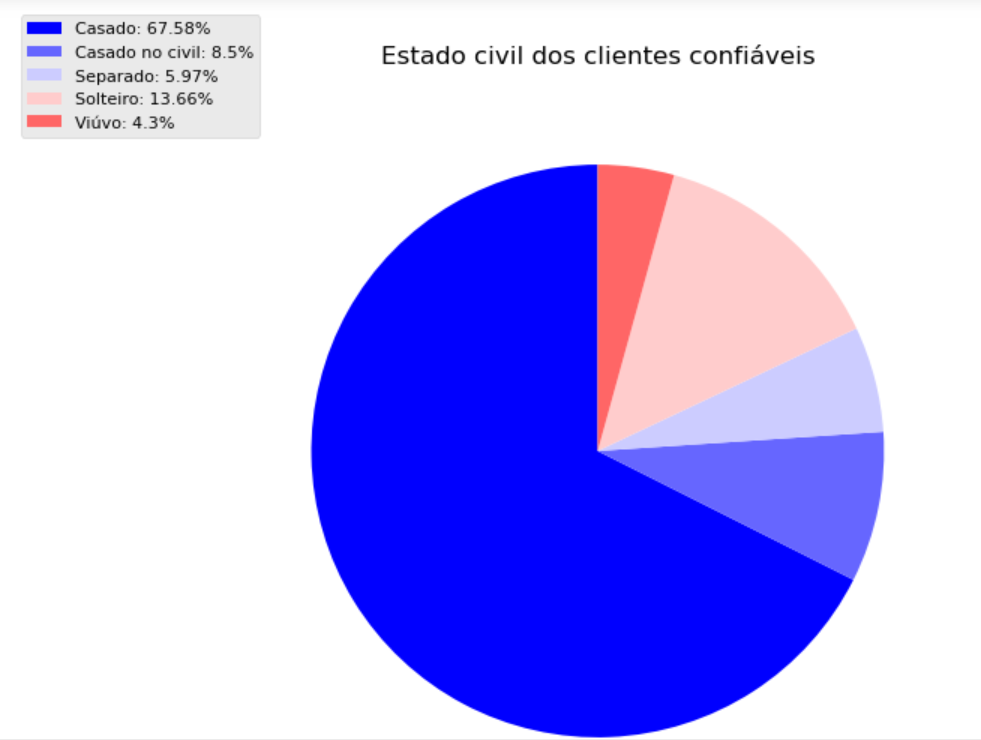






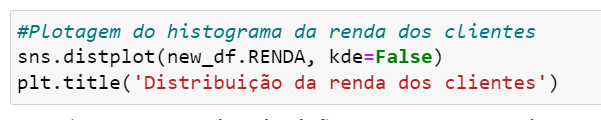
Os resultados da análise do estado civil também serão apesentados em gráficos.

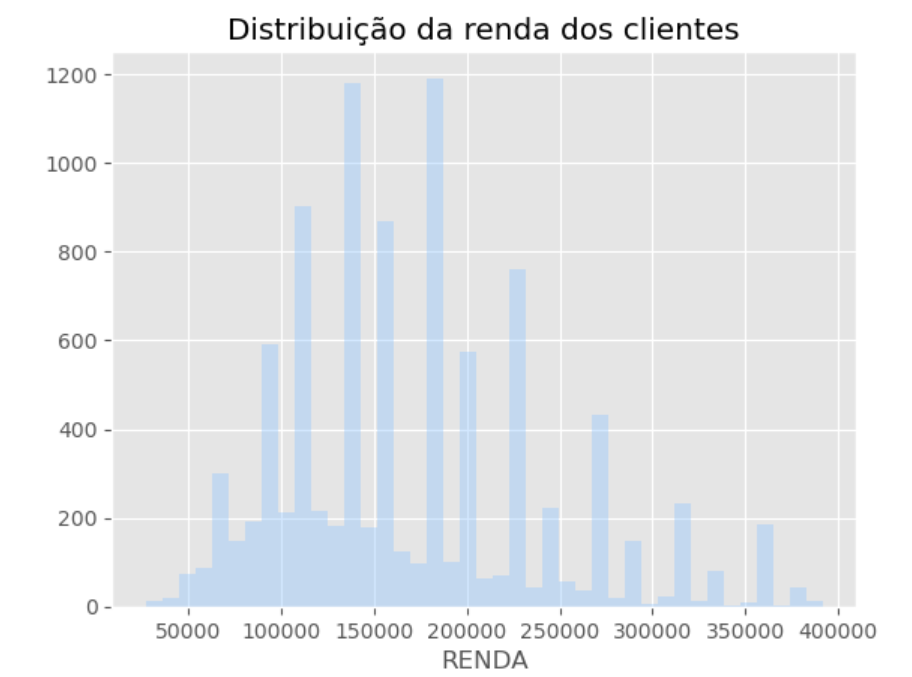


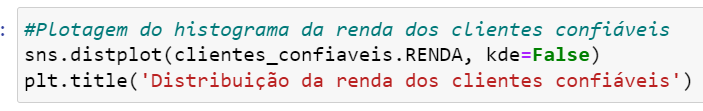


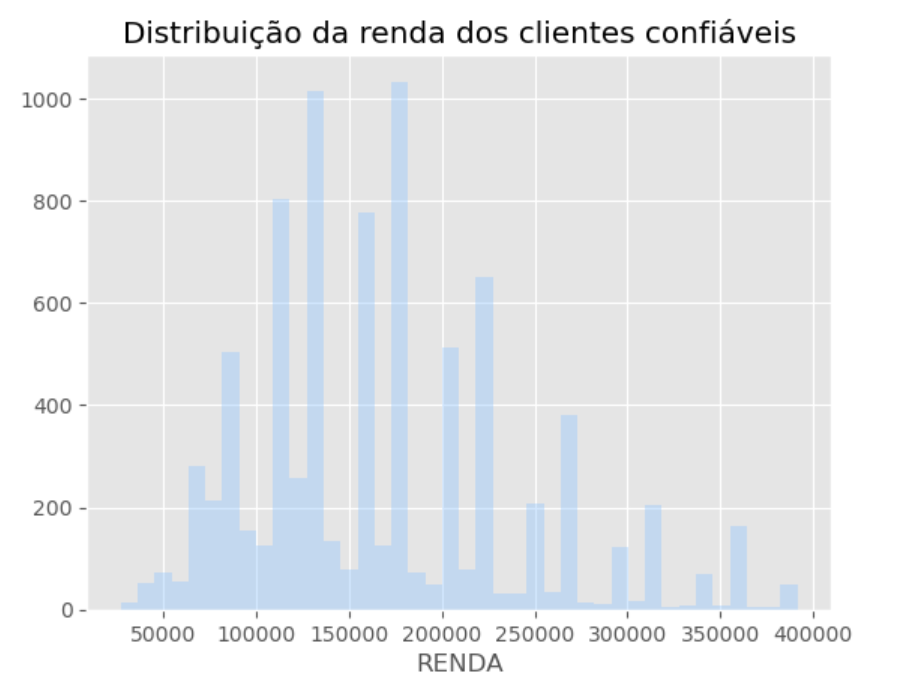
Percebe-se por meio dos gráficos e percentuais que a maioria dos clientes é casado, com 67,26%, número que se eleva quando avaliamos os clientes confiáveis, com 67,58%.

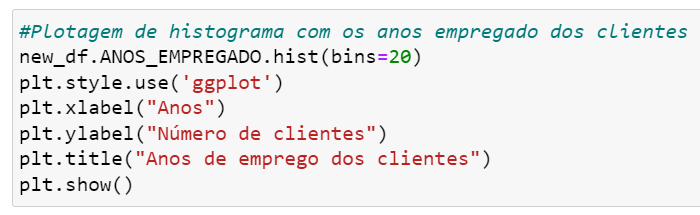
Para a distribuição de renda, anos de emprego e a posse de imóvel, carro e emprego, os percentuais entre o total de clientes e os clientes confiáveis estão também semelhantes, conforme os gráficos apresentados.

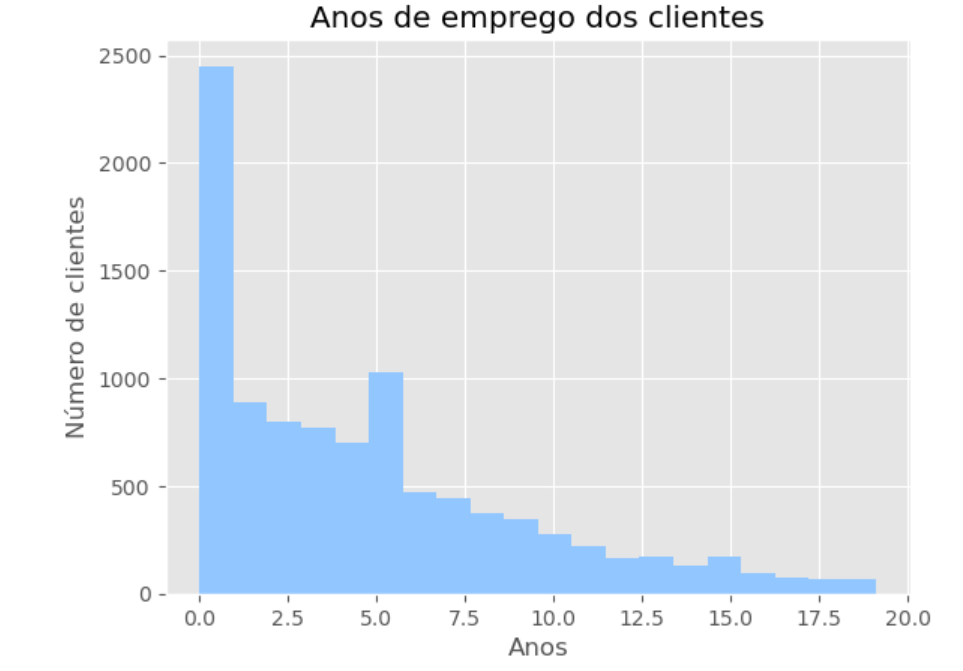


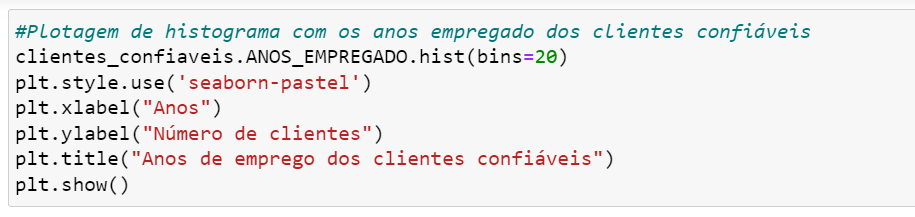


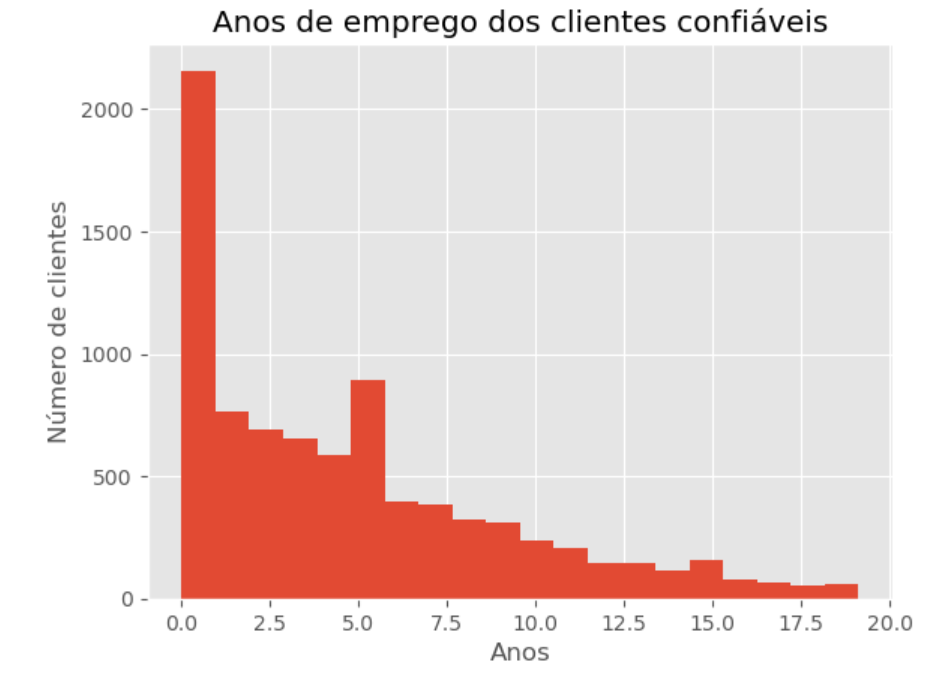


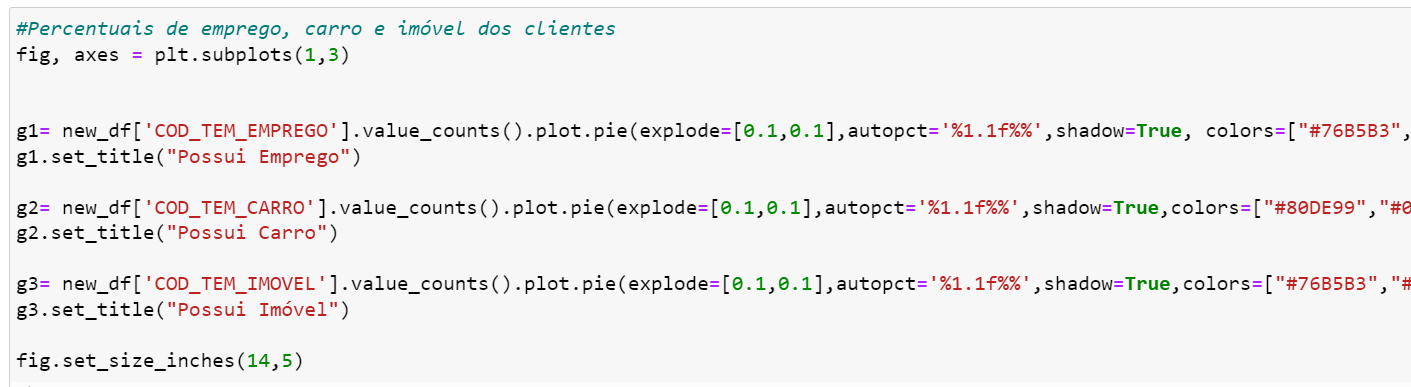


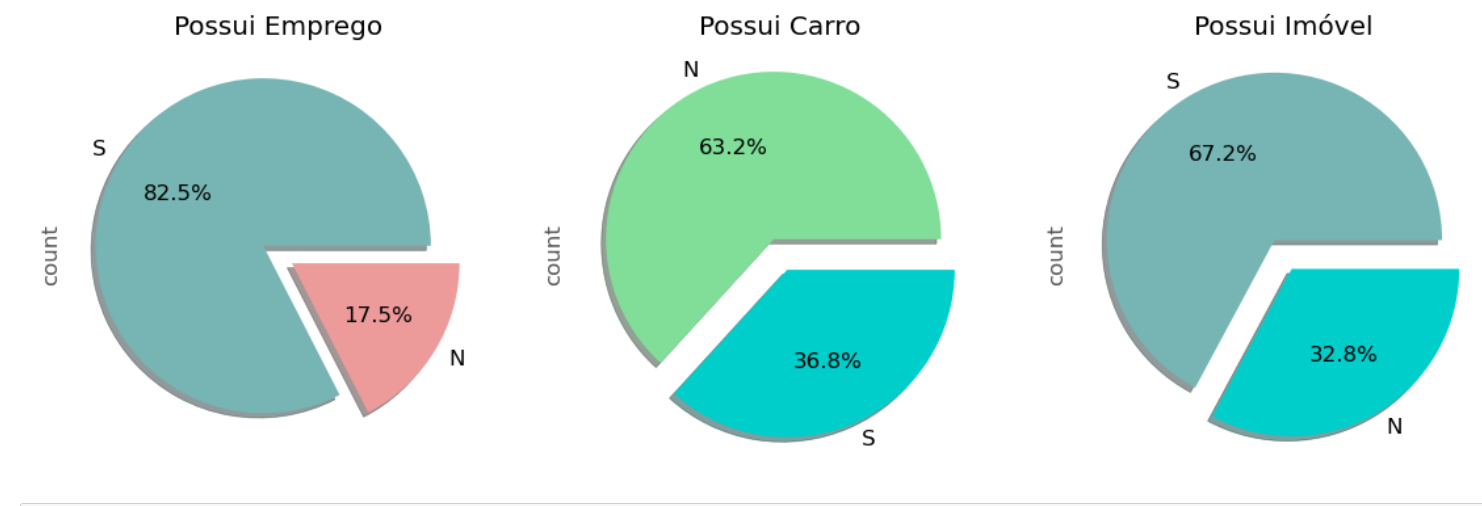


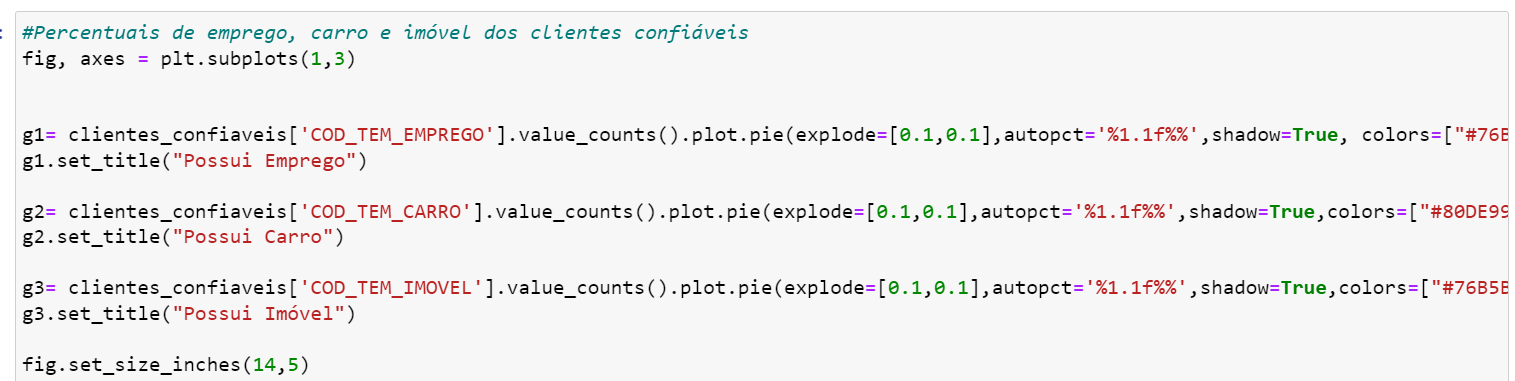


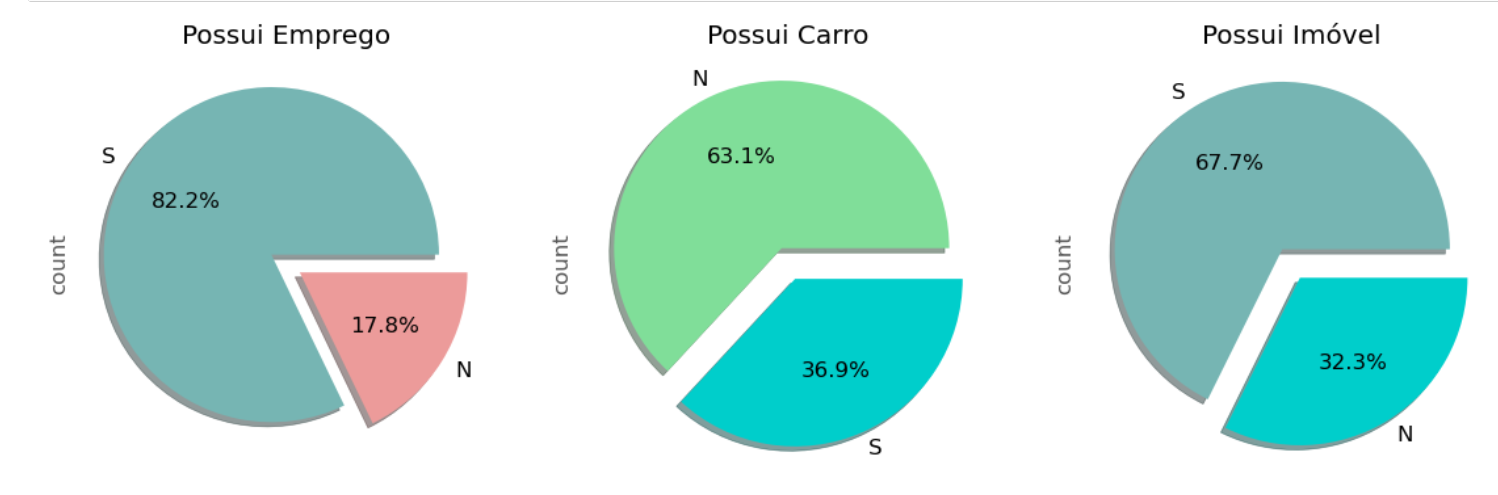




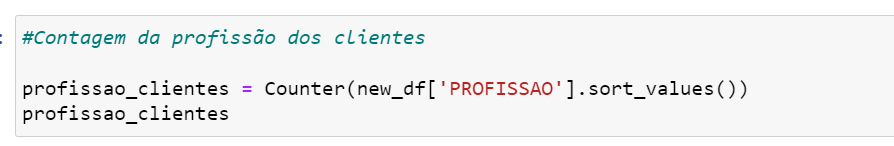






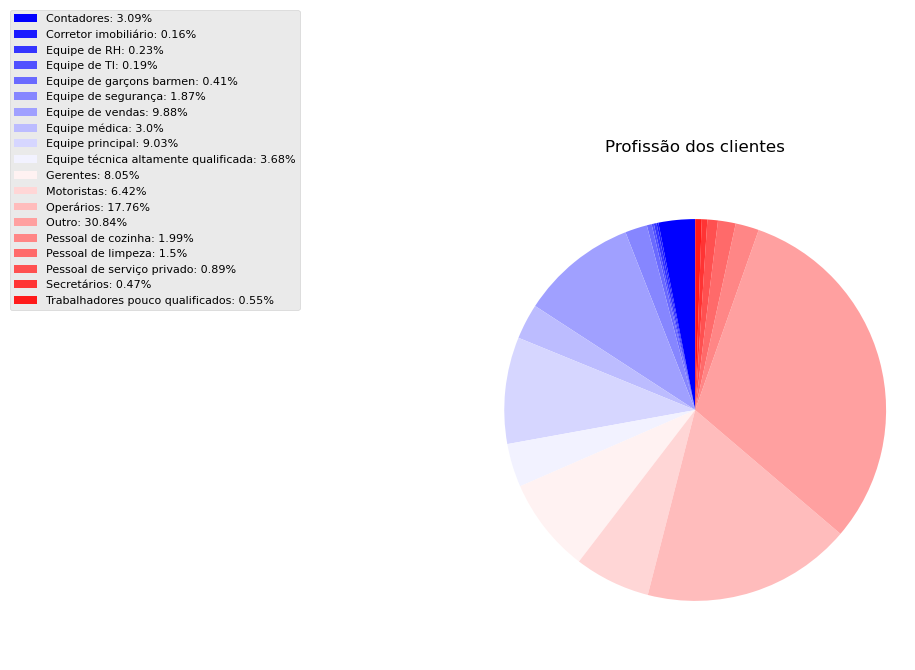


Ao analisar a profissão, podemos perceber que há um grande percentual para as profissões que não foram categorizadas, aproximadamente 30%, para todos os clientes e os clientes confiáveis.

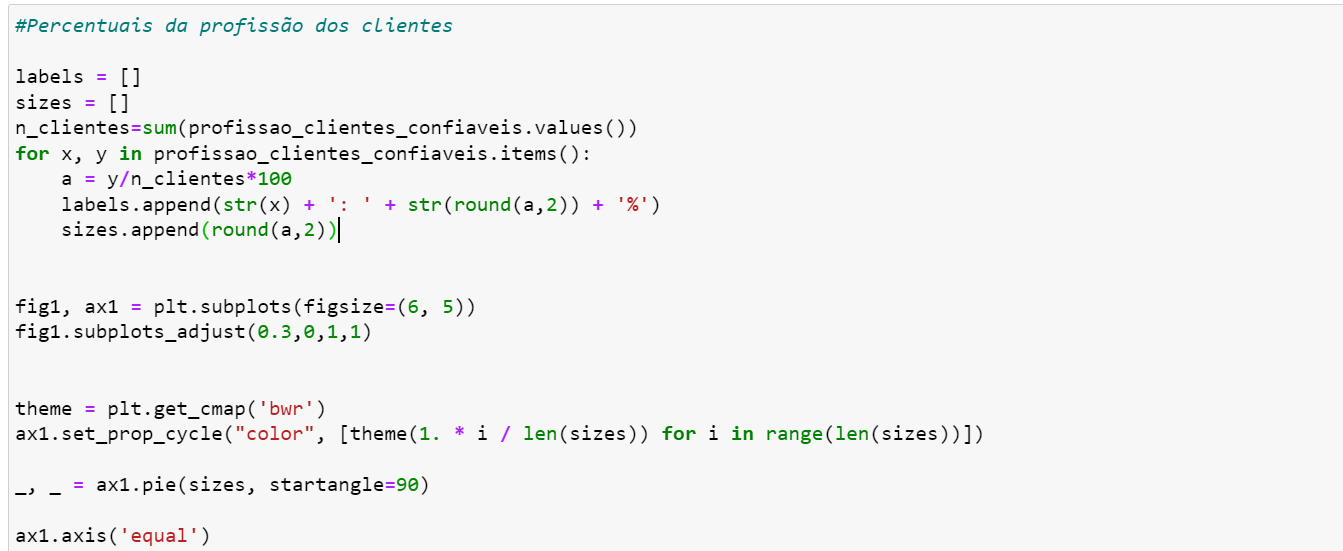


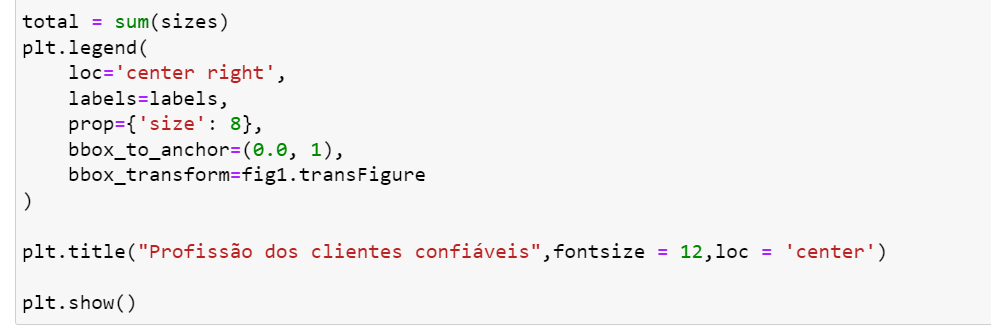


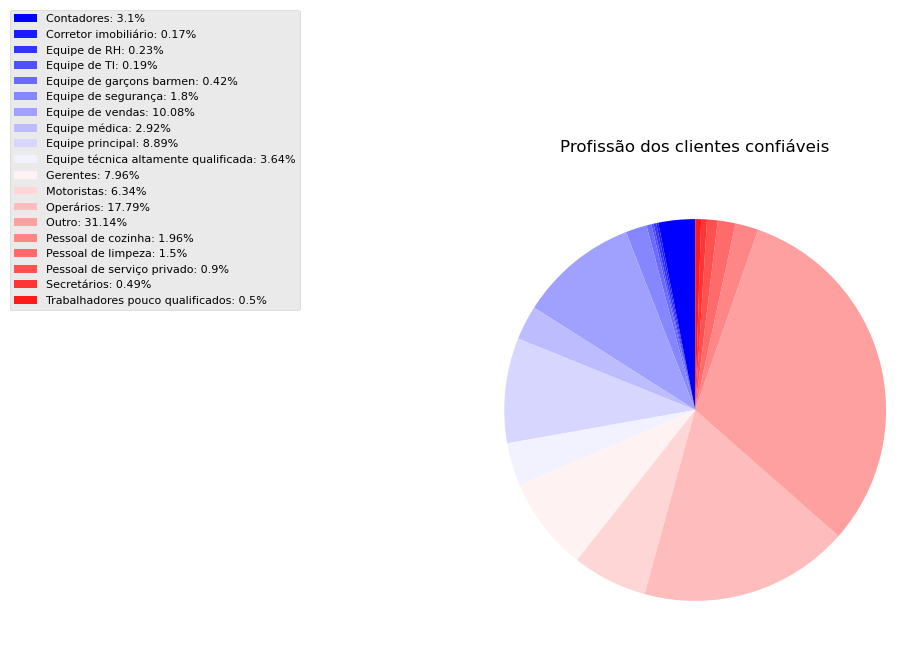












Da análise de todas as informações desse dataframe, percebe-se que os dados dos clientes, confiáveis ou não, são semelhantes, o que indica um comportamento padrão nos dados do dataframe.

**5. Criação de Modelos de Machine Learning**

O aprendizado de máquina (Machine Learning) é o estudo de algoritmos de computador que melhoram automaticamente através da experiência e do uso de dados. Pode ser classificado em: a) aprendizado supervisionado, em que o computador é apresentado com exemplos de entradas e suas saídas desejadas, rotuladas, e o objetivo é aprender uma regra geral que mapeia entradas em saídas; b) aprendizado não supervisionado, em que nenhum rótulo é dado ao computador, deixando-o sozinho para encontrar a estrutura em sua entrada. O aprendizado não supervisionado pode ser um objetivo em si (descobrir padrões ocultos nos dados) ou um meio para um fim (aprendizado de recursos); e c) aprendizado por reforço, em que o computador interage com um ambiente dinâmico no qual deve realizar uma determinada tarefa. À medida que navega no espaço do problema, o programa recebe um feedback, que ele tenta maximizar.

O objetivo do estudo é identificar quais clientes são confiáveis dado um conjunto de características e, para isso, optou-se por utilizar algoritmos de classificação, que são cálculos preditivos usados para atribuir dados a categorias predefinidas, analisando um conjunto de dados de treinamento.

Nesse estudo serão utilizados 6 tipos de algoritmos de classificação, são eles: Árvore de Decisão, Regressão Logística, Naïve Bayes, Gradiente Descendente, KNN (K - Nearest Neighbors) e Randon Forest.

Árvores de Decisão é uma tabela de decisão sob a forma de árvore com nós e folhas sequenciais e interligados que classificam as instâncias, ordenando-as com base nos valores dos recursos. Cada nó em uma árvore de decisão representa uma característica em uma instância a ser classificada, e cada ramo representa um valor que o nó pode assumir. As instâncias são classificadas começando no nó raiz com base em seus valores de recursos. Trata-se de um dos modelos mais práticos e mais utilizados em inferência por indução.

A regressão logística está entre as técnicas mais utilizadas quando se deseja entender as correlações entre variáveis de entrada e saída de um modelo, normalmente são as primeiras a serem testadas em modelos de machinelearning. É uma técnica recomendada para situações em que a variável dependente é de natureza dicotômica ou binária. Quanto às independentes, tanto podem ser categóricas ou não. Ela é mais útil no entendimento da influência de diversas variáveis independentes em uma saída única variável. A desvantagem é que funciona apenas quando a variável prevista é binária, assume que todos os preditores são independentes uns dos outros e assume que os dados estão livres de valores ausentes.

O algoritmo Naive Bayes é uma técnica de classificação com uma suposição de independência entre as variáveis preditoras. O classificador Naive(do inglês ingênuo) Bayesassume que a presença de uma característica particular em uma classe não está relacionada com a presença de qualquer outra. As principais vantagens deste algoritmo é que ele requer uma pequena quantidade de dados de treinamento para estimar os parâmetros necessários e é extremamente rápido em comparação com métodos mais sofisticados. A maior desvantagem é que ele é conhecido por não ser um bom avaliador.

O algoritmo gradiente descendente é um dos algoritmos de maior sucesso em problemas de Machine Learning. O método consiste em encontrar, de forma iterativa, os valores dos parâmetros que minimizam determinada função de interesse. As vantagens desse algoritmo são sua eficiência e facilidade de implementação, mas ele requer vários hiperparâmetros e é sensível ao dimensionamento de recursos.

O algoritmo k-nearest neighbor, muitas vezes abreviado k-nn, é uma abordagem para classificação de dados que estima a probabilidade de um ponto de dados ser membro de um grupo ou de outro, dependendo de qual grupo os pontos de dados mais próximos a ele estão. Esse algoritmo é de fácil implementação, robusto a variações nos dados de treinamento e efetivo quando temos muitos dados de treinamento. O ponto negativo desse algoritmo é que exige muito recurso computacional.

Por fim, o algoritmo Random Forest que é um algoritmo de aprendizagem supervisionada que cria uma “floresta” de um modo aleatório. A “floresta” é uma combinação (ensemble) de árvores de decisão, na maioria dos casos treinados com o método de bagging. A ideia principal do método de bagging é que a combinação dos modelos de aprendizado o que aumenta o resultado geral. Dizendo de outro modo, o algoritmo cria várias árvores de decisão e as combina para obter uma predição com maior acurácia e mais estável. As dificuldades desse modelo são sua dificuldade de implementação e complexidade do algoritmo.