UNIVERSIDADE DE SÃO PAULO
ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

EDUARDO PALHARES JÚNIOR

Sistemas especialistas aplicados na análise de cíclos de mercado: um estudo comparativo utilizando indicadores antecedentes para classificação das fases do ciclo econômico brasileiro

São Paulo
2021

Tese de autoria de Eduardo Palhares Júnior, sob o título **"Sistemas especialistas aplicados na análise de cíclos de mercado: um estudo comparativo utilizando indicadores antecedentes para classificação das fases do ciclo econômico brasileiro"**, apresentada à Escola Politécnica da Universidade de São Paulo, para obtenção do título de Duotor em Ciências pelo Programa de Pós-graduação em Engenharia Elétrica, na área de concentração Sistemas Eletrônicos, aprovada em ____ de _____ de _____ pela comissão julgadora constituída pelos doutores:

Prof. Dr. Flávio Almeida de Magalhães Cipparrone
USP
Presidente

Prof. Dr. Afonso de Campos Pinto
FGV

Profa. Dra. Elia Matsumoto
FGV

Escreva aqui sua dedicatória, se desejar, ou remova esta página...

Agradecimentos

"Escreva aqui uma epígrafe, se desejar, ou remova esta página..."

(Autor da epígrafe)

Resumo

PALHARES JÚNIOR, Eduardo. **Sistemas especialistas aplicados na análise de ciclos econômicos**: um estudo comparativo utilizando indicadores antecedentes para classificação das fases do ciclo econômico brasileiro. 2021. 92 f. Tese (Doutorado em Ciências) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2021.

Este trabalho busca realizar um estudo comparativo entre diversas técnicas de machine learning, aplicados na análise das fases do ciclo econômico brasileiro. Para tanto, foram utilizados diversos indicadores macroeconômicos para construir um modelo que fosse capaz de identificar e prever os pontos de virada do ciclo econômico, como o inicio de uma recessão ou de uma recuperação. A discretização das variáveis mostrou-se determinante na qualidade do processo de classificação, por conta da diversidade dos dados e da natureza não-linear do fenômeno analisado. As diferentes técnicas utilizadas reforçam um dilema, pois muitas vezes os melhores resultados vem de métodos de difícil interpretação, de modo que o modelo torna-se uma caixa preta.

Palavras-chaves: Aprendizago de máquina. Classificação. Ciclo econômico.

Abstract

PALHARES JUNIOR, Eduardo. **Expert systems applied in the analysis of economic cycles**: a comparative study using leading indicators to classify the phases of the Brazilian economic cycle. 2021. 92 p. Thesis (Philosophiæ Doctor) – Polytechnic School, University of São Paulo, São Paulo, 2021.

This work proposes a comparative study between several machine learning techniques, applied in the analysis of the phases of the Brazilian economic cycle. To this end, several macroeconomic indicators were used to build a model that was able to identify and predict the turning points of the economic cycle, such as the beginning of a recession or a recovery. The discretization of the variables proved to be decisive in the quality of the classification process, due to the diversity of the data and the non-linear nature of the analyzed phenomenon. The different techniques used reinforce a dilemma, because often the best results come from methods that are difficult to interpret, so that the model becomes a black box.

Keywords: Machine Learning. Classification. Economic Cycle.

Lista de figuras

Lista de algoritmos

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

Sigla/abreviatura 1	Definição da sigla ou da abreviatura por extenso
Sigla/abreviatura 2	Definição da sigla ou da abreviatura por extenso
Sigla/abreviatura 3	Definição da sigla ou da abreviatura por extenso
Sigla/abreviatura 4	Definição da sigla ou da abreviatura por extenso
Sigla/abreviatura 5	Definição da sigla ou da abreviatura por extenso
Sigla/abreviatura 6	Definição da sigla ou da abreviatura por extenso
Sigla/abreviatura 7	Definição da sigla ou da abreviatura por extenso
Sigla/abreviatura 8	Definição da sigla ou da abreviatura por extenso
Sigla/abreviatura 9	Definição da sigla ou da abreviatura por extenso
Sigla/abreviatura 10	Definição da sigla ou da abreviatura por extenso

Lista de símbolos

Γ	Letra grega Gama
Λ	Lambda
ζ	Letra grega minúscula zeta
$\in$	Pertence

Sumário

1 Introdução

Uma percepção comum do funcionamento das economias modernas mostra que elas possuem uma taxa de crescimento que seguem tendências, com fases de expansão e recessão. Os períodos de expansão trazem crescimento econômico e geralmente aumentam a qualidade de vida e os padrões de consumo, como poder de compra, aumento de salários e a capacidade de pagar por boa educação e saúde, enquanto que nos períodos de recessão, a tendência é que ocorra o oposto. À medida que as economias param de crescer, os salários tendem a se estabilizarem e, com isso, ocorre uma redução do poder de compra que às vezes pode levar a dificuldades de acesso à saúde e educação básica. Essas flutuações no crescimento afetam não apenas os indivíduos, mas também as empresas, com redução das demandas, das oportunidades econômicas e da lucratividade. É evidente que essas mudanças no crescimento econômico também afetam os governos, como agentes macroeconômicos, com uma pressão crescente de pessoas e empresas necessitando de auxílio para reverter a tendência de queda.

Em razão desses fatores, torna-se clara a importância que as pessoas, empresas e governos tem em relação a alguma metodologia que consiga prever quando essas mudanças irão acontecer. Portanto, a principal preocupação desta tese é encontrar uma maneira de prever esses eventos com tempo de espera suficiente para que esses agentes possam tomar as precauções cabíveis. Como a análise de todas as economias globais é impraticável devido às diferenças estruturais entre elas, esse trabalho terá como foco a economia do Brasil, já que é a principal economia da américa latina e a 8ª economia do mundo, de acordo com os últimos Relatório do Fundo Monetário Internacional (IMF, 2020). Dessa forma, uma recessão no Brasil terá forte impacto em outras economias.

1.1 Motivação

A classificação de variáveis econômicas para inferir desacelerações econômicas e recessões tem sido usado por um longo tempo em pesquisas na área da economia, pelo menos desde Burns e Mitchell (1946). Ao longo do tempo, várias variáveis distintas foram propostas como indicadores econômicos, como discutido em Estrella e Mishkin (1995), os quais são reconhecidos como úteis na identificação de um ponto de inflexão da economia,

sendo que a mais conhecida continua sendo a curva de juros. (KAUPPI; SAIKKONEN, 2008; RUDEBUSCH; WILLIAMS, 2009)

Existem também diversos pesquisadores acadêmicos que se concentram em produzir análises do estado da economia, como por exemplo o Índice de Atividade Nacional do Fed de Chicago (CFNAI) do Federal Bank of Chicago CFNAI (2020), que consiste na média ponderada dos 85 indicadores mensais da atividade económica dos Estados Unidos. A ideia por trás de sua abordagem é que existe algum fator comum a todos os vários indicadores de inflação, que se utilizado como índice, torna-se útil para prever a inflação.

Nos resultados apresentados em Stock e Watson (1999) mostrou-se que o CFNAI fornece um indicador útil sobre a atividade econômica atual e futura e a inflação nos Estados Unidos. No entanto, a maioria dos economistas segue uma ampla variedade de indicadores cuja usabilidade não pode ser determinada para avaliar o estado futuro da economia, uma vez que, até agora, nenhum deles se mostrou totalmente confiável no passado.

Assim, esta tese pretende desenvolver modelos de aprendizado de máquina que analisem diversos indicadores econômicos e sinalizem a possibilidade de um ponto de inversão do crescimento econômico, no caso, o início de uma fase de recessão em diversos horizontes temporais.

1.2 Objetivos

Frente a motivação descrita anteriormente, foi estabelecido um conjunto de objetivos com o intuito de alcançar êxito na problemática descrita. Esses objetivos são baseados em um conjunto de suposições necessárias para tornar a solução para o problema factível:

- Não se pretende adivinhar as datas de início/fim de uma recessão;
- Não se pretende explorar a causa nem a intensidade de cada recessão;
- Somente será analisada a economia do Brasil.

Definidos as hipóteses das quais o trabalho não tem a pretensão de se aprofundar, definem-se a seguir os objetivos a serem abordados nesse trabalho:

- Selecionar e analisar diversos indicadores da economia brasileira;
- Identificar e antecipar os pontos de virada do ciclo econômico;

- Analisar a diferença de desempenho entre diversos modelos;

Ao cumprir estes objetivos, espera-se fazer um estudo exploratório sobre o assunto e desenvolver um método que possa prever, com um razoável avanço de tempo, uma mudança para uma recessão económica e, se possível, superando o estado da arte atual.

1.3 Estrutura do documento

1. **Introdução:**
   Proporciona uma visão geral do trabalho, fornecendo sua motivação, objetivos e contribuições.
2. **Problema de pesquisa:**
   Estabelece a base teórica utilizada nesta tese, no nível econômico, mas também nos modelos usados para recuperar os resultados. Apresenta o estado da arte sobre o assunto.
3. **Arquitetura:**
   Apresenta as metodologias adotadas e a arquitetura de software utilizada para realização do trabalho.
4. **Implementação:**
   Discute aspectos mais técnicos relativos as estratégias utilizadas na implementação, bem como as potencialidades e limitações.
5. **Resultados:**
   Apresenta os resultados alcançados pelos métodos propostos, bem como os métodos de validação utilizados.
6. **Conclusão:**
   Resume o desenvolvimento do trabalho, e estabelece os melhores resultados e estratégias estabelecidas por este trabalho. Além disso, descreve as possibilidades de trabalhos futuros nesta área.

2 Problema de pesquisa

2.1 Estado da arte

A previsão de uma recessão econômica é um objetivo há muito desejada pelos economistas, mas também por muitas outras áreas de estudo, incluindo engenharia e ciência da computação. Com a disseminação e o desenvolvimento da computação pessoal, alguns modelos matemáticos antigos tornaram-se cada vez mais fáceis de aplicar e utilizar em grandes quantidades de dados e informações. Este desenvolvimento conduziu ao nascimento e desenvolvimento de áreas de estudo como o Machine Learning (ML), Data Mining (DM) e as Redes Neurais (NN), que passaram a dar respostas a questões há muito tempo em aberto, como discutido em Gorgulho *et al.* (2011) e em Canelas *et al.* (2013). As combinações dessas ideias guiaram muitos cientistas e economistas ao domínio da informática para prever recessões econômicas. (SILVA *et al.*, 2015; ALMEIDA *et al.*, 2018).

Mesmo sendo reconhecido como um mercado altamente competitivo, esses estudos ainda estão rodeados de segredos e mistérios, onde nem todas as metodologias da literatura, conjuntos de dados ou mesmo resultados são apresentados e explicados. Estas omissões apresentam-se como um dos principais motivos para a exploração deste trabalho, onde o objetivo não é só mostrar bons resultados, mas principalmente explicar os detalhes para que outras pessoas possam utilizar a desenvolver esse conhecimento. Portanto, para estabelecer um campo de comparação, este capítulo usa a literatura mais completa disponível e está dividido em três áreas de relevância: os modelos, os dados e as análise de resultados onde são discutidas as métricas de avaliação.

2.1.1 Dados

Os dados usados para alimentar os diferentes modelos são os primeiros e provavelmente uma das partes mais importantes na concepção de um processo para descobrir quando uma recessão vai acontecer. É aqui que parte da literatura começa a ocultar seus conjuntos de dados ou mesmo onde a informação foi adquirida. Este último ponto é muito importante porque as informações macroeconômicas geralmente estão sujeitas a mudanças e atualizações pelas estruturas que as produzem. Desta forma, não se trata apenas de ter a quantidade certa de informações, mas também de ter a informação certa.

Como esse tipo de abordagem ainda é inédita no contexto da economia brasileira, será realizada uma discussão com referência ao que se observou em relação à outras economias globais. Cada autor utiliza diversos indicadores diferentes, mas como afirmado no contexto econômico, algumas áreas parecem ser de maior relevância do que outras.

Estrella e Mishkin (1995), alguns dos primeiros a explorar este tema, propõe variáveis mais ligadas a taxas de juros e disponibilidade de dinheiro, ao invés de produção e condições de trabalho. Indicadores como a curva de rendimentos, os preços das ações do Dow Jones e S&P500, e as variações do estoque monetário, representam as principais contribuições nesta área. Ainda assim, eles ainda usaram outras informações de diferentes áreas, como licenças de habitação ou o índice de preço ao consumidor.

Berge (2013) utiliza outros indicadores macroeconômicos como a produção industrial, as horas semanais médias da indústria de transformação ou as solicitações de seguro desemprego. Essas abordagens seguem um conjunto menor de indicadores, geralmente entre dez e vinte indicadores diferentes, mas nem todos os autores usam esse método de baixa informação. Por exemplo, Kisinbay e Baba (2011) do FMI utilizam até 166 indicadores macroeconômicos divididos em cinco categorias: renda e produção, emprego, construção, taxas de juros e, finalmente, preços nominais e salários.

Diferente dos casos anteriores, o Wells Fargo Securities Economics Group aposta em uma abordagem com um grande número de indicadores. Eles começam recuperando 500.000 indicadores do Federal Reserve Bank de St. Louis (FRED), então, usando apenas a data de início de 1972, reduzem o conjunto de dados para 5.889 variáveis e, eventualmente, usando procedimentos adicionais, reduzem o conjunto de dados para uma contagem final de 192 indicadores diferentes.(SILVIA, 2018)

Essas abordagens apresentam duas grandes diferenças, uma com um baixo número de indicadores econômicos, onde as variáveis são escolhidas de acordo com uma análise econômica e diferem do ponto de vista de cada autor. O outro usa um grande número de indicadores econômicos e não faz suposições reais sobre os dados, mas requer um árduo processo de recuperação e preparação dos mesmos.

2.1.2 Modelos

Estabelecidos os dados a serem usados no processo de previsão, existem vários modelos matemáticos propostos para resolver este tipo de problema de classificação binária. A maior parte da literatura segue os mesmos princípios básicos sobre o assunto, acrescentando algumas alterações nos métodos de tratamento e previsão de dados, mas quase sempre utilizando modelos multivariados.

Os métodos mais comumente usados são os modelos probit (LIU; MOENCH, 2016) e logit (KISINBAY; BABA, 2011). Esses modelos são considerados boas soluções para problemas de classificação binária e, portanto, muitos autores os utilizaram como parte de seu sistema para desenvolver a previsão final. Mas mesmo com os resultados positivos mostrados por esses modelos, a comunidade continuou tentando diferentes abordagens para o problema e, ultimamente, empresas como a Wells Fargo Securities também tentaram abordagens da Random Forest e Gradient Boosting para lidar com seus dados(SILVIA, 2018). Esses modelos são o que há de mais moderno, não apenas na área de detecção de recessão, mas em quase todos os problemas que tratam de classificação binária.

Mas, além dos próprios modelos, Kauppi e Saikkonen (2008) usaram várias outras estratégias para melhorar seus resultados, interferindo nos dados ou mesmo nos modelos. Uma das estratégias para prever recessões com algum tempo de antecedência era utilizar algum modelo específico, e então usar seus resultados e prevê-los com um novo modelo univariado. Essa técnica permitiu que eles comprendessem e conseguissem interpretar as previsões obtidas. Eles também apresentaram uma alternativa mais complicada para o objetivo direto, um modo iterativo de prever vários trimestres à frente e tentar decifrar se havia uma recessão entre esses períodos.

Mas não apenas sobre os métodos de previsão do futuro essas mudanças são aplicadas, Berge (2013) também aplica várias médias de modelo em seus modelos de previsão para tentar melhorar seu poder de predição. Ele utiliza uma média ponderada normal, uma média ponderada diferente chamada Bayesian Model Average, e também uma solução alternativa de escolha do melhor modelo por meio de um algoritmo de boost.

Todos os exempos apresentados não são apenas soluções de ponta para detectar uma recessão económica, mas também configuram alguns dos testes e transformações efectuadas nos dados e modelos para melhorar os seus resultados.

2.1.3 Análise de resultados

Em relação aos resultados e às métricas do estado da arte na detecção de recessão econômica, quando nos referimos a economia norte-americana, nenhum benchmark é usado como alternativa às datas de recessão nos Estados Unidos fornecidas pelo NBER. Essas datas fornecem um termo de comparação em toda a literatura e são amplamente aceitas como o sinal a ser batido. Outras economias como a da Europa e do Japão também dispõe de séries históricas especificamente voltadas para identificar o que é consensualmente aceito como período de recessão econômica.

Dentro das bases de dados utilizadas como referência para esse trabalho, não foi encontrado nenhum indicador específico que classifique objetivamente os períodos de recessão, portanto, o critério de decisão para esse fenômeno será estabelecido pelo autor. Quanto às métricas utilizadas, tendem a variar de autor para autor, sendo utilizadas diferentes explicações.

As métricas mais utilizadas tendem a ser a curva Receiver Operating Characteristic (ROC) e a Area Under the Curve (AUC). Autores como Liu e Moench (2016) utilizam essas métricas para avaliar seus trabalhos, recuperando resultados de seus diversos testes com valores de AUC em torno de 0,8. Esta métrica não é usada apenas neste trabalho, mas também no trabalho recente de Wells Fargo Securities, com resultados de AUC em torno dos valores de 0,9, embora esses valores não pareçam se ajustar aos gráficos que representam.(SILVIA, 2018)

Trabalhos mais antigos, como Estrella e Mishkin e Baba e Kışınbay, usaram outros tipos de métricas para avaliar seus trabalhos. Estrella e Mishkin (1995) usa o Pseudo R^2 enquanto que Kisinbay e Baba (2011) utiliza o Quadratic Probability Score (QPS) e o Log Probability Score (LPS). Essas métricas não são muito utilizadas nos trabalhos mais recentes da área.

Esta revisão estabelece o estado da arte da detecção de recessão nos Estados Unidos, estabelecendo alguns benchmarks para esta tese e fornecendo algumas ideias para a implementação e validação do sistema. Apesar das diferenças em relação a economia Brasileira, as metodologias aplicadas nesses benchmarks podem ser facilmente transpostas para o problema proposto nesse trabalho.

2.2 Séries temporais

A análise, modelagem e previsão de séries temporais têm sido aplicadas a diversos problemas nas últimas décadas e têm sido objeto de diversos estudos e pesquisas. Seu objetivo é criar e desenvolver um modelo que possa descrever com precisão uma série ao longo do tempo e no futuro usando predições ou previsões. Esta previsão é geralmente descrita como o ato de prever o futuro olhando para o passado e compreendendo-o. Como tal, a análise de séries temporais se tornou muito importante nas áreas de engenharia, economia e finanças. Estes dois últimos também são considerados os mais desafiadores de se trabalhar, devido às suas características ruidosas, não estacionárias e deterministicamente caóticas, conforme discutido em Tay e Cao (2001).

Dada a importância da análise de séries temporais para a compreensão do comportamento dos sinais financeiros e econômicos, é fundamental entender seus componentes e como seus modelos funcionam na previsão de seus comportamentos futuros. Esta seção examinará apenas os componentes das séries temporais, usados principalmente em modelos univariados, deixando para os próximos capítulos as análises de modelos multivariados. Esta distinção é necessária para diferenciar a evolução e projeção de cada indicador econômico da utilização de vários sinais econômicos para tentar apurar um diferente, neste caso, um sinal de recessão.

As séries temporais podem ser definidas como uma sequência ordenada de valores de uma variável em intervalos de tempo igualmente espaçados (CROARKIN; TOBIAS, 2012). Podemos destacar 2 motivos principais para estudar modelos de séries temporais:

- Obter uma compreensão das forças e estrutura subjacentes que produziram os dados observados
- Ajustar um modelo e promover a previsão, monitoramento de feedback e controle de feedforward.

A análise de séries temporais pode ser dividida em duas categorias principais, dependendo do tipo de modelo que pode ser ajustado:

Modelo Cinético: Os dados são ajustados como $x_t = f(t)$. As medições ou observações são vistas em função do tempo.

Modelo Dinâmico: Os dados são ajustados como $x_t = f(x_{t-1}, x_{t-2}, x_{t-3} \cdots)$.

Como tantas séries temporais são não estacionárias, segundo Desikan e Srivastava (2005), é necessário entender que elas podem ser decompostas em quatro componentes principais:

Tendência: mudança monotônica no nível médio da série temporal.

Ciclo de troca: longas ondas, mais ou menos regulares, em torno de uma linha de tendência.

Sazonalidade: flutuação na série temporal que se repete durante períodos específicos de menos de um ano e geralmente causada por fatores como clima, férias ou feriados nacionais.

Resíduos: representa todas as influências nas séries temporais que não são explicadas pelos componentes anteriores.

Na prática geral, os dados são primeiro analisados e examinados, normalmente realizando correlações, autocorrelação ou mesmo análises básicas de plotagem, para descobrir esses quatro componentes. Esses processos são importantes porque fornecem informações úteis na escolha de quais modelos aplicar para atingir os objetivos propostos. Eles também demonstram quais transformações são necessárias para realizar nos dados para que possam ser usados pelos diferentes previsores. Algumas transformações geralmente são feitas para modelos univariados, como forçar a estacionaridade, removendo a tendência e a sazonalidade do sinal. Outros são mais típicos de modelos multivariados, como alterar a granularidade dos sinais, adicionar lag ou normalizar os dados.

2.3 Classificação

Converter o problema da detecção de recessão econômica em um problema matemático e probabilístico requer um exame dos resultados possíveis e favoráveis. É bastante natural concluir que existem apenas dois resultados possíveis disponíveis, ou há uma recessão econômica ou não. Isso leva à postulação de que esse problema não é apenas uma questão de classificação, ou seja, o resultado pertence a uma categoria específica, mas também é binário, uma vez que existem apenas duas categorias possíveis às quais os resultados podem ser atribuídos. Esta dedução permite reduzir os modelos possíveis a aplicar aos indicadores disponíveis.

Como esse é um conceito com definição abstrata na economia alvo desse estudo, optou-se por uma abordagem um pouco mais abrangente, com o intuito de não restringir o conceito. Considerando as componentes principais das séries temporais, o objetivo principal é propor um método que consiga identificar antecipadamente os momentos em que ocorrer uma quebra de tendência ou então uma movimentação brusca que desrespeite a sazonalidade.

O primeiro passo é entender quais informações estão disponíveis e quais estão faltando. A partir das análises econômicas, pode-se estabelecer que os dados de entrada e saída estão facilmente disponíveis, o que significa que a entrada, os indicadores econômicos, possuem valores bem definidos e não precisam ser inferidos. Com relação aos dados de saída, foi feita uma interpretação baseada em critérios matemáticos.

No entanto, os capítulos financeiros também fornecem o primeiro obstáculo, a transformação ou correlação entre a entrada e a saída. Esse conhecimento é importante ao construir modelos de previsão, uma vez que a saída é uma variável desconhecida e a única informação disponível é a entrada.

Para inferir sobre o processo de transformação de uma entrada em uma saída, uma das principais áreas de estudo é o Aprendizado de Máquina (ML), como apresentando por Alpaydin (2014). Principalmente as técnicas de aprendizado de máquina utilizam os dados disponíveis para inferir sobre algumas outras informações, que no estudo desta tese se apresentam como uma boa solução para descobrir datas de recessão com base em sinais econômicos.

Modelos de classificação binária de aprendizado de máquina tendem a conceber uma regra, a partir dos dados disponíveis, principalmente por meio de uma análise multivariada, que estabelece quando uma determinada informação resultaria em zero ou uma resposta. Nesta tese, o resultado poderia representar:

- 1 - recessão
- 0 - sem recessão

Para atingir este objetivo podem ser utilizadas duas abordagens multivariadas principais, a abordagem linear e a não linear. No entanto, como estamos trabalhando com um conceito mais amplo de estados possíveis, vamos estender o conceito de classificação binária para a classificação n-ária, já que podemos ter mais de 2 estados possíveis para

classificar. Portanto, o conceito de classificação binária será estendido para diversas classes, utilizando a ideia de classificação multiclasse One-vs-all.

2.3.1 Regressão linear

Um modelo é definido como linear quando cada um de seus termos é uma constante ou o produto de um parâmetro com uma variável preditora. Esses modelos tendem a seguir uma equação da forma,

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{1}$$

onde β_0 é uma constante, β é um parâmetro e x é uma variável preditora. Aparentemente, pode parecer que os modelos lineares são incapazes de ajustar curvas, mas esse não é o caso. Uma vez que a linearidade é fixada nos parâmetros e não nas variáveis preditoras, bastaria incluir uma variável de natureza não linear, como por exemplo:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^1 + \beta_4 x_1 x_2 \tag{2}$$

e ainda assim ser linear (NELDER; WEDDERBURN, 1972). Isso demonstra que os modelos lineares podem ser aplicados a um problema de classificação binária mesmo com algumas modificações. Segue uma definição mais formal para o conceito apresentado por Heij *et al.* (2004):

Para uma variável dependente binária, o modo de regressão

$$y_i = x_i'\beta + \varepsilon_i = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i,\ E[\varepsilon_i] = 0 \tag{3}$$

é chamado de modelo de probabilidade linear. Como $E[\varepsilon_i] = 0$ e y_i podem assumir apenas os valores zero e um, segue-se que $x_i'\beta = E[y_i] = 0 \cdot P[y_i = 0] + 1 \cdot P[y_i = 1]$ de modo que

$$P[y_i = 1] = E[y_i] = x_i'\beta \tag{4}$$

2.3.2 Regressão logística

As probabilidades podem ser limitadas a valores entre zero e um usando um modelo não linear. Seja F uma função com valores que variam entre zero e um, e seja

$$P[y_i = 1] = F(x_i'\beta) \tag{5}$$

Para facilitar a interpretação deste modelo, a função F é sempre considerada monotonicamente não decrescente. Neste caso, se $b_j > 0$, então um incremento em x_{ji} leva a um incremento (ou pelo menos não a um decremento) da probabilidade de $y_i = 1$. Ou seja, coeficientes positivos (negativos) correspondem a efeitos positivos (negativos) sobre a probabilidade de sucesso. Uma escolha óbvia para a função F é uma função de distribuição cumulativa, como pode ser visto na figura 1

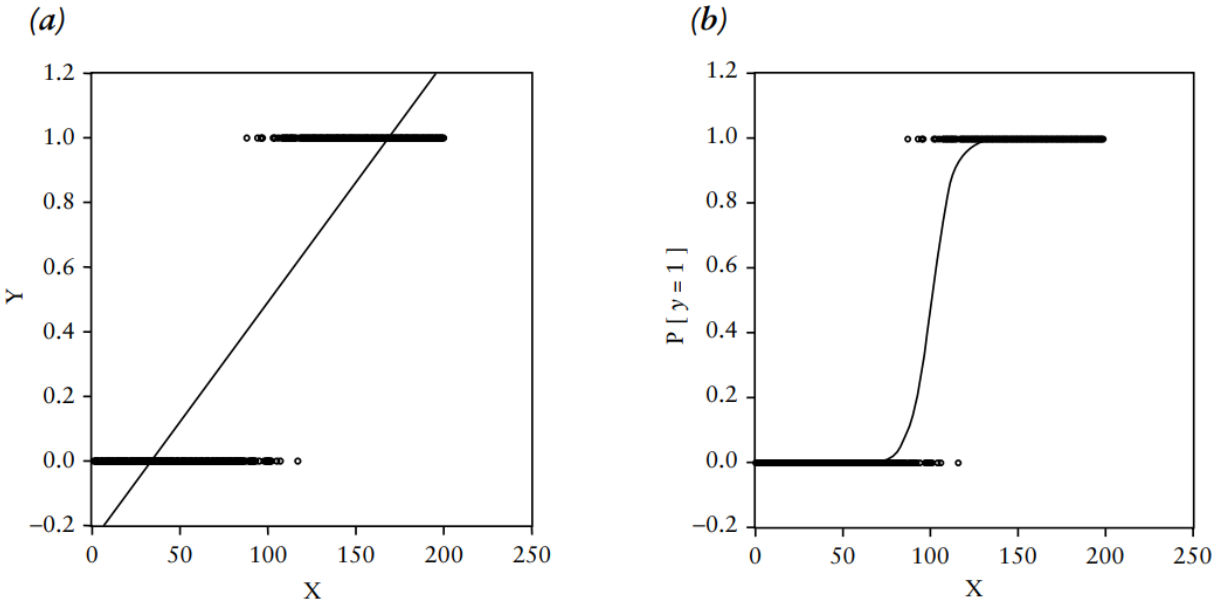

Figura 1 – Variável dependente binária (y assume valor 0 ou 1) com modelo de probabilidade linear (a) e com modelo de probabilidade não linear em termos de uma função de distribuição cumulativa (b), para uma única variável explicativa (x).

O modelo 5 depende não apenas da escolha das variáveis explicativas x, mas também da forma da função de distribuição F. Esta escolha corresponde a assumir uma distribuição específica para os efeitos individuais não observados (na função índice ou nas utilidades) e determina a forma da função de resposta marginal por meio da função de densidade f correspondente. Na prática, muitas vezes escolhe-se a densidade normal padrão

$$f(t) = \phi(t) = \frac{1}{\sqrt{2}} e^{-\frac{1}{2}t^2}$$

ou a densidade logisitica

$$f(t) = \lambda(t) = \frac{e^t}{(1+e^t)^2}$$

O modelo 5 com a distribuição normal padrão é denominado modelo probit, e o modelo com distribuição logística denomina-se modelo logit. Uma vantagem do modelo

logit é que a função de distribuição cumulativa $F = \Lambda$ pode ser calculada explicitamente, como

$$\Lambda(t) = \int_{-\infty}^{t} \lambda(s)ds = \frac{e^t}{1+e^t} = \frac{1}{1+e^{-t}} \tag{6}$$

enquanto a função de distribuição cumulativa $F = \Phi$ do modelo probit deve ser calculada numericamente pela aproximação da integral

$$\Phi(t) = \int_{-\infty}^{t} \phi(s)ds = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}s^2} ds \tag{7}$$

Na prática, isso não apresenta problemas reais, pois existem algoritmos de integração numérica muito precisos. Em geral, as diferenças entre os dois modelos não são tão grandes, a menos que as caudas das distribuições sejam importantes.

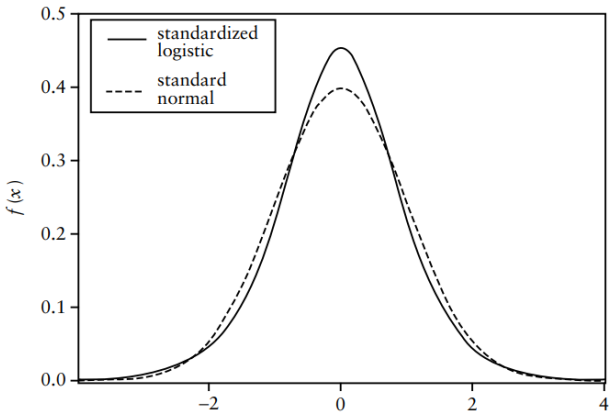

Figura 2 – Densidades da distribuição normal padrão (linha tracejada) e da distribuição logística (linha contínua, escalonada de forma que ambas as densidades tenham desvio padrão igual a 1). Em comparação com a densidade normal, a densidade logística apresenta valores maiores em torno da média ($x = 0$) e também em ambas as caudas (para valores de x distantes de 0).

2.4 Conceitos econômicos

Alguns conceitos básicos a respeito de economia são necessários para que se compreender não só o problema de pesquisa, mas também as soluções propostas nesse trabalho. Para iniciar essa discussão é necessária uma definição mais precisa do termo economia.

Economia representa o consumo e produção de recursos escassos em uma determinada área, relativos a bens e serviços, que existem para satisfazer as necessidades de seus participantes. Existem vários tipos de economias que se distinguem principalmente pela propriedade privada ou social dos meios de produção e pelo mercado ou pela decisão de planejamento da alocação de recursos. Para o propósito deste estudo, segundo os conceitos

definidos por Rosser J. Barkley e Rosser (2018), é importante esclarecer que a economia dos Brasil é baseada no capitalismo de mercado, ou seja, é baseada na propriedade privada, com um método de decisão de mercado para alocação de recursos. Em outras palavras, produtores e consumidores determinam o que é produzido e vendido. Os produtores são os proprietários dos produtos e definem seu preço, e os consumidores são os proprietários do que compram e decidem quanto desejam pagar.

Essas economias e seus fatores que afetam podem ser analisados e estudados por uma ciência social chamada economia, que pode ser dividida em duas subdivisões principais, microeconomia e macroeconomia. A microeconomia estuda o comportamento de famílias e empresas, e sua decisão sobre o que comprar e produzir, respectivamente, e as quantidades compradas e produzidas. Por outro lado, a macroeconomia lida com a economia como um todo, examinando fatores mais amplos como o consumo nacional, a produção nacional, o nível geral de preços ou mesmo o desemprego (AGARWALA, 2009). Para uma análise das tendências econômicas globais, torna-se claro que um estudo macroeconômico é necessário, sendo, portanto, objeto de estudo desta tese.

2.4.1 Macroenonomia

A macroeconomia é um dos estudo fundamentais para a compreensão da estrutura e do desempenho de uma economia. Ele usa as despesas e consumos associados a uma nação ou região, a quantidade economizada e gasta por todas as famílias ou mesmo a produtividade do trabalho de um país como base para sua análise. Como discutido em Rittenberg e Tregarthen (2009), a maioria dos economistas usa o Produto Interno Bruto (PIB), uma medida da produção total, para determinar se a economia de uma nação está crescendo ou encolhendo. Mas não apenas o PIB é usado para avaliar o estado da economia, como pode ser visto em OECD... (2018). Neste trabalho, é verificado que outros indicadores como crescimento industrial, fabricação de novos pedidos e crescimento do volume de vendas no varejo também podem ser usados para essa finalidade. Isso nos fornece a primeira postulação básica de que os estudos macroeconômicos atuais utilizam diversos modelos e indicadores, relacionados à produção, trabalho ou negócios, para analisar seu objeto de estudo.

A macroeconomia também introduz o conceito de ciclos econômicos, para descrever as flutuações no crescimento econômico ao longo do tempo, com base em expansões e contrações (recessões) da economia, conforme descrito mais adiante. Desse modo, fica claro que a utilização de indicadores macroeconômicos é um requisito para analisar a economia e suas variações, tornando-os indispensáveis para a resolução do trabalho proposto nesta tese.

2.4.2 Ciclos econômicos

Como afirmado anteriormente, um ciclo econômico descreve a tendência e sua variação da economia. Uma abordagem mais específica para definir um ciclo econômico foi proposta por Burns e Mitchell (1946) e ainda é usada hoje:

> "Os ciclos econômicos são um tipo de flutuação encontrada na atividade econômica agregada das nações que organizam seu trabalho principalmente em empresas: um ciclo consiste em expansões ocorrendo quase ao mesmo tempo em muitas atividades econômicas, seguidas por recessões gerais semelhantes, contrações e avivamentos que se fundem na fase de expansão do próximo ciclo; esta sequência de mudanças é recorrente, mas não periódica; na duração os ciclos econômicos variam de mais de um ano a dez ou doze anos; eles não são divisíveis em ciclos mais curtos de caráter semelhante com amplitudes que se aproximam de si mesmos."

Embora seja uma definição ampla e a maioria dos economistas tenda a usar análises mais específicas, geralmente medindo a tendência do PIB, como o National Bureau of Economic Research (NBER) que utiliza uma variedade mais ampla de parâmetros para definir cada ciclo (NBER, 2010). Esta informação é importante porque o NBER tem sido amplamente aceito como a autoridade líder tanto na definição das datas do ciclo de negócios quanto na pesquisa econômica por trás disso, fornecendo um padrão nesta área de especialização.

Considerando essa explicação, a economia só pode flutuar entre dois estados distintos, expansão ou recessão, simplificados como crescimento econômico ou retração econômica, respectivamente. Isso leva à conclusão de que durante uma expansão, os indicadores macroeconômicos, como PIB, emprego ou produção, tendem a crescer e durante uma recessão, eles tendem a diminuir. Portanto, é seguro supor que a evolução dos indicadores macroeconômicos, ao longo de um período de tempo, pode descrever não só o estado atual da economia, mas também em que fase do ciclo ela se encontra atualmente.

2.4.3 Ponto de virada do ciclo

O comportamento cíclico da economia tem sido estudado há vários anos por diversos economistas e instituições, mas como todos os estudos do comportamento cíclico, seus interesses se concentram principalmente nos pontos de inversão, não nas fases estáveis. Normalmente, após esses períodos de mudança que medidas corretivas críticas e diferentes métodos econômicos costumam ser aplicados, resultando ocasionalmente em mudanças estruturais na economia, geralmente após uma recessão. Portanto, a previsão de um ponto de inversão levando a uma fase de recessão é um dos objetivos mais cobiçados da economia.

Essas mudanças na economia tendem a ocorrer após uma recessão, devido aos seus efeitos prejudiciais para a sociedade. As recessões pesam muito nas famílias, principalmente devido ao desemprego, perda de poder de compra, declínio nas condições de bem-estar e educação, que em última instância tendem a levar a mortes prematuras, suicídios, depressões, baixo rendimento escolar, pobreza infantil e diminuição das taxas de natalidade, como discutido em Adrian e Education (2010).

A combinação desses eventos negativos tende a mudar os hábitos das famílias, tornando-os mais instruídos economicamente sobre as causas da recessão anterior e capazes de lidar com perdas de renda e patrimônio no futuro (O'NEILL; XIAO, 2012). Isso também é verdade para empresas e governos, que também devem se adaptar a essas mudanças na estrutura da economia.

Para detectar esses pontos de inflexão, e de acordo com as evidências anteriores, os economistas usam um conjunto de indicadores macroeconômicos e métodos de avaliação. Nesse sentido, é seguro assumir que a variação desses indicadores, relacionados ao consumo, produção ou emprego, pode ser usada para descrever e identificar a fase atual do ciclo econômico e um possível ponto de inflexão. Como discutido em Advisors (2017), o conjunto de indicadores relacionados a essas áreas é vasto e apresenta diversos desafios, como sinais contraditórios, erros de amostragem de pesquisas e o peso atribuído a cada indicador.

É claro que nenhum indicador é suficiente, e mesmo uma combinação deles pode ser insuficiente para avaliar o estado da economia. No entanto, existem alguns indicadores que surgem em numerosa literatura como preditores muito fortes, como a curva de rendimento, pedidos iniciais de seguro-desemprego, novas licenças de habitação, ganhos mensais de emprego ou produção industrial. Os trabalhos a seguir referem-se as discussões mais

relevantes a respeito dessa diversidade de indicadores. (ESTRELLA; MISHKIN, 1995; KAUPPI; SAIKKONEN, 2008; RUDEBUSCH; WILLIAMS, 2009; LIU; MOENCH, 2016; KISINBAY; BABA, 2011; BERGE, 2013)

Para detectar esses pontos de inflexão, os atuais economistas e pesquisadores utilizam principalmente modelos probabilísticos, pois são atualmente as melhores ferramentas de previsão disponíveis. Os modelos probit, um tipo específico de modelos binários de classificação, são os modelos probabilísticos mais comumente usados para estimar a probabilidade de uma observação se enquadrar em uma das duas categorias disponíveis, neste caso, recessão ou não recessão. Esses modelos serão discutidos em capítulos posteriores.

Em conclusão, a detecção de pontos de inflexão nos ciclos de negócios é feita por meio de diversos indicadores macroeconômicos, principalmente nas áreas de consumo, produção e emprego, e a variação desses sinais. Os métodos usados para atingir essas previsões são baseados em modelos probabilísticos com ênfase em modelos de classificação binários. A maior parte da comunidade econômica usa essas avaliações de pontos de inflexão para detectar recessões devido aos seus efeitos sobre a economia. Essas afirmações e as dos subcapítulos anteriores definem o contexto econômico mais fundamental usado nesta tese.

3 Arquitetura

Introdução

Este capítulo apresenta a arquitetura proposta para classificação dos ciclos de mercado da economia brasileira, utilizando algoritmos de machine learning com dados de indicadores macroeconômicos. O objetivo é discutir as etapas de coleta e tratamento dos dados, bem como apresentar e comparar as diferentes técnicas e sua performance em diferentes cenários.

A arquitetura proposta para o problema foi pensada de forma modular, de modo a facilitar futuras melhorias específicas.

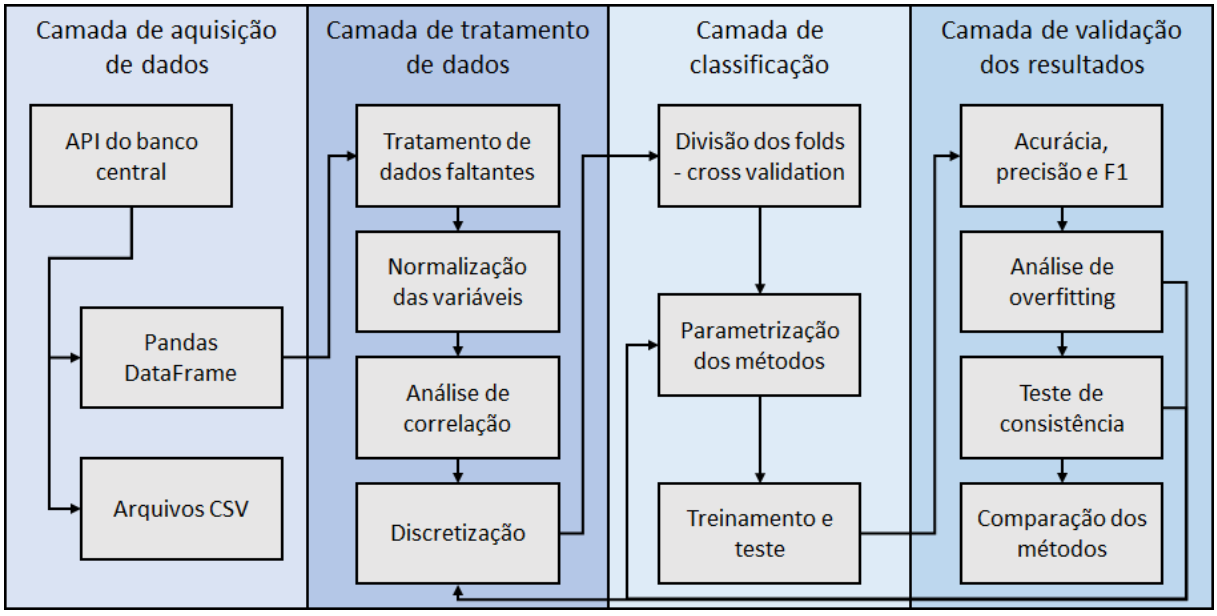

Figura 3 – Arquitetura proposta para o problema

A figura 3 mostra as camadas de implementação, fornecendo uma divisão lógica desde aquisição até a validação dos resultados. Essa abordagem facilita a compreensão geral do problema, mas ainda preserva abstração suficiente para inserir subcamadas.

É importante ressaltar que as realimentações contribuem na eficácia de forma dinâmica, além de ajudas a distinguir quando temos limitações do método ou quando temos uma concepção equivocada em alguma camada anterior.

3.1 Camada de aquisição de dados

A camada de aquisição consiste em uma função que acessa o site do banco central e retorna a série histórica escolhida, em formato de dataframe. Além disso, foi implementada uma função que exporta essas informações para um backup em CSV.

3.1.1 Função aquisição

A função aquisição foi implementada baseada em uma API do banco central, que retorna a série escolhida dentro de um intervalo de tempo pré-determinado.

$$F(codigo\_bcb, data\_inicial, data\_final) \tag{8}$$

Como podemos verificar, a função possuí 3 inputs que devem ser definidos manualmente. O tamanho do output vai depender da granularidade temporal da série, que pode variar bastante (desde trimestral até diária).

3.1.2 Arquivos CSV

Toda a arquitetura está estruturada utilizando os dataframes Pandas, já que foi implementado através de uma aplicação online. No entanto, muitas vezes se torna abstrato analisar os dados nesse formato, devido às limitações da interface do Google Colab.

Com o intuito de manter uma cópia dos dados (para o caso de alguma série ser descontinuada), ou ainda conseguir analisar os dados utilizando diferentes softwares, foi implementada uma função que exporta as séries em um arquivo CSV para ser armazenado localmente.

3.2 Camada de tratamento de dados

A camada de tratamento de dados é a parte mais sensível do projeto, uma vez que as escolhas tomadas nessa etapa causam grande impacto na previsão final. A qualidade da previsão depende muito essencialmente da qualidade dos dados, ou seja, mesmo uma boa técnica vai trazer resultados ruins se os dados forem tratados de forma inadequada.

3.2.1 Dados faltantes

Quando lidamos com séries de dados temporais, é comum a existência de dados incompletos ou preenchidos de forma inadequada. No entanto, como os indicadores utilizados são compilados oficialmente, esse problema é praticamente nulo.

Ainda assim, algumas séries são descontinuadas com o tempo, seja por obsolescência ou por mudanças na metodologia. Nesse segundo caso, as séries costumam ganhar outra numeração e escala, tornando o conteúdo pretérito não comparável. Essa característica afetou significamente o trabalho, tornando necessário remover alguns dos indicadores com maior grau de correlação com a variável alvo.

3.2.2 Normalização das variáveis

Os indicadores utilizados derivam de diversos agentes e, além disso, se referem a fenômenos diferentes. Podemos destacar 4 categorias diferentes de indicadores utilizados:

- Preço;
- Atividade econômica;
- Confiança;
- Monetário.

Dessa forma, eles possuem diversidade no que se refere a alguns aspectos:

- Unidades de medida;
- Período de análise;
- Granularidade de tempo;

3.2.2.1 Unidades de medida:

As variáveis que representam índices são disponiblizados em pontos (cada um em uma escala individual), enquanto as variáveis monetárias são cotadas em u.m.c (mil). Ainda existem outros indicadores que são dados em função da variação percentual mensal.

Para homogeinizar as análises, são propostas algumas transformações nas variáveis, de modo que elas sejam representadas em função de sua variação percentual mensal.

3.2.2.2 Período de análise:

Algumas variáveis possuem dados catalogados desde os anos 80, mas para homogeinizar os períodos, as medidas são consideradas à partir do início do indicador mais recente utilizado (janeiro de 2002). O período final é sempre a última medida vigente disponibilizada, mantendo as análises sempre atualizadas.

Uma das variáveis com maior correlação com o problema teve sua série descontinuada durante as pesquisas, e apesar da sua importância, decidiu-se por remove-la da análise.

3.2.2.3 Granularidade de tempo:

A periodicidade das variáveis possuí grande variabilidade, porém, como as variáveis que se mostraram mais importantes possuem intervalo mensal, optou-se por manter apenas as variáveis que tivessem essa janela de representação.

3.2.3 Análise de correlação

A correlação foi proposta com o objetivo de auxiliar na filtragem das variáveis. Para tanto, diversas rodadas de teste devem ser realizadas, combinando diferentes variáveis, enquanto que o critério de prioridade de cada variável basea-se principalmente na matriz de correlação de Pearson.

3.2.4 Discretização

A discretização é uma etapa bastante sensível do problema, tendo papel crucial na qualidade dos resultados. Como pode ser notado na figura 3, há uma realimentação partindo da última camada, já que é possível inferir problemas na discretização através da análise de overfitting.

Portanto, é destinada uma seção especificamente para tratar dos detalhes desta etapa.

3.3 Camada de classificação

A camada de classificação envolve os passos relativos a divisão do conjunto de treinamento e teste, bem como a escolha e parametrização dos métodos utilizados para classificação.

3.3.1 Conjunto de treinamento - cross validation

A etapa de treinamento tipicamente está muito suscetível aos fenômenos de underfitting/overfitting. Quando a quantidade de dados não é muito grande, esse problema pode ser agravado.

Como a periodicidade dos dados escolhida foi mensal e alguns indicadores só possuem dados a partir de 2002, pode-se considerar que essa é uma base de dados de tamanho modesto para essa classe de problemas.

Por conta disso, torna-se fundamental implementar a metodologia de cross validation, cujos detalhes serão explicados em uma seção posterior.

3.3.2 Parametrização dos métodos

A parametrização dos métodos escolhidos também é muito importante para melhorar os resultados de teste dentro das métricas estabelecidas, bem como contornar problemas de convergência e overfitting.

Essa etapa trouxe algumas surpresas interessantes, o que acabou ampliando bastante a compreensão do problema como um todo. Alguns métodos apresentaram muita sensibilidade à discretização ao ponto de não convergirem, no entanto, quando convergiram apresentaram os melhores resultados. Isso reforça a ideia de que a alta não-linearidade presente no fenômeno deixa o problema muito dependente de uma boa modelagem.

Como pode ser verificado na figura 3, essa etapa também é realimentada pela camada de validação, uma vez que os parâmetros tem que ser ajustados por tentativa e erro a partir da observação de resultados. O nível de detalhamento para tornar essa discussão clara, exige uma discussão que será feira em seção posterior.

3.3.3 Treinamento e teste

As etapas de treinamento e teste foram realizadas utilizando diferentes métodos que possuem características distintas, com o objetivo de se avaliar os seguintes aspectos:

- Precisão/acurácia de classificação;
- Consistência de resultados;
- Custo computacional;
- Interpretabilidade.

Os algoritmos de aprendizagem automática notadamente têm conquistado ótimos resultados em diversos seguimentos da ciência, auxiliando na predição de fenômenos complexos. No entanto, em muitas situações, esses bons resultados dependem se uma base de dados imensa e de muito custo computacional, o que tornaria a aplicação pouco útil para aplicações em tempo real.

Além disso, técnicas muito abstratas podem alcançar excelentes resultados sem oferecer a possibilidade de uma interpretação humana, o que acaba tornando impraticável uma adesão massiva já que os sistemas ficariam reféns de possíveis cenários não previstos. Por conta disso, optou-se por incluir técnicas que, mesmo sendo simples e supostamente menos eficientes, possuem uma interpretabilidade mais natural, com o objetivo de tentar explicar o fenômeno com mais facilidade. Todas essas especificidades serão abordadas com mais profundidade em seções posteriores.

3.4 Camada de validação dos resultados

Essa camada tem como principal objetivo verificar se os resultados obtidos pelos métodos aplicados possuem qualidade e consistência.

3.4.1 Métricas de avaliação

As métricas quantitativas utilizadas derivam da matriz de confusão, e a discussão detalhada será aprofundada em seção posterior.

- Acurácia;

- Precisão;
- Revocação;
- F1-Score.

Apesar de outros aspectos qualitativos também serem passíveis de discussão (como custo computacional e interpretabilidade), a literatura não dispõe de uma metodologia objetiva para quantificar e comparar esses critérios. Dessa forma, ainda que representem aspectos importantes da análise, essa discussão não será o foco deste trabalho, ainda que possam ser utilizados como eventual critério de desempate.

3.4.2 Análise de overfitting

A análise de overfitting é realizada comparando os resultados obtidos no conjunto de treinamento e de teste. Quando há uma diferença significativa entre essas medidas, o método está se especializando muito em relação aos dados de treinamento e não consegue compreender bem dados novos.

Existem diversos fatores que podem influênciar esse aspecto, mas especificamente para esse problema, percebeu-se que a discretização e a paramtrização dos métodos são os fatores que mais influenciam. Dessa forma, é proposta uma realimentação no diagrama 3, com o objetivo de otimizar essas etapas.

3.4.3 Teste de consistência

Quando os métodos não estão bem adaptados à modelagem, é possível que ocorra variabilidade nos resultados quando testados de forma recorrente, devido à influência de pequenos vieses.

Para contornar esse problema, tanto o treinamento quanto os testes são repetidos várias vezes, enquanto que os resultados apresentados nos capítulos posteriores referem-se à média de todas as observações. Ainda assim, só devem ser considerados aceitáveis os métodos que possuem baixa variância nesse aspecto.

4 Implementação

Com a arquitetura do problema devidamente definida, desde os processos isolados até as conexões lógicas entre as partes, o objetivo dessa seção é apresentar as etapas de implementação das etapas citadas.

Toda a análise foi feita utilizando códigos escritos em Python 3.7 através da IDE (ambiente de desenvolvimento integrado) do Google Colab. A escolha por essa ferramenta se deu pela facilidade de testar alguns trechos de código separadamente, uma vez que sua arquitetura permite os testes de código em bloco, evitando ter de rodar repetidamente códigos que envolviam download de grandes quantidades de dados, o que que além de gastar tempo poderia incorrer em bloqueios de acesso pela API. Além disso, sua integração com os serviços Google permitiu que os testes fossem feitos sem a necessidade de instalação local do software, permitindo análises pelo celular mesmo em contextos de baixa disponibilidade de acesso à internet, além de tornar a versionamento e compartilhamento com outros colaboradores, bastante natural.

4.1 Seleção de indicadores

A etapa de escolha de indicadores é de fundamental importância, no entanto, não existe um consenso na literatura de como fazer isso. Comparando diversos estudos realizados em diferentes países, é notável perceber que cada economia é melhor representada por certos parâmetros e segmentos que possuem correlação com características intrínsecas de suas atividades econômicas. Em outras palavras, não há uma metodologia para tal, de modo que a escolha tem que seguir alguns critérios subjetivos.

A abordagem mais comum consiste em incluir os principais indicadores relativos a diversos seguimentos da economia, e filtra-los posteriormente através de estudos de correlação e/ou estatísticos. Dessa forma, tomou-se como referência os principais indicadores amplamente utilizados em outros estudos, e buscou-se um equivalente na economia brasileira. Cabe ressaltar que além dos indicadores sem equivalente direto, muitos indicadores importantes sofreram mudanças de metodologia e suas séries foi descontinuada e reiniciada com outras métricas, tornando inviável uma conexão que mantivesse a continuidade.

Os indicadores apresentados abaixo foram utilizados nas primeiras etapas de análise, como candidatos ao modelo definitivo.

- Preço

  PIB: Produto Interno Bruto mensal - Valores correntes (R$ milhões)

  IBOV Valor das empresas listadas na Bovespa

  IPA Índice de preços ao produtor amplo

  IPEM Indicador da produção - extrativa mineral

  IPIT Indicadores da produção - indústria de transformação

  IPBC Indicadores da produção - bens de capital

  IPBCD Indicadores da produção - bens de consumo duráveis

- Atividade Econômica

  IVVV Índice volume de vendas no varejo - Automóveis, motocicletas, partes e peças - Brasil

  VVCCL Vendas de veículos pelas concessionárias - Comerciais leves

  VVCC Vendas de veículos pelas concessionárias - Caminhões

- Confiança

  IEF Índice de Expectativas Futuras

  ICC Índice de Confiança do Consumidor

- Monetário

  Spub Saldos das operações de crédito das instituições financeiras sob controle público - Total

  Spriv Saldos das operações de crédito das instituições financeiras sob controle privado - Total

  M1 Meios de pagamento - M1 (média dos dias úteis do mês) - Novo

  M2 Meios de pagamento - M2 (média dos dias úteis do mês) - Novo

Os nomes foram expressos de forma completa como descrito nas bases de dados do BCB, porém, foram propostas algumas siglas para facilitar a análise e interpretação das próximas etapas. A tabela 1 sintetiza as siglas propostas e demais informações relevantes

a respeito das variáveis propostas, que foram fundamentais para a etapa de tratamento das variáveis.

Tabela 1 – Metadados dos indicadores macroeconômicos da base do SGS-BCB

Indicador	Código	Categoria	Início	Unidade	Fonte
PIB	4380	Preço	jan/90	R$ (milhões)	BCB-DEPEC
IBOV	7849	Preço	jan/96	u.m.c. (milhões)	BM&FBOVESPA
IPA	7450	Preço	set/94	Var. % mensal	FGV
IPEM	21861	Preço	jan/02	Índice	IBGE
IPIT	21862	Preço	jan/02	Índice	IBGE
IPBC	21863	Preço	jan/02	Índice	IBGE
IPBCD	21866	Preço	jan/02	Índice	IBGE
IVVV	1548	Atividade	jan/00	Índice	IBGE
VVCCL	7385	Atividade	jan/90	Unidades	FENABRAVE
VVCC	7386	Atividade	jan/90	Unidades	FENABRAVE
IEF	4395	Confiança	mar/99	Índice	Fecomercio
ICC	4393	Confiança	mar/99	Índice	Fecomercio
Spub	2007	Monetário	jun/88	R$ (milhões)	BCB-DSTAT
Spriv	2043	Monetário	jun/88	R$ (milhões)	BCB-DSTAT
M1	27788	Monetário	dez/01	u.m.c. (milhões)	BCB-DSTAT
M2	27810	Monetário	jul/01	u.m.c. (milhões)	CNI

4.2 Preparação dos dados

4.2.1 Dados faltantes

Como pode ser observado na tabela 1 as séries propostas tem início em datas bastante distintas. Considerar uma maior quantidade de dados é quase sempre desejável, no entanto, nesse caso específico de usar como referência a série que tem início primeiro, causaria problemas referentes aos dados faltantes das outras séries e prejudicaria o treinamento do modelo.

Por conta disso, o período de referência considerado nesse estudo coincide com o início das séries mais jovens. Como todas as séries são completas, após sincronizar as datas de início, não houve problemas adicionais com dados faltantes.

4.2.2 Normalização das variáveis

O primeiro passo para conseguir comparar o comportamento das variáveis é transforma-las de modo que suas grandezas sejam comparáveis. Como o objetivo é analisar a variação percentual do PIB, decidiu-se utilizar a variação percentual como unidade de medida para todas as variáveis.

Foram realizados também alguns estudos utilizando a variação percentual acumulada, porém, os resultados não foram relevantes.

4.2.3 Análise de correlação

A análise de correlação primária utilizada se baseou na matriz dos coeficientes de Pearson. No entanto, esse não foi um critério de exclusão, uma vez que os métodos propostos podem ser capazes de capturar alguma correlação não-linear. Dessa forma, apesar da baixa correlação linear entre algumas variáveis propostas, ainda assim foram mantidas em alguns testes.

A utilização ou não de uma variável baseou-se em critérios de eficácia dos métodos propostos. Porém, mesmo em um fenômeno não-linear, a análise de correlação pode ser importante na interpretabilidade do modelo.

4.2.4 Discretização

Na etapa de discretização foram testados diferentes cenários que acabaram resultando em respostas bastante distintas. A discussão acerca desta diferença torna-se relevante uma vez que essa variabilidade aparentemente possuí um ponto de equilíbrio que pode ser interpretado de forma empírica. Isso acaba se tornando um elementos adicional, reforçando a idéia de que o problema foi modelado de forma correta.

A idéia principal é mapear cada valor do conjunto de dados através de um valor discreto pré-estabelecido. Dessa forma, o classificador seria treinado com valores pré-definidos das variáveis de entrada para identificar valores pré-definidos para a variável de saída. A quantidade de classes escohidas foram fator determinante na qualidade dos resultados obtidos, portanto, são representados à seguir os 3 cenários propostos.

4.2.4.1 Classificação binária

A primeira discretização proposta para o dataset foi buscando um modelo de classificação binária:

$$\begin{cases} x \mapsto 0, & \text{se } \Delta x < 0; \\ x \mapsto 1, & \text{se } \Delta x \geq 0. \end{cases} \tag{9}$$

Apesar de ser uma aproximação bastante grosseira, o objetivo desta etapa foi testar e validar alguns pedaços do código desenvolvido. Escolher a discretização mais simples possível, reduziu a complexidade de variáveis a se ajustarem ao código completo e ajudou a perceber outros problemas.

Apesar da aproximação grosseira, trouxe resultados bastante promissores que foram importantes para verificar a robustez de cada método em relação à discretização. No entanto, assim que o código foi validado, foi proposta uma nova discretização.

4.2.4.2 Classificação multiclasse - 3 classes

Buscando contornar os problemas encontrados na primeira discretização, foi proposto uma nova divisão do intervalo buscando separar movimentos de baixa, alta e lateralização. A principal diferença se refere ao critério de divisão intervalar baseada na média e no desvio padrão.

$$\begin{cases} x \mapsto -1, & \text{se } \Delta x \leq \mu_x - \sigma_x; \\ x \mapsto 0, & \text{se } \mu_x - \sigma_x < \Delta x < \mu_x + \sigma_x; \\ x \mapsto 1, & \text{se } \Delta x \geq \mu_x + \sigma_x. \end{cases} \tag{10}$$

O racional por traz dessa escolha intervalar parte da idéia de encontrar movimentos que tenham magnitude relevante em relação à média. Para tanto, é proposto o incremento de 1 desvio padrão, de modo que os casos de lateralização representem 68,27%. Com esse intervalo proposto, foram obtidos resultados bastante satisfatórios.

Outro fato relevante que cabe ressaltar é que as predições em relação aos cenários de crescimento implicaram em melhores resultados, quando comparados aos resultados de decrescimento. Isso reforça a ideia de que a modelagem está correta, pois coincide com o que ocorre na vida real, uma vez que movimentos de queda costumam ser mais súbitos e difíceis de se identificar previamente.

4.2.4.3 Classificação multiclasse - 5 classes

Com o objetivo de deixar o modelo ainda mais sofisticado, foram propostas outras discretizações baseando-se na idéia de separar movimentos fortes, fracos e neutros. Dessa forma, foram estabelecidos alguns intervalos que serão discutidos à seguir:

$$\begin{cases} x \mapsto -2, & \text{se } \Delta x \leq \mu_x - 2 \cdot \sigma_x; \\ x \mapsto -1, & \text{se } \mu_x - 2 \cdot \sigma_x < \Delta x \leq \mu_x - \sigma_x; \\ x \mapsto 0, & \text{se } \mu_x - \sigma_x < \Delta x < \mu_x + \sigma_x; \\ x \mapsto 1, & \text{se } \mu_x + \sigma_x \leq \Delta x < \mu_x + 2 \cdot \sigma_x; \\ x \mapsto 2, & \text{se } \Delta x \geq \mu_x + 2 \cdot \sigma_x. \end{cases} \tag{11}$$

O primeiro intervalo sugerido baseou-se na idéia de incluir sub-intervalos ainda mais incomuns, além dos 95% dos dados centrais. Essa abordagem apresentou um desempenho preditivo muito pior em relação as abordagens anteriores. Como os intervalos mais extremos representam eventos bastante raros, algumas das técnicas utilizadas nem sequer conseguiam prever certas classes mais raras. Além disso, como algumas variáveis são relativamente desbalanceadas em relação à média, isso também pode ter influenciado negativamente nos resultados.

Buscando ajustar o intervalo e minimizar esse efeito das classes muito raras, foi proposto um novo intervalo:

$$\begin{cases} x \mapsto -2, & \text{se } \Delta x \leq \mu_x - 1,6745 \cdot \sigma_x; \\ x \mapsto -1, & \text{se } \mu_x - 1,6745 \cdot \sigma_x < \Delta x \leq \mu_x - 0,6745 \cdot \sigma_x; \\ x \mapsto 0, & \text{se } \mu_x - 0,6745 \cdot \sigma_x < \Delta x < \mu_x + 0,6745 \cdot \sigma_x; \\ x \mapsto 1, & \text{se } \mu_x + 0,6745 \cdot \sigma_x \leq \Delta x < \mu_x + 1,6745 \cdot \sigma_x; \\ x \mapsto 2, & \text{se } \Delta x \geq \mu_x + 1,6745 \cdot \sigma_x. \end{cases} \tag{12}$$

Nesse novo intervalo proposto, a classe intermediária definida como lateralização ($\mu_x - 0,6745 \cdot \sigma_x < \Delta x < \mu_x + 0,6745 \cdot \sigma_x$), representa 50% dos dados centrais. Somando-se as classes relativas a alta e queda leve, temos um total de 90% em torno dos dados centrais, o que faz com que as classes limite se tornem bem menos raras. Isso foi suficiente tanto para evitar que algumas técnicas de predição falhassem, como também para melhorar a qualidade das predições.

Mesmo com essa mudança, os resultados ainda representam uma piora significativa quando comparada a abordagem utilizando 3 classes. Alguns motivos propostos para tentar explicar essa piora são:

- Maior complexidade do intervalo de dados
- Quantidade de classes inadequada ao fenômeno analisado
- Valores inadequados para os intervalos

Devido à baixa eficiência de predição e à dificuldade de se escolher um intervalo adequado, optou-se por uma abordagem envolvendo menos classes e cuja escolha do intervalo seguisse um critério mais homogenio. Portanto, o foco do trabalho será aboradar o problema utilizando apenas 3 classes.

4.3 Classificação

Para realizar um estudo exploratório das recessões econômicas e também poder prevê-las, as transformações nos sinais são muito importantes. As transformações de sinais macroeconômicos tentam diminuir os problemas de usar um conjunto de dados pequeno, mas também fornecem algumas informações econômicas relevantes sobre a importância dos pontos de dados anteriores para a detecção de recessão econômica.

Com a implementação dessas soluções de algoritmos de aprendizado de máquina, geralmente surgem alguns problemas de baixo desempenho devido a overfitting ou underfitting. Como o conjunto de dados tem muito poucas informações, é provável que isso aconteça, mesmo com a análise da matriz de correlação para remover alguns sinais que provavelmente produziriam análises excessivamente ajustadas. Para neutralizar essa probabilidade, um sistema de validação cruzada foi implementado no Algoritmo de Classificação.

4.3.1 Validação cruzada

O princípio básico por trás de um sistema de validação cruzada é dividir os dados em conjuntos de treinamento e conjuntos de teste. Ao fazer isso, os modelos não ajustam todos os dados em uma única etapa de treinamento, impedindo-os de ajuste excessivo ou insuficiente.

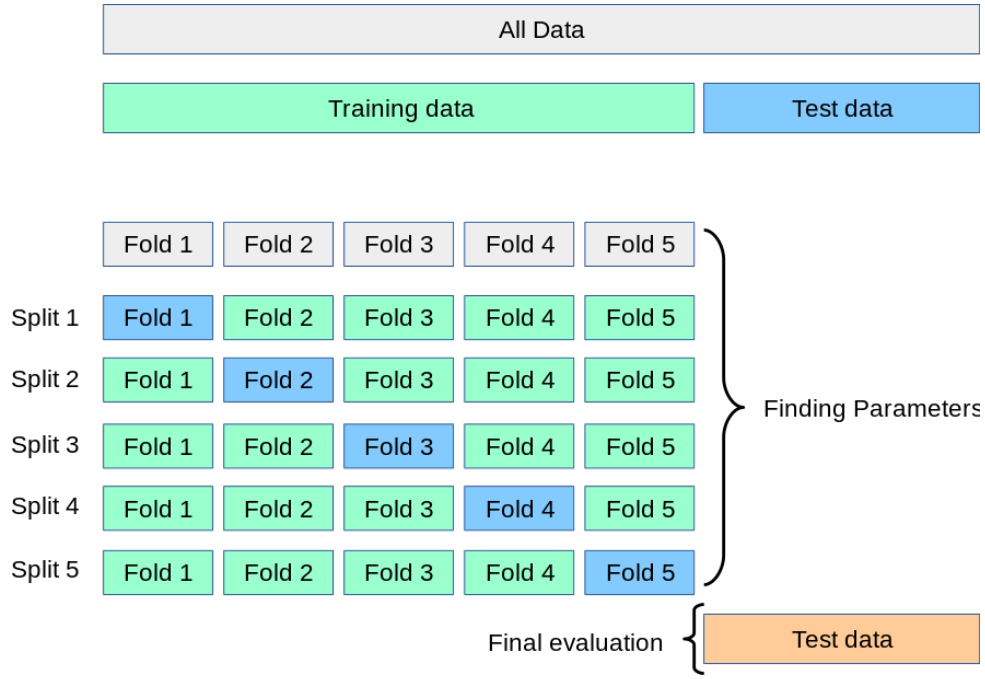

Figura 4 – Estrutura do processo de validação cruzada baseada em k-folds

O método baseado em k-fold divide os dados em k conjuntos iguais e realiza k treinamentos, onde cada uma das k etapas desconsidera o conjunto de testes da vez durante o treinamento. Após concluir uma etapa de treinamento, o conjunto de teste que foi retirado é usado para validar a eficácia do treinamento realizado. Repetindo esse procedimento k vezes, cada subconjunto do conjunto de dados terá sido utilizado em uma etapa diferente de validação, inibindo dessa forma o risco de viés inerente a um subconjunto específico do conjunto de dados.

4.3.2 Métodos de classificação

4.3.2.1 Nearest Neighbors

O princípio por trás dos métodos Nearest Neighbors é encontrar um número predefinido de amostras de treinamento mais próximas de distância para o novo ponto e prever o rótulo destes. O número de amostras pode ser uma constante definida pelo usuário (aprendizagem entre os k vizinho mais próximos) ou varia com base na densidade local de pontos (aprendizado baseado nos vizinho em um certo raio). A distância pode, em geral, ser qualquer medida métrica: distância padrão eclidiana é a escolha mais comum. Os

métodos baseados em vizinhos são conhecidos como métodos de aprendizagem de máquina não generalizada, uma vez que simplesmente retomam todos os seus dados de treinamento.

A classificação baseada em vizinhos é um tipo de aprendizado baseado em instância ou aprendizado de não-generalização: não tenta construir um modelo interno geral, mas simplesmente armazena instâncias dos dados de treinamento. A classificação é calculada a partir de uma maioria simples votação dos vizinhos mais próximos de cada ponto: um ponto de consulta é atribuído a classe de dados que tem mais representantes dentro dos vizinhos mais próximos do ponto.

4.3.2.2 Naive Bayes

Os métodos Naive Bayes são um conjunto de algoritmos de aprendizagem supervisionada baseados na aplicação do teorema de Bayes com a suposição "ingênua"de independência condicional entre cada par de características dado o valor da variável de classe. O teorema de Bayes afirma que, dada a variável de classe y e vetor dependente dos features x_1 até x_n:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots, x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Usando a premissa ingênua de independência condicional de que

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y)$$

para todo i, essa relação é simplificada para

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Desde que $P\left(x_1, \cdots, x_n\right)$ seja constante dada a entrada, podemos usar a seguinte regra de classificação:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\therefore$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

e podemos usar a estimativa máxima de posteriori (mapa) para estimar $P(y)$ e $P\left(x_i|y\right)$; O primeiro é então a frequência relativa da classe y no conjunto de treinamento.

Apesar de suas suposições aparentemente simplificadas demais, os classificadores Bayes ingênuos funcionaram muito bem em muitas situações do mundo real, como a famosa classificação de documentos e filtragem de spam. Eles requerem uma pequena quantidade de dados de treinamento para estimar os parâmetros necessários.

Classificadores Naive Bayes podem ser extremamente rápidos em comparação com métodos mais sofisticados. O desacoplamento das distribuições de características condicionais de classe significa que cada distribuição pode ser estimada independentemente como uma distribuição unidimensional. Isso, por sua vez, ajuda a aliviar os problemas decorrentes da "maldição da dimensionalidade".

No classificador Naive Bayes escolhido nesse trabalho, a verossimilhança das features é considerada gaussiana:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Os parâmetros σ_y e μ_y são estimados usando a máxima verossimilhança.

4.3.2.3 Decision Tree

Árvores de decisão (DTs) são um método de aprendizado supervisionado não paramétrico usado para classificação e regressão. O objetivo é criar um modelo que preveja o valor de uma variável de destino, aprendendo regras de decisão simples inferidas dos recursos de dados. Uma árvore pode ser vista como uma aproximação constante por partes.

As árvores de decisão aprendem com os dados a se aproximar de uma curva senoidal com um conjunto de regras de decisão se-então-senão. Quanto mais profunda a árvore, mais complexas são as regras de decisão e mais adequado é o modelo.

Algumas vantagens das árvores de decisão são:

- Simples de entender e interpretar, as árvores podem ser visualizadas graficamente;
- Requer pouca preparação de dados, em relação a outras técnicas;
- O custo computacional é logarítmico em relação ao número de pontos usados para treinar o modelo;
- Capaz de lidar com problemas de múltiplas saídas;
- Os resultados possuem interpretabilidade simples em relação a outras técnicas;
- É possível validar modelo por meio de testes estatísticos, tornado possível mensurar a confiabilidade do modelo;

- Apresenta um bom desempenho, mesmo que suas hipóteses são violadas.

As desvantagens das árvores de decisão incluem:

- Árvores muito complexas não generalizam bem os dados por conta do overfitting;
- Podem ser instáveis devido a variabiblidade dos dados;
- São representadas por funções não-contínuas (constantes por partes);
- As soluções são sub-ótimas e requerem abordagem heurística;
- Há problemas difíceis de modelar, como multiplexadores, paridade e XOR;
- É preciso balancear os dados para não gerar uma árvore com viés.

4.3.2.4 Random Forest

As florestas aleatórias são uma forma de calcular a média de várias árvores de decisão profundas, treinadas em diferentes partes do mesmo conjunto de treinamento, com o objetivo de reduzir a variância. Isso ocorre às custas de um pequeno aumento no viés e alguma perda de interpretabilidade, mas geralmente aumenta muito o desempenho no modelo final.

As florestas são como reunir esforços de algoritmos de árvore de decisão. Fazer o trabalho em equipe de muitas árvores, melhorando assim o desempenho de uma única árvore aleatória. Embora não sejam exatamente a mesma coisa, as florestas fornecem os efeitos parecidos com o de uma validação cruzada K-fold.

No geral, cada árvore no conjunto é construída a partir de uma amostra retirada com substituição (ou seja, uma amostra bootstrap) do conjunto de treinamento. Além disso, ao dividir cada nó durante a construção de uma árvore, a melhor divisão é encontrada em todos os recursos de entrada ou em um subconjunto aleatório de tamanho pré-definido.

O objetivo dessas duas fontes de aleatoriedade é diminuir a variância do estimador florestal. Na verdade, as árvores de decisão individuais geralmente exibem alta variação e tendem a se sobre-ajustar. A aleatoriedade injetada nas florestas produz árvores de decisão com erros de previsão um tanto dissociados. Tirando uma média dessas previsões, alguns erros podem ser cancelados. As florestas aleatórias alcançam uma variação reduzida combinando diversas árvores, às vezes ao custo de um ligeiro aumento no viés. Na prática, a redução da variância é frequentemente significativa, resultando em um modelo geral melhor.

4.3.2.5 Logistic Regression

Em estatística , o modelo logístico (ou modelo logit) é usado para modelar a probabilidade de uma determinada classe ou evento existir. Isso pode ser estendido para modelar várias classes de eventos, onde cada classe detecdata está atribuída a uma probabilidade dentro de um intervalo binário.

Assim como em um problema de otimização, a classe binária l_2 ponderada pela regressão logística minimiza a seguinte função custo:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$$

Em um modelo de regressão logística binária, a variável dependente possui dois níveis (categóricos), e para o caso de saídas com mais de dois valores são modeladas por regressão logística multinomial. O próprio modelo de regressão logística simplesmente modela a probabilidade de saída em termos de entrada e não realiza classificação estatística. No problema proposto, foram escolhidos valores de corte para que as entradas sejam classificadas dentro dessas faixas de corte de probabilidades, transpondo a abordagem para um problema de classificação.

4.3.2.6 Support Vector Classification

As máquinas de vetores de suporte (SVMs) são um conjunto de métodos de aprendizado supervisionado usados para classificação, regressão e detecção de outliers.

Dados vetores de treinamento $x_i \in \mathbb{R}^p, i = 1, \ldots, n$, em duas classes, e um vetor $y \in \{1, -1\}^n$, nosso objetivo é encontrar $\omega \in \mathbb{R}^p$ e $b \in \mathbb{R}$ de modo que a previsão dada por $sign\left(\omega^T \phi(x) + b\right)$ está correto para a maioria das amostras.

O SVC resolve o seguinte problema primário:

$$\begin{aligned} \min_{\omega,b,\zeta} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n}\zeta_i \\ \text{sujeito a } y_i\left(\omega^T\phi\left(x_i\right) + b\right) \geq 1 - \zeta_i \\ \zeta_i \geq 0, i = 1, \ldots, n \end{aligned}$$

Intuitivamente, estamos tentando maximizar a margem (minimizando $\|\omega\|^2 = \omega^T\omega$), enquanto causando uma penalidade quando uma amostra é classificada incorretamente

ou dentro do limite da margem. Idealmente, o valor seria para todas as amostras, o que indica uma previsão perfeita. Mas geralmente os problemas nem sempre são perfeitamente separáveis com um hiperplano, então permitimos que algumas amostras fiquem à distância de seu limite de margem correto. O termo de penalidade C controla a intensidade dessa penalidade e, como resultado, atua como um parâmetro de regularização.

Depois que o problema de otimização for resolvido, a saída de "função de decisão"para uma determinada amostra x se torna:

$$\sum_{i \in SV} y_i \alpha_i K\left(x_i, x\right) + b$$

e a classe prevista corresponde ao seu sinal. Precisamos apenas somar sobre os vetores de suporte (isto é, as amostras que estão dentro da margem) porque os coeficientes duplos α_i são zero para as outras amostras.

Podemos destacar as principais vantagens dessa metodologia:

- Eficaz em problemas que possuem muitas dimensões de features;
- Ainda eficaz nos casos em que o número de dimensões é maior do que o número de amostras;
- Usa um subconjunto de pontos de treinamento na função de decisão (chamados vetores de suporte), portanto, também é eficiente em termos de memória;
- É bastante versátil, pois a função de decisão pode sem implementada utilizando diferentes Kernels personalizados.

Dentre as princiapis desvantagens das máquinas de vetores de suporte, podemos destacar:

- Nos casos em que o número de features for muito maior do que o número de amostras, é preciso evitar o ajuste excessivo na escolha das funções do kernel, além de que o termo de regularização é crucial.
- Os SVMs não fornecem estimativas de probabilidade diretamente, elas são calculadas usando uma validação cruzada.

4.3.2.7 Neural Network

A rede neural utilizada nesse trabalho é baseada na arquitetura Multi-layer Perceptron (MLP), um algoritmo de aprendizado supervisionado que aprende uma função

$f(\cdot) : R^m \rightarrow R^o$ treinando em um conjunto de dados, onde m é o número de dimensões para entrada e o é o número de dimensões para saída. Dado um conjunto de recursos $X = x_1, x_2, ..., x_m$ e um alvo , ele pode aprender um aproximador de função não linear para classificação ou regressão. É diferente da regressão logística, pois entre a camada de entrada e a de saída pode haver uma ou mais camadas não lineares, chamadas camadas ocultas. A Figura 5 mostra um MLP de uma camada oculta com saída escalar

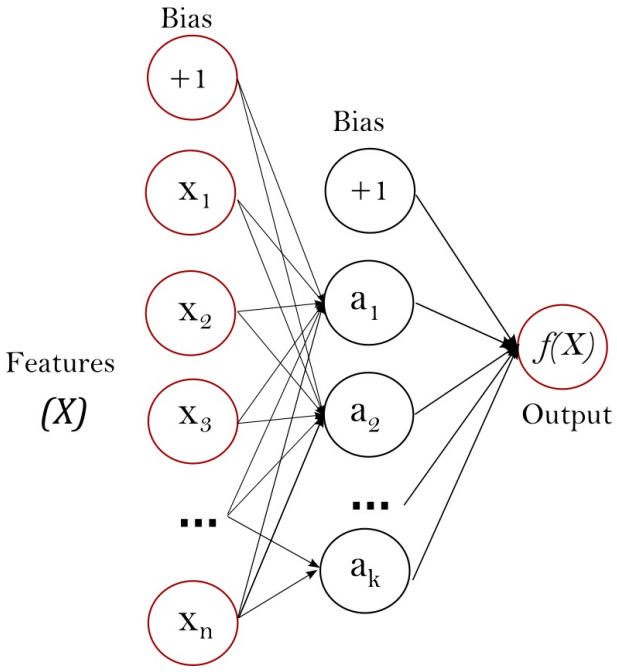

Figura 5 – MLP de uma camada escondida

A camada mais à esquerda, conhecida como camada de entrada, consiste em um conjunto de neurônios $\{x_i|x_1, x_2, ..., x_m\}$ representando os recursos de entrada. Cada neurônio na camada oculta transforma os valores da camada anterior com uma soma linear ponderada $w_1x_1 + w_2x_2 + ... + w_mx_m$, seguido por uma função de ativação não linear $g(\cdot) : R \rightarrow R$ como a função tangente hiperbólica. A camada de saída recebe os valores da última camada oculta e os transforma em valores de saída.

As principais vantagens do Perceptron multicamadas são:

- Capacidade de aprender modelos não lineares;
- Capacidade de aprender modelos em tempo real (aprendizado on-line).

Dentre as principais desvantagens, podemos destacar:

- MLP com camadas ocultas tem uma função de perda não convexa onde existe mais de um mínimo local. Portanto, inicializações de peso aleatório diferentes podem levar a uma precisão de validação diferente;
- O MLP requer o ajuste de vários hiperparâmetros, como o número de neurônios ocultos, camadas e iterações;
- O MLP é sensível ao dimensionamento de recursos.

4.3.3 Métricas de avaliação

Para analisar e conseguir comparar os resultados das várias técnicas, é preciso diferenciar o efeito qualitativo que cada métrica nos informa. Para tanto, é fundamental compreender a diferença entre precisão e acurácia. A figura 6 nos fornece idéia intuitiva de como esses 2 conceitos se relacionam.

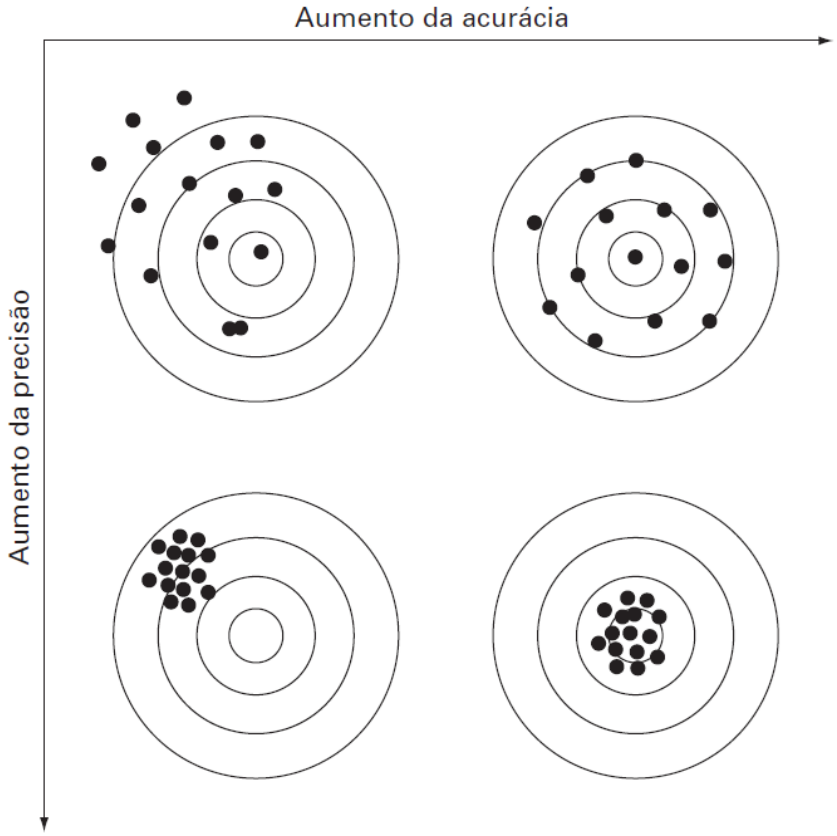

Figura 6 – Diferença entre acurácia e precisão

Para compreender a relação matemática entre esses dois conceitos, será utilizado como base a matriz de confusão.

4.3.3.1 Matriz de confusão

Em análise preditiva, a matriz de confusão é uma tabela com duas linhas e duas colunas que relata o número de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos. Isso permite uma análise mais detalhada do que a mera proporção de classificações corretas (precisão). A precisão produzirá resultados enganosos se o conjunto de dados estiver desequilibrado; isto é, quando o número de observações em diferentes classes variam muito.

Tabela 2 – Exemplo de tabela de confusão com 2 classes.

		Reality	
	Total	positive	negative
Prediction	positive	TP	FP
	negative	FN	TN

TP - verdadeiro positivo (true positive); **FN** - falso negativo (false negative);

FP - falso positivo (false possitive); **TN** - verdadeiro negativo (true negative).

Podemos extender o mesmo raciocínio quando consideramos mais classes, lembrando que o objetivo é sempre maximizar os elementos da diagonal principal, que caracterizam a quantidade de acertos na predição de uma determinada classe.

Tabela 3 – Exemplo de tabela de confusão com 3 classes.

		Reality		
	Total	Class 1	Class 2	Class 3
Prediction	Class 1	T1	$F_2 1$	$F_3 1$
	Class 2	$F_1 2$	T2	$F_3 2$
	Class 3	$F_1 3$	$F_2 3$	T3

$T1$ - acertos em relação à classe 1; $F_3 2$ - classe 3 classificada como classe 2;

$F_2 1$ - classe 2 classificada como classe 1; $T3$ - acertos em relação à classe 3;

$F_3 1$ - classe 3 classificada como classe 1; $F_1 3$ - classe 1 classificada como classe 3;

$T2$ - acertos em relação à classe 2; $F_2 3$ - classe 2 classificada como classe 3;

$F_1 2$ - classe 1 classificada como classe 2;

4.3.3.2 Acurácia

Acurácia consiste no grau de conformidade de um valor medido ou calculado em relação à sua definição ou com respeito a uma referência padrão. Com relação ao problema proposto, podemos definir a acurácia como a proporção de acerto no processo de classificação.

Considerando-se a definição com 2 classes, exemplificado na tabela 2, temos:

$$acc_2 = \frac{TP + TN}{TP + FP + FN + TN} \tag{13}$$

Extendendo o conceito para o problema proposto na tabela 3, é possível calcular a acurácia para 3 classes:

$$acc_3 = \frac{T1 + T2 + T3}{T1 + F_2 1 + F_3 1 + T2 + F_1 2 + F_3 2 + T3 + F_1 3 + F_2 3} \tag{14}$$

Essa generalização pode ser feita, independente da quantidade de classes.

4.3.3.3 Precisão e revocação

Precisão é o grau de variação de resultados de uma medição, tendo como base o desvio-padrão de uma série de repetições da mesma análise. Diferente da acurácia, a precisão será avaliada em relação a cada classe predita (em relação a cada linha da tabela). Analogamente, a revocação é avaliada em relação a cada classe verificada (em relação a cada coluna da tabela).

Para compreender melhor a diferença conceitual entre previsão e revocação, segue-se um exemplo:

Precisão - "Quantos elementos selecionados são relevantes".

Revocação - "Quantos elementos relevantes foram selecionados".

onde é possível perceber o caráter complementar de ambas as definições.

Do ponto de vista matemático, ao tomar como base a definição com 2 classes, as expressões 15 e 16 referem-se (respectivamente) a precisão e revocação em relação à classe positiva:

$$pre_2(P) = \frac{TP}{TP + FP} \tag{15}$$

$$rec_2(P) = \frac{TP}{TP + FN} \tag{16}$$

Extendendo o conceito para o problema proposto na tabela 3, na formulação com 3 classes e, tomando como referencia a classe 1, as expressões 17 e 18 precisão e revocação são dadas (respectivamente) por:

$$pre_3(1) = \frac{T1}{T1 + F_2 1 + F_3 1} \tag{17}$$

$$rec_3(1) = \frac{T1}{T1 + F_1 2 + F_1 3} \tag{18}$$

4.3.3.4 Score-F1

Existem várias métricas que podem ser usadas para avaliar um modelo de classificação binária, e a precisão é uma das mais simples de entender. A precisão pode ser útil, mas não leva em consideração as sutilezas dos desequilíbrios de classe ou custos diferentes de falsos negativos e falsos positivos.

Nesse sentido, algumas das vantagens de se utiliza a métrica Score-F1 são:

- quando existem cutos diferentes de falsos positivos ou falsos negativos;
- quando há grande desequilíbrio entre as classes.

A precisão tem a vantagem de ser muito facilmente interpretável, mas a desvantagem de não ser robusta quando os dados estão distribuídos de maneira desigual ou quando há um custo mais alto associado a um tipo específico de erro.

O Score-F1 também é avaliado separadamente em relação a cada classe, à partir da média harmônica entre precisão e revocação.

$$F_1 = \frac{2}{\frac{1}{pre} + \frac{1}{rec}} = 2 \times \frac{pre \times rec}{pre + rec} \tag{19}$$

5 Resultados

O objetivo desse capítulo é apresentar os resultados obtidos pelos algoritmos de classificação em diferentes cenários e configurações de discretização e conjunto de variáveis, buscando compreender a correlação de algumas variáveis na identificação de ciclos econômicos.

Um ponto de interesse a se observar é que após as discretizações, algumas variáveis aprensetaram baixa correlação. Portanto, foram realizados testes com um conjunto de variáveis completo e outros testes somente com variáveis de alta correlação, para avaliar se essas variáveis menos significantes eram irrelevantes ou poderiam atrapalhar o modelo de alguma forma.

5.1 Análise das variáveis

O primeiro passo natural é analisar a correlação entre as variáveis propostas, através da matriz de coeficientes de Pearson.

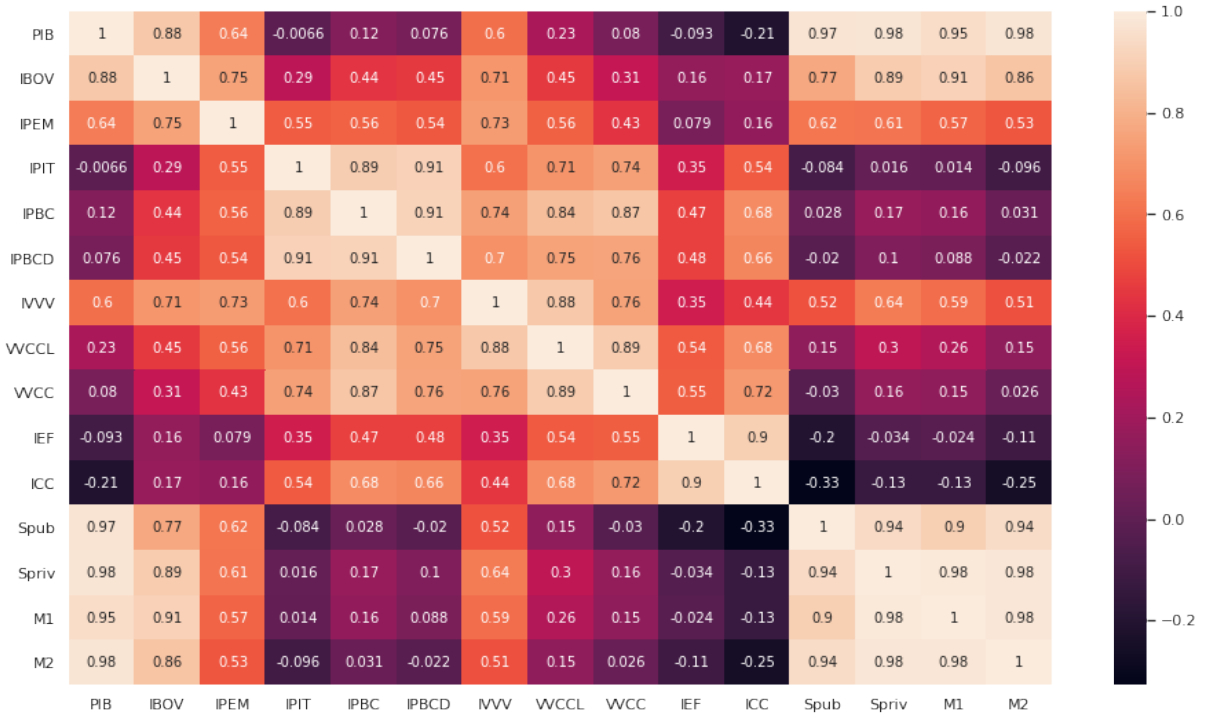

Figura 7 – Matriz de correlação das variáveis brutas

A abordagem utilizando as variáveis brutas é pouco significativa, pois como podemos observar na tabela 1, as variáveis possuem unidades de medida distintas. A maneira sugerida

para torna-las comparáveis é realizar uma transformação que retorna a variação percentual entre 2 períodos consecutivos para cada uma das variáveis.

$$f(x_i) = \Delta x_i = x_i - x_{i-1} \tag{20}$$

Outro fato relevante é que a variável denotada por "IBOV"(que representa a soma do valor das empresas listadas na Bovespa), apesar de ser considerada uma variável relevante, foi retirada da análise uma vez que a série foi descontinuada. Uma proposta para trabalhos futuros é tentar replicar a metodologia para calcular os dados faltantes e poder incluir essa séria como variável explicativa para o modelo.

Aplicando a transformação 20 nas séries de dados propostas pela tabela 1, é possível recalcular a matriz de coeficientes de Pearson:

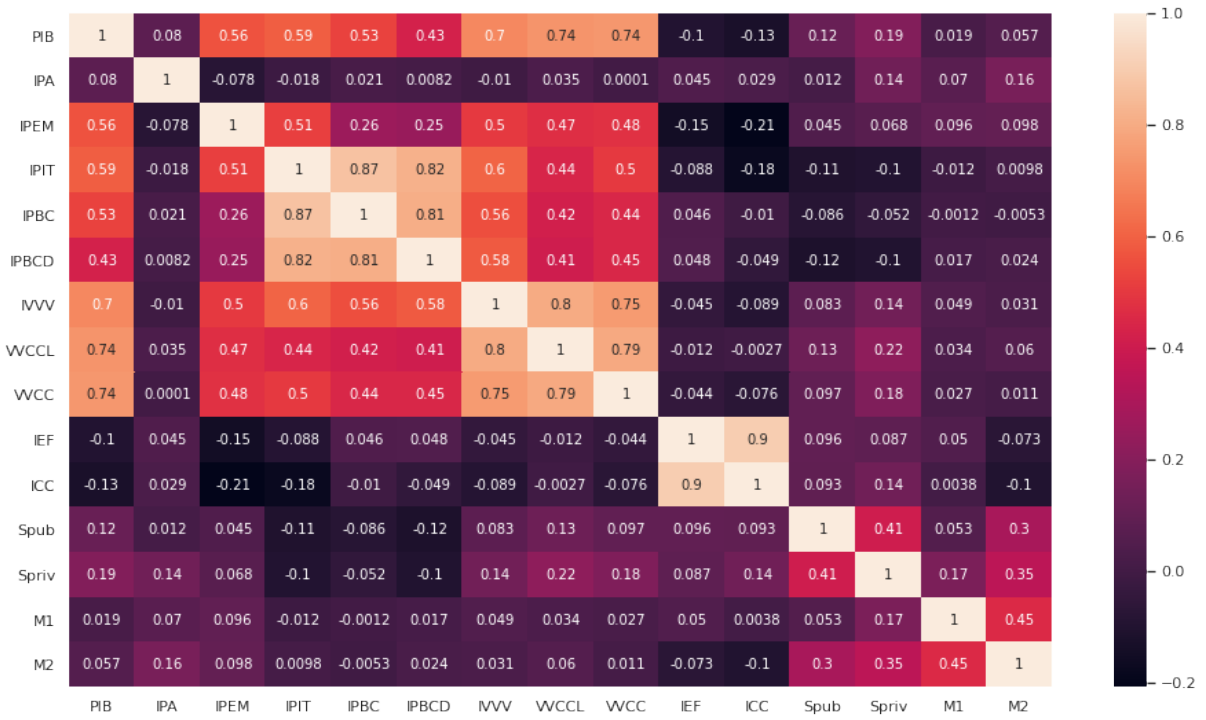

Figura 8 – Matriz de correlação das variáveis em variação percentual

É possível notar que foi incluída uma nova variável denotada por "IPA"(índice de preços por atacado-mercado), uma vez que ela já é fornecida diretamente na unidade de medida de variação percentual, ou seja, não necessita a aplicação da transformação 20.

5.2 Cenários de classificação

Antes de apresentar os resultados é importante ressaltar que algumas técnicas propostas são bastante sensíveis à discretização, de modo que a classificação não foi possível. Apesar de aparentemente negativo, esse comportamento ajudou na identificação de melhorias que tornasse a modelagem mais precisa e coerente com a realidade.

Outras técnicas são bastante sensíveis aos parâmetros intrínsecos, especialmente no que se refere ao overfitting. Portanto, serão apresentadas os setups utilizados em cada cenário, pois pode ser alvo de um estudo posterior.

Para o processo de validação cruzada, foi utilizado um split no intervalo de 30%, sendo que o split dos folds dentro do conjunto de treinamento foi de 20%. Como o conjunto de dados possuia inicialmente 231 recorrências, o dataset foi dividido da seguinte forma:

Conjunto de treinamento: 128 medidas

Conjunto de validação: 33 medidas

Conjunto de teste: 70 medidas

O método de classificação baseado em redes neurais MLP possui grande variabilidade de eficacia em relação a sua configuração. Dessa forma, foi proposto um algoritmo que calcula os parâmetros ótimos dentro de um cenário de discretização. Cada setup será detalhado da subseção correspondente.

A estrutura de dados utilizada para imlpementar os métodos de seleção, treinamento, classificação e avaliação foram as listas ligadas.

5.2.1 Classificação binária

Ao aplicar a discretização binária, descrita pela equação 9, é importante recalcular a matriz de coeficientes de Pearson para as novas variáveis propostas.

$$\begin{cases} x \mapsto 0, & \text{se } \Delta x < 0; \\ x \mapsto 1, & \text{se } \Delta x \geq 0. \end{cases}$$

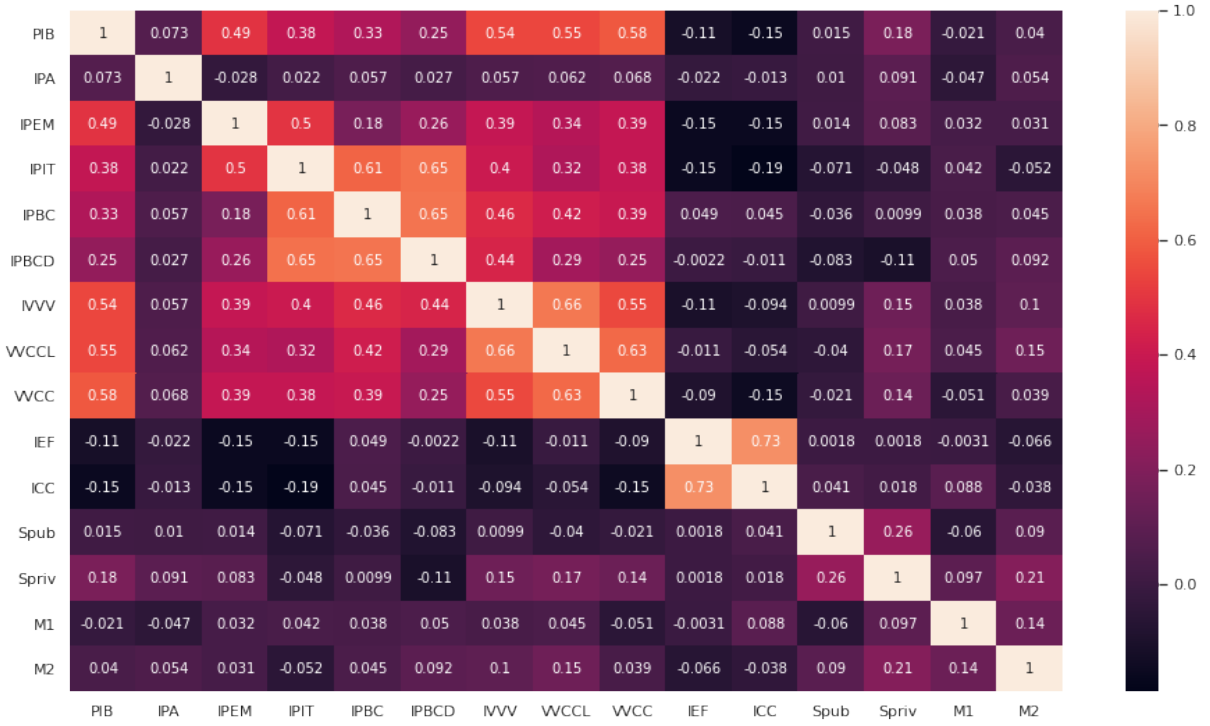

Figura 9 – Matriz de correlação das variáveis discretizadas utilizando 2 classes.

Comparando as figuras 8 e 9, é possível notar que o processo de discretização causou uma redução na correlação geral entre as variáveis, o que reforça a importância da metodologia respeitar a sensibilidade intrinseca do problema proposto.

Nessa configuração de discretização e utilizando todas as variáveis propostas, a rede neural foi configurada utilizando os seguintes parâmetros:

Função de ativação Relu

Solver SGD

Número de camadas ocultas 21

Tabela 4 – Acurácia da classificação binária na base completa

Métodos	Treino	Teste	ε
Nearest Neighbors	78,79%	90,00%	11,21%
Naive Bayes	75,76%	87,14%	11,39%
Decision Tree	72,73%	80,00%	7,27%
Random Forest	72,73%	88,57%	15,84%
Logistic Regression	75,76%	84,29%	8,53%
Support Vector Classification	72,73%	88,57%	15,84%
Neural Network	72,73%	85,71%	12,99%

Figura 10 – Acurácia da classificação binária na base completa

Analisando a tabela 4, é possível notar que ambas as técnicas propostas atingiram um nível parecido de acurácia tanto na etapa de treino quanto na etapa de testes. A coluna erro é calculada à partir da diferença entre os resultados entre as 2 etapas sugeridas, ou seja.

$$\varepsilon = Train - Test \tag{21}$$

Apesar da diferença relativamente significativa entre treino e teste, não é muito coerente atribuir esse erro ao overfitting, uma vez que a acurácia foi mais alta no conjunto de teste. Como ambos os métodos tiveram um aumento similar (tanto proporcional quanto absoluto), e considerando que a quantidade de dados disponíveis é relativamente pequena, provavelmente isso decorre de um erro de viés.

Além da análise de acurácia relativa a cada método, é fundamental avaliar o modelo em relação as sutilizas de cada classe escolhida. A tabela 5 apresenta os valores de precisão e revocação para cada uma das classes, mas o parâmetro considerado para fins de discussão será somente o score-F1, uma vez que ele é robusto em relação a desequilibrios entre classes ou ainda os custos diferentes entre falsos positivos e falsos negativos.

Tabela 5 – Avaliação da classificação binária na base completa

Métodos	Classe	Precisão	Revocação	F_1-Score
Nearest Neighbors	0	82,14%	92,00%	86,79%
	1	95,24%	88,89%	91,95%
Naive Bayes	0	78,57%	88,00%	83,02%
	1	92,86%	86,67%	89,66%
Decision Tree	0	70,37%	76,00%	73,08%
	1	86,05%	82,22%	84,09%
Random Forest	0	84,00%	84,00%	84,00%
	1	91,11%	91,11%	91,11%
Logistic Regression	0	71,88%	92,00%	80,70%
	1	94,74%	80,00%	86,75%
Support Vector Classification	0	79,31%	92,00%	85,19%
	1	95,12%	86,67%	90,70%
Neural Network	0	82,61%	76,00%	79,17%
	1	87,23%	91,11%	89,13%

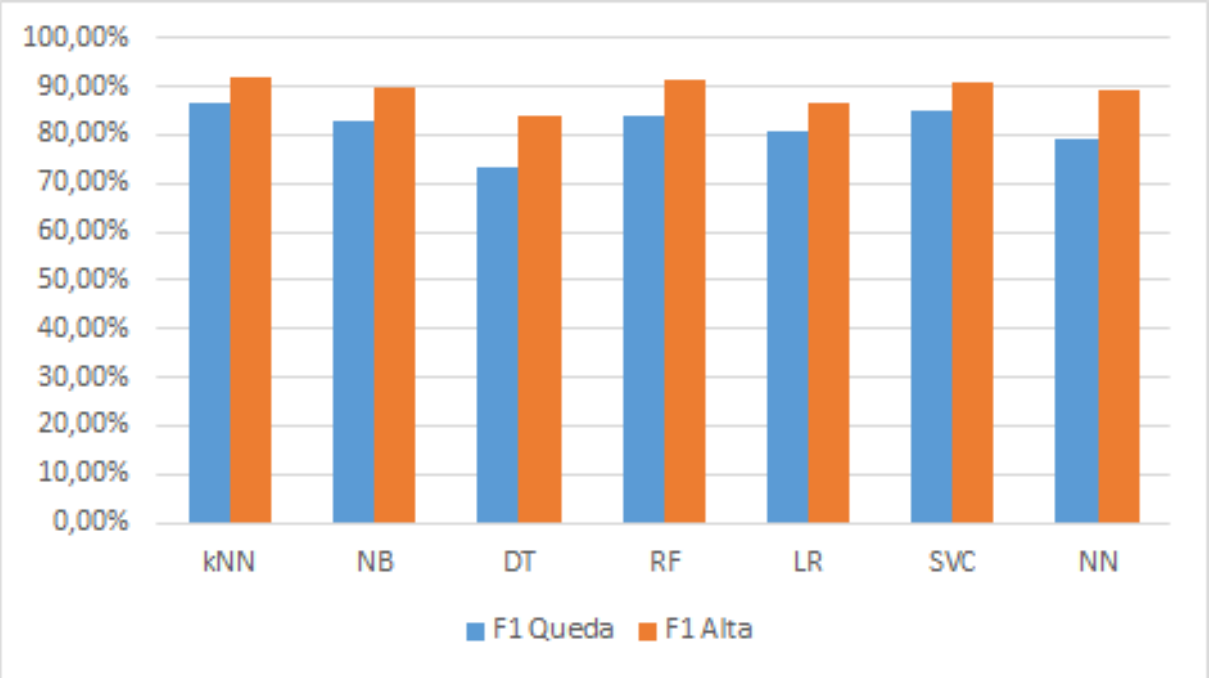

Figura 11 – Score-F1 da classificação binária na base completa

Assim como foi apresentado em relação a acurácia, é possível notar que em relação ao score-F1, ambos os métodos tiveram resultado semelhante. Outro fato relevante que fica nítido é que é mais fácil prever movimentos de alta do que movimentos de queda.

Buscando eliminar a influência de variáveis de baixa correlação, foi proposta uma segunda base de dados considerando somente os indicadores com mais alta correlação, cuja matriz de coeficientes de Pearson é dada por:

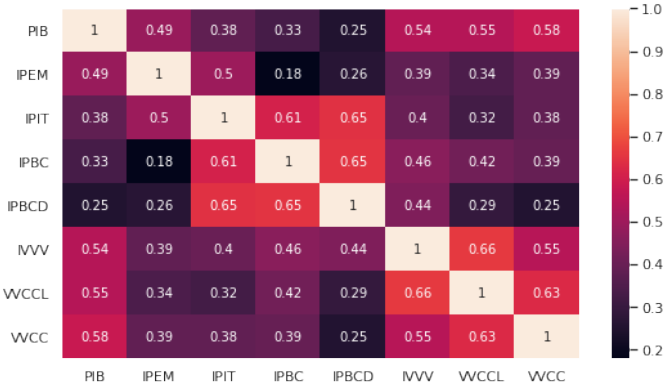

Figura 12 – Matriz de correlação das variáveis de maior correlação, discretizadas utilizando 2 classes.

Nessa configuração de discretização e utilizando somente as variáveis com maior correlação, a rede neural foi configurada utilizando os seguintes parâmetros:

Função de ativação Tanh

Solver SGD

Número de camadas ocultas 18

Tabela 6 – Acurácia da classificação binária na base restrita

Métodos	Treino	Teste	ε
Nearest Neighbors	69,70%	90,00%	20,30%
Naive Bayes	72,73%	87,14%	14,42%
Decision Tree	75,76%	88,57%	12,81%
Random Forest	72,73%	90,00%	17,27%
Logistic Regression	72,73%	87,14%	14,42%
Support Vector Classification	75,76%	90,00%	14,24%
Neural Network	75,76%	85,71%	9,96%

Comparando a tabela 4 com 6, vemos que apesar de uma pequena melhora em relação ao dataset mais restrito, ela pode ser desconsiderada uma vez que fica dentro da mergem de erro da medida. Inclusive, no conjunto de teste a variabilidade foi praticamente inexistente, o que reforça a idéia de que apesar de discretização grosseira, o nível de

aprendizagem foi satisfatório. Nesse cenário, utilizar a base restrita se justifica, uma vez que alcança o mesmo resultado, com um custo computacional menor.

Figura 13 – Acurácia da classificação binária na base restrita

Tabela 7 – Avaliação da classificação binária na base restrita

Métodos	Classe	Precisão	Revocação	F_1-Score
Nearest Neighbors	0	82,14%	92,00%	86,79%
	1	95,24%	88,89%	91,95%
Naive Bayes	0	76,67%	92,00%	83,64%
	1	95,00%	84,44%	89,41%
Decision Tree	0	84,00%	84,00%	84,00%
	1	91,11%	91,11%	91,11%
Random Forest	0	84,62%	88,00%	86,27%
	1	93,18%	91,11%	92,13%
Logistic Regression	0	75,00%	96,00%	84,21%
	1	97,37%	82,22%	89,16%
Support Vector Classification	0	87,50%	84,00%	85,71%
	1	91,30%	93,33%	92,31%
Neural Network	0	86,21%	76,00%	79,17%
	1	87,23%	91,11%	89,13%

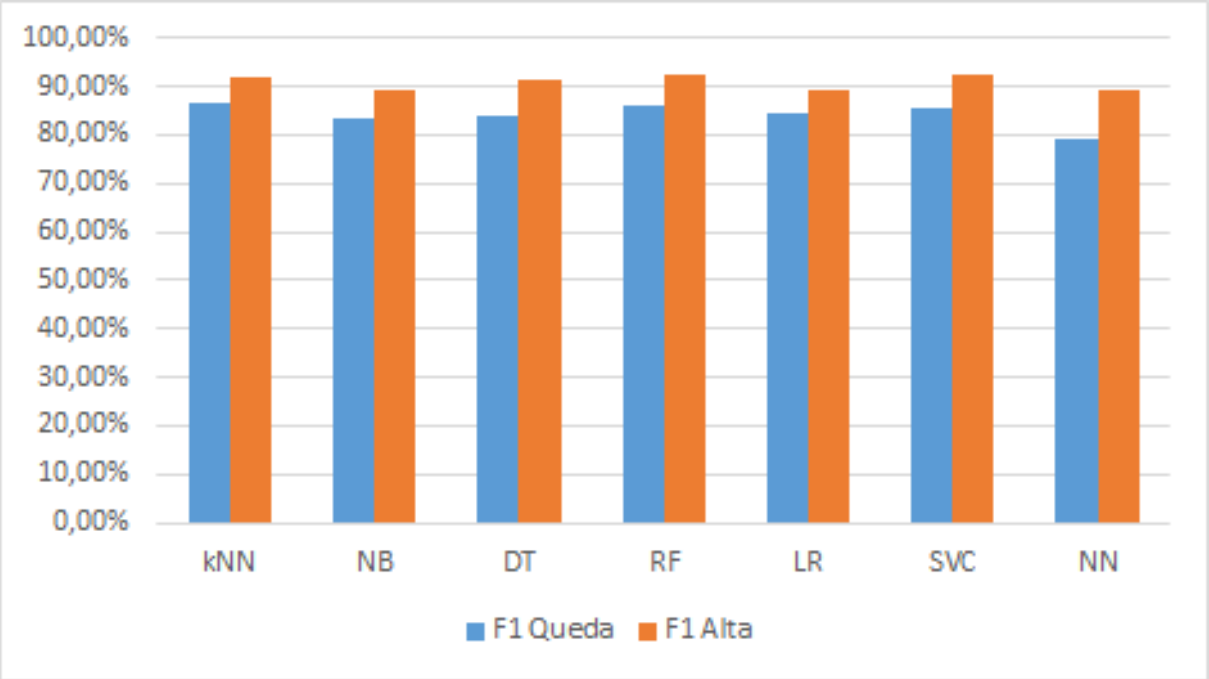

Figura 14 – Score-F1 da classificação binária na classe restrita

Em relação ao Score-F1, há uma pequena melhora quando usada a base restrita, que em termos relativos é pouco significativa, mas já justifica a redução das variáveis, visto que temos uma excelente resposta com um modelo mais simplificado.

5.2.2 Classificação multiclasse - 3 classes

A discretização proposta a seguir possuí uma maior complexidade intrinseca, tendo em vista que o espaço de saída possuí mais possibilidades. No entanto, essa abordagem faz muito mais sentido prático, já que é fundamental que os períodos de lateralização sejam identificados.

Assim como no caso binário, ao aplicar a discretização utilizando 3 classes, é importante recalcular a matriz de coeficientes de Pearson para as novas variáveis propostas. A discretização proposta pode ser descrita pela equação 10, onde μ_x é a média e σ_x é o desvio padrão da variável x em questão.

$$\begin{cases} x \mapsto -1, & \text{se } \Delta x \leq \mu_x - \sigma_x; \\ x \mapsto 0, & \text{se } \mu_x - \sigma_x < \Delta x < \mu_x + \sigma_x; \\ x \mapsto 1, & \text{se } \Delta x \geq \mu_x + \sigma_x. \end{cases}$$

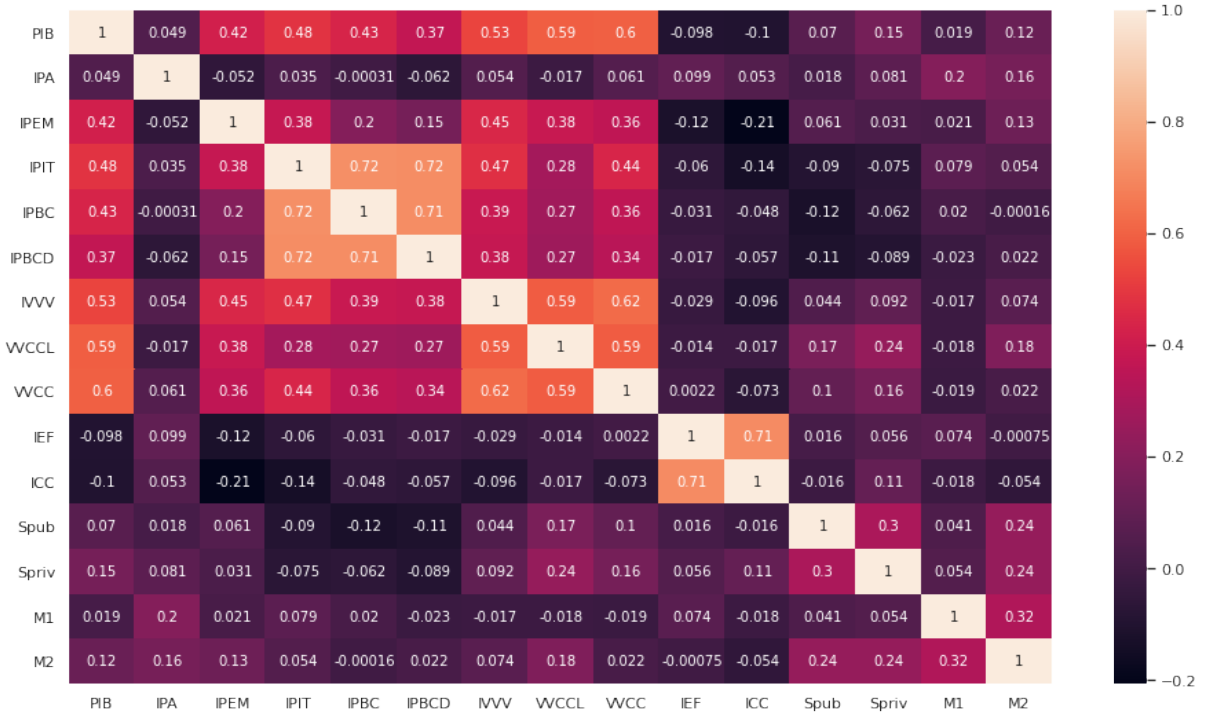

Figura 15 – Matriz de correlação das variáveis discretizadas utilizando 3 classes.

Nessa configuração de discretização e utilizando todas as variáveis propostas, a rede neural foi configurada utilizando os seguintes parâmetros:

Função de ativação Identity

Solver SGD

Número de camadas ocultas 13

Tabela 8 – Acurácia da classificação com 3 classes na base completa

Métodos	Treino	Teste	ε
Nearest Neighbors	78,79%	81,43%	2,64%
Naive Bayes	84,85%	78,57%	-6,28%
Decision Tree	60,61%	72,86%	12,25%
Random Forest	78,79%	82,86%	4,07%
Logistic Regression	81,82%	68,57%	-12,82%
Support Vector Classification	81,82%	81,00%	-0,82%
Neural Network	81,82%	84,00%	2,18%

Comparando a tabela 4 com 8, podemos perceber que a variabilidade entre os métodos aumentou ligeiramente, mas ainda dentro de um range bastante limitado.

Em relação a etapa de treinamento é possível notar um aumento médio no nível de acurácia. Já em relação a etapa de testes houve uma queda média, tornando o resultado mais coerente com o resultado que era esperado.

Figura 16 – Acurácia da classificação com 3 classes na base completa

Em média o valor absoluto de ε diminuiu, o que significa que o modelo se ajustou melhor ao fenomeno observado. Além disso a maior variabilidade em relação à classificação binária ocorreu com os métodos que possuem uma abordagem mais simples, como a árvore de decisão e a regressão logistica. Como essa discretização multiclasse possui maior complexidade, é esperado que os métodos mais sofisticados sejam capazes de lidar com as não linearidades de forma mais eficiente.

Comparando a tabela 4 com 9, percebemos que a inclusão de mais classes, trouxe um aumento significativo na variabilidade do Score-F1. Em grande parte isso pode ser explicado pela baixa quantidade de dados disponíveis, já que em classes com menor frequência, cada classificação errada tem peso relativo mais significativo.

O gráfico 17 deixa claro que além da maior dificuldade em se prever os eventos quie se distancias da medida central (alta e baixa), há uma dificuldade maior em se prever as quedas. Isso corrobora o que se observa empiricamente, uma vez que os movimentos de correção e queda costumam ser mais repentinos e abruptos.

Tabela 9 – Avaliação da classificação com 3 classes na base completa

Métodos	Classe	Pre	Rec	Score-F1
Nearest Neighbors	-1	80,00%	36,36%	50,00%
	0	79,66%	97,92%	87,85%
	1	100,00%	54,55%	70,59%
Naive Bayes	-1	62,50%	45,45%	52,63%
	0	82,35%	87,50%	84,85%
	1	72,73%	72,73%	72,73%
Decision Tree	-1	60,00%	54,55%	57,14%
	0	82,22%	77,08%	79,57%
	1	53,33%	72,73%	61,54%
Random Forest	-1	71,43%	45,45%	55,56%
	0	84,62%	91,67%	88,00%
	1	81,82%	81,82%	81,82%
Logistic Regression	-1	57,14%	72,73%	64,00%
	0	88,24%	62,50%	73,17%
	1	45,45%	90,91%	60,61%
Support Vector Classification	-1	80,00%	36,36%	50,00%
	0	79,66%	97,92%	87,85%
	1	100,00%	54,55%	70,59%
Neural Network	-1	83,33%	45,45%	58,82%
	0	83,64%	95,83%	89,32%
	1	88,89%	72,73%	80,00%

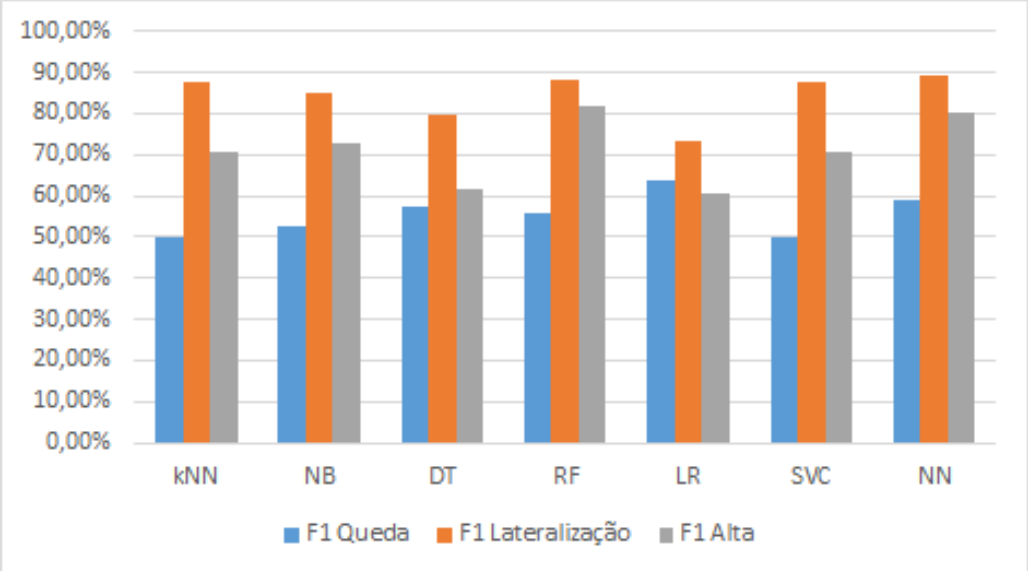

Figura 17 – Score-F1 da classificação com 3 classes na base completa

Assim como na discretização binária, foi proposta outra base de dados considerando somente os indicadores com mais alta correlação, cuja matriz de coeficientes de Pearson é dada por:

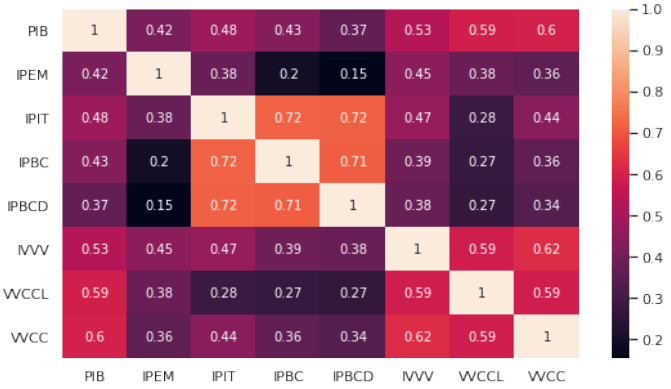

Figura 18 – Matriz de correlação das variáveis de maior correlação, discretizadas utilizando 3 classes.

Nessa configuração de discretização e utilizando somente as variáveis com maior correlação, a rede neural foi configurada utilizando os seguintes parâmetros:

Função de ativação Tanh

Solver Adam

Número de camadas ocultas 5

Tabela 10 – Acurácia da classificação binária na base restrita

Métodos	Train	Test	ε
Nearest Neighbors	78,79%	81,43%	2,64%
Naive Bayes	78,79%	80,00%	1,21%
Decision Tree	69,70%	75,71%	6,02%
Random Forest	69,70%	75,71%	6,02%
Logistic Regression	78,79%	75,71%	-3,07%
Support Vector Classification	75,76%	84,29%	8,53%
Neural Network	75,76%	82,86%	7,10%

Comparando as tabelas 8 e 10, percebemos uma ligeira melhor em algumas técnicas e ligeira piora para outras técnicas, dentro de um range que pode ser considerado margem de erro para a medida. Sob esse ponto de vista, faz sentido restringir a base uma vez que foi igualmente satisfatório, quando se utilizou um modelo mais simples.

Figura 19 – Acurácia da classificação com 3 classes na base restrita

Tabela 11 – Avaliação da classificação com 3 classes na base restrita

Métodos	Classe	Pre	Rec	Score-F1
Nearest Neighbors	-1	66,67%	36,36%	47,06%
	0	83,02%	91,67%	87,13%
	1	81,82%	81,82%	81,82%
Naive Bayes	-1	60,00%	54,55%	57,14%
	0	85,42%	85,42%	85,42%
	1	75,00%	81,82%	78,26%
Decision Tree	-1	66,67%	36,36%	47,06%
	0	79,25%	87,50%	83,17%
	1	63,64%	63,64%	63,64%
Random Forest	-1	66,67%	36,36%	47,06%
	0	79,25%	87,50%	83,17%
	1	63,64%	63,64%	63,64%
Logistic Regression	-1	58,33%	63,64%	60,87%
	0	87,80%	75,00%	80,90%
	1	58,82%	90,91%	71,43%
Support Vector Classification	-1	83,33%	45,45%	58,82%
	0	84,91%	93,75%	89,11%
	1	81,82%	81,82%	81,82%
Neural Network	-1	80,00%	36,36%	50,00%
	0	83,33%	93,75%	88,24%
	1	81,82%	81,82%	81,82%

Se comparamos o efeito da base restrita no Score-F1, essa melhora ainda existe mas é menos significativa do que no caso da discretização binária.

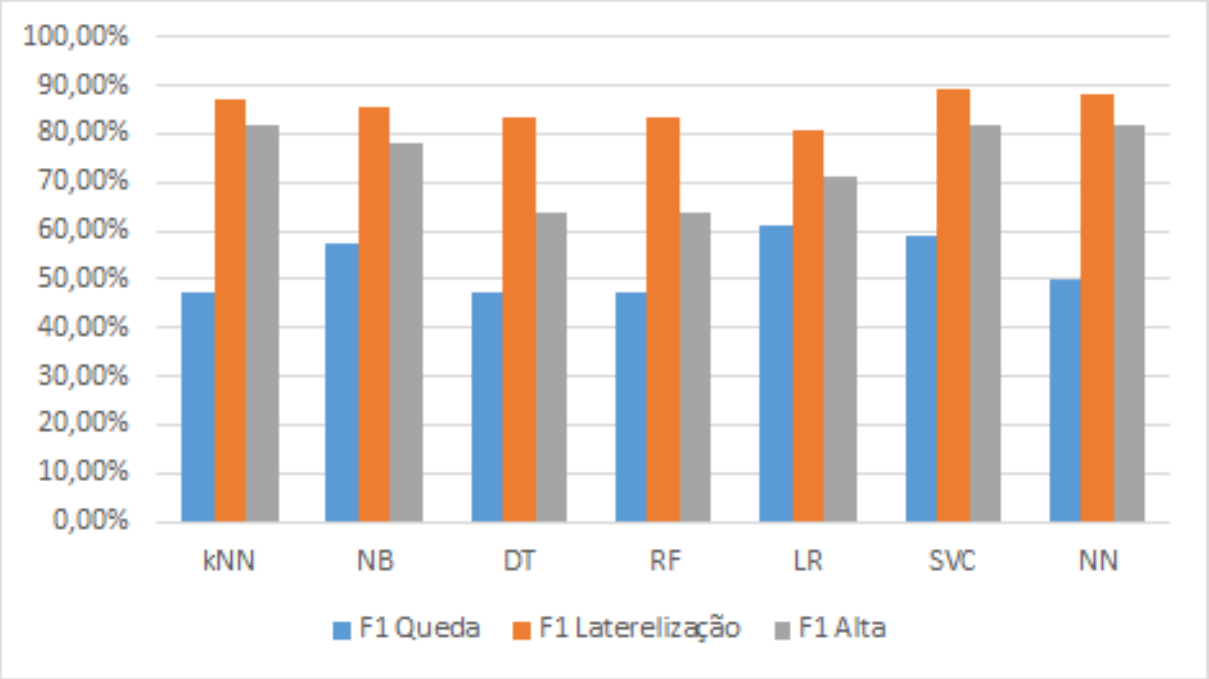

Figura 20 – Score-F1 da classificação com 3 classes na base completa

Ainda assim é possível notar que a redução de variáveis melhorou o desempenho, e se considerarmos que o modelo também se tornou mais simples, podemos dizer que o modelo é bem descrito pelo conjunto de variáveis da base restrita.

5.2.3 Classificação multiclasse - 5 classes

A abordagem de discretização utilizando 5 classes tem o objetivo principal de identificar outliers muito significativos em uma classe separada. Portanto, foi um incluído outro subintervalo nas extremidades, utilizando um incremento de 2 desvios-padrão, descrito pela equação 11, onde μ_x é a média e σ_x é o desvio padrão da variável x em questão.

$$\begin{cases} x \mapsto -2, & \text{se } \Delta x \leq \mu_x - 2 \cdot \sigma_x; \\ x \mapsto -1, & \text{se } \mu_x - 2 \cdot \sigma_x < \Delta x \leq \mu_x - \sigma_x; \\ x \mapsto 0, & \text{se } \mu_x - \sigma_x < \Delta x < \mu_x + \sigma_x; \\ x \mapsto 1, & \text{se } \mu_x + \sigma_x \leq \Delta x < \mu_x + 2 \cdot \sigma_x; \\ x \mapsto 2, & \text{se } \Delta x \geq \mu_x + 2 \cdot \sigma_x. \end{cases}$$

Como a abordagem para seleção do intervalo é empírica, foi proposto um intervalo que extendesse a idéia já bem sucedidade da divisão em 3 classes. No entanto, essa estratégia não se mostrou satisfatória, pois a restrição das classes limite foram tão severas, tornando a classe tão rara que o classificador nunca escolhia essa classe.

Tabela 12 – Acurácia da classificação com 5 classes standard na base restrita

Métodos	Train	Test	ε
Nearest Neighbors	69,70%	72,86%	3,16%
Naive Bayes	78,79%	67,14%	-11,65%
Decision Tree	60,61%	71,43%	10,82%
Random Forest	78,89%	74,29%	-4,50%
Logistic Regression	78,89%	58,57%	-20,22%
Support Vector Classification	75,76%	77,14%	1,39%
Neural Network	72,73%	75,71%	2,99%

Considerando somente a acurácia, podemos ter uma falsa impressão de que apesar da ligeira piora, a abordagem é valida. Isso por sí só já seria um argumento para desconsiderar a abordagem, uma vez que com 3 classes os resultados foram satisfatórios. No entanto, olhando para os valores de Score-F1 percebeu-se que o problema era mais grave, pois todos os métodos tiveram Precisão e Revocação nula para as classes extremais. Os métodos Support Vector Classification e Neural Network nem sequer conseguiram concluir os testes, tendo em vista o grau de erro induzido pela discretização.

Durante os testes, percebeu-se que a inclusão de um intervalo de tempo maior aumentou significativamente a qualidade dos métodos. No entanto, realmente o intervalo escolhido foi muito extremo, já que restringi-lo a $\pm 2\sigma$ garante que apenas 2,5% dos dados sejam considerados em cada ponta (95% dos dados pertencem às classes centrais). Portanto, decidiu-se uma nova abordagem intervalar com o objetivo de extender um pouco os limites de raridade para uma classe. Portanto foi proposta nova divisão intervalar, descrita pela equação 11, onde μ_x é a média e σ_x é o desvio padrão da variável x em questão.

$$\begin{cases} x \mapsto -2, & \text{se } \Delta x \leq \mu_x - 1,6745 \cdot \sigma_x; \\ x \mapsto -1, & \text{se } \mu_x - 1,6745 \cdot \sigma_x < \Delta x \leq \mu_x - 0,6745 \cdot \sigma_x; \\ x \mapsto 0, & \text{se } \mu_x - 0,6745 \cdot \sigma_x < \Delta x < \mu_x + 0,6745 \cdot \sigma_x; \\ x \mapsto 1, & \text{se } \mu_x + 0,6745 \cdot \sigma_x \leq \Delta x < \mu_x + 1,6745 \cdot \sigma_x; \\ x \mapsto 2, & \text{se } \Delta x \geq \mu_x + 1,6745 \cdot \sigma_x. \end{cases}$$

Essa configuração garante que a classe 0 (central) contemple 50% dos dados, enquanto que as classes -1 e 1 (classes intermediárias) contemplem 20% cada uma. Dessa forma, as 3 classes centrais correspondem a 90% dos dados e cada classe extrema (representadas por -2 e 2) correspondem a 5% cada uma, o que ainda torna a ocorrência de seus eventos relativamente raros.

Com essa nova discretização, é importante recalcular a matriz de coeficientes de Pearson para as novas variáveis propostas.

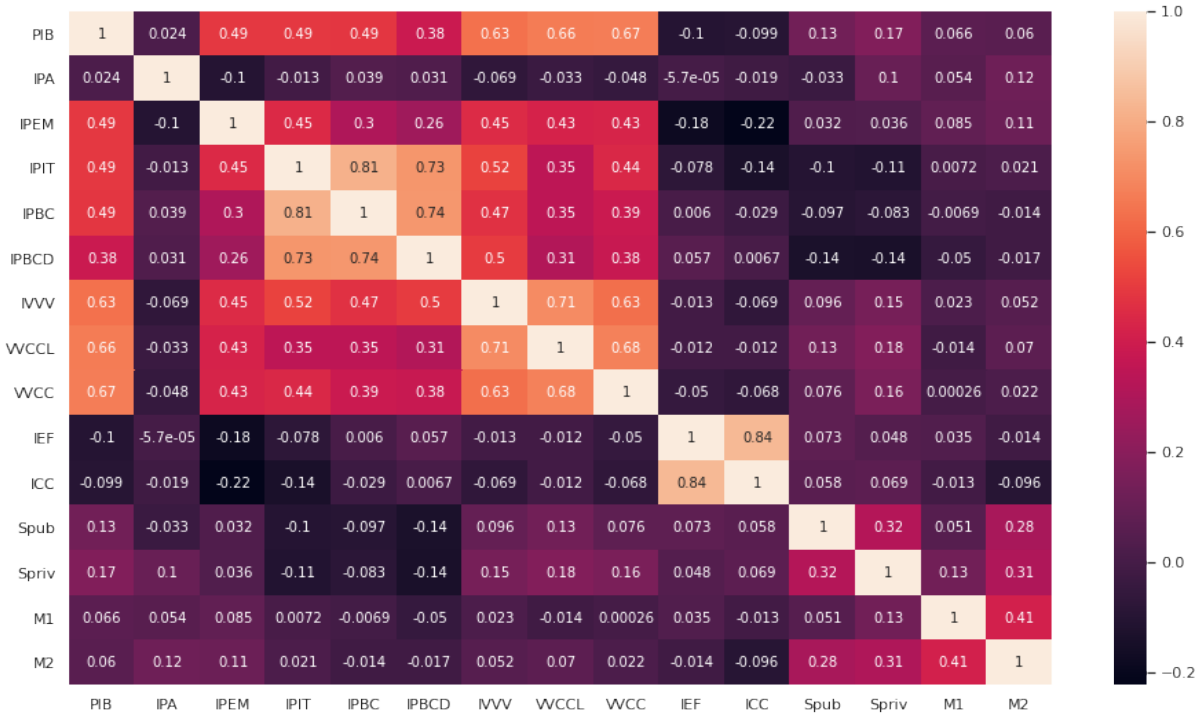

Figura 21 – Matriz de correlação das variáveis discretizadas utilizando 5 classes.

Nessa configuração de discretização e utilizando todas as variáveis propostas, a rede neural foi configurada utilizando os seguintes parâmetros:

Função de ativação Identity

Solver SGD

Número de camadas ocultas 16

Comparando a tabela 8 com 9 vemos uma nítida piora nos níveis de acurácia. Era esperado que uma modelagem ainda mais complexa implicasse em uma maior taxa de erro, mas os níveis alcançados se mostraram excessivamente insatisfatórios. Apesar da quantidade reduzida de dados influenciar para um resultado precário em um modelo mais

complexo, talvez esse fenômeno seja realmente melhor representado utilizando somente 3 classes.

Tabela 13 – Acurácia da classificação com 5 classes na base completa

Métodos	Train	Test	ε
Nearest Neighbors	51,52%	55,71%	4,20%
Naive Bayes	57,58%	50,00%	-7,58%
Decision Tree	51,52%	48,57%	-2,94%
Random Forest	54,55%	65,71%	11,17%
Logistic Regression	63,64%	57,14%	-6,49%
Support Vector Classification	69,70%	54,29%	-15,41%
Neural Network	63,64%	58,57%	-5,06%

Figura 22 – Acurácia da classificação com 5 classes na base completa

Através da tabela 14, podemos analisar o comportamento do Score-F1 para diferentes métodos. Apesar de todos terem convergido, algumas classes foram classificadas erradas em 100% das vezes, em métodos mais sofisticados como o Support Vector Classification e as redes neurais.

Essa abordagem reforçou a idéia de que classificar os períodos de baixa é bem mais difícil do que os de alta. A baixa eficácia e alta variabilidade com todos os métodos reforça a idéia de que o problema deve estar no modelo, ou seja, a discretização que não se adequa bem ao fenômeno estudado.

Tabela 14 – Avaliação da classificação com 5 classes na base completa

Métodos	Classe	Pre	Rec	Score-F1
Nearest Neighbors	-2	25,00%	25,00%	25,00%
	-1	28,57%	13,33%	18,18%
	0	59,52%	78,13%	67,57%
	1	71,43%	58,82%	64,52%
	2	33,33%	50,00%	40,00%
Naive Bayes	-2	12,50%	25,00%	16,67%
	-1	25,00%	13,33%	17,39%
	0	63,64%	65,63%	64,62%
	1	58,82%	58,82%	58,82%
	2	25,00%	50,00%	33,33%
Decision Tree	-2	50,00%	50,00%	50,00%
	-1	25,00%	13,33%	17,39%
	0	60,71%	53,13%	56,67%
	1	46,15%	70,59%	55,81%
	2	25,00%	50,00%	33,33%
Random Forest	-2	50,00%	50,00%	50,00%
	-1	55,56%	33,33%	41,67%
	0	73,53%	78,13%	75,76%
	1	65,00%	76,47%	70,27%
	2	33,33%	50,00%	40,00%
Logistic Regression	-2	40,00%	50,00%	44,44%
	-1	53,85%	46,67%	50,00%
	0	70,37%	59,38%	64,41%
	1	55,00%	64,71%	59,46%
	2	20,00%	50,00%	28,57%
Support Vector Classification	-2	0,00%	0,00%	0,00%
	-1	45,45%	33,33%	38,46%
	0	55,32%	81,25%	65,82%
	1	58,33%	41,18%	48,28%
	2	0,00%	0,00%	0,00%
Neural Network	-2	50,00%	25,00%	33,33%
	-1	41,67%	33,33%	37,04%
	0	62,86%	68,75%	65,67%
	1	65,00%	76,47%	70,27%
	2	0,00%	0,00%	0,00%

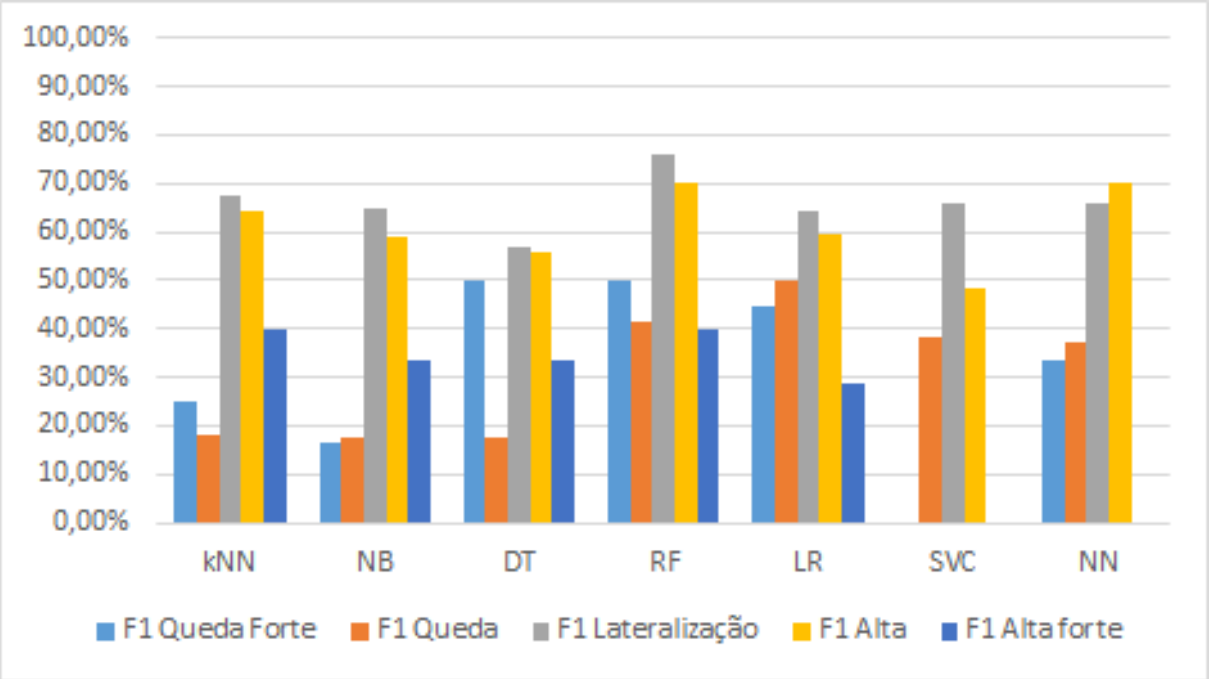

Figura 23 – Acurácia da classificação com 5 classes na base completa

Apesar dos resultados insatisfatórios, assim como no caso binário e com 3 classes, foi proposta uma base de dados mais restrita que considera apenas as variáveis com maior correlação. A matriz de coeficientes de Pearson nessa abordagem é dada por:

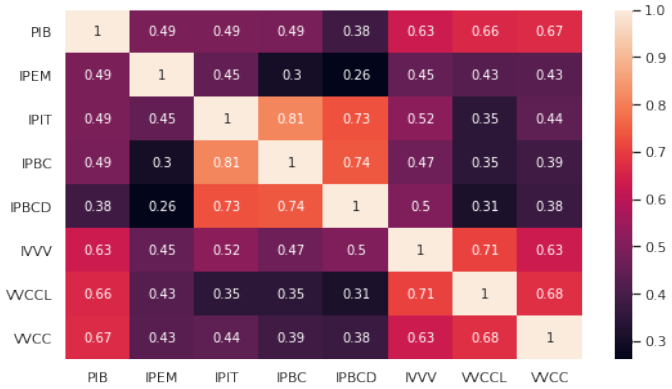

Figura 24 – Matriz de correlação das variáveis de maior correlação, considerando 5 classes

Nessa configuração de discretização e utilizando todas as variáveis propostas, a rede neural foi configurada utilizando os seguintes parâmetros:

Função de ativação Identity

Solver SGD

Número de camadas ocultas 2

Tabela 15 – Acurácia da classificação com 5 classes na base restrita

Métodos	Train	Test	ε
Nearest Neighbors	51,52%	55,71%	4,20%
Naive Bayes	57,58%	51,43%	-6,15%
Decision Tree	63,64%	52,86%	-10,78%
Random Forest	54,55%	60,00%	5,45%
Logistic Regression	57,58%	58,57%	1,00%
Support Vector Classification	60,61%	57,14%	-3,46%
Neural Network	57,58%	60,00%	2,42%

Em relação à acurácia, houve pouca melhora/piora de modo que podemos considerar dentro da margem de erro intrínseco. Assim como nos outros casos, olhando somente pelo ponto de vista da acurácia, faz sentido reduzir a base para apenas as variáveis de mais alta correlação.

Figura 25 – Acurácia da classificação com 5 classes na base restrita

Analisando a influência dessa base mais restrita em relação ao Score-F1, a tabela 16 mostra que o efeito foi mais significativo. Apesar de respostas com tendência média similar, os resultados aprensentam diferenças suficientes para não serem considerados resultados equivalentes. No entanto, cada um possuí vantagens e desvantagens em relação ao outro, tornando-se equivocado dizer que um foi melhor que o outro.

Tabela 16 – Avaliação da classificação com 5 classes na base restrita

Métodos	Classe	Pre	Rec	Score-F1
Nearest Neighbors	-2	33,33%	25,00%	28,57%
	-1	45,45%	33,33%	38,46%
	0	60,47%	81,25%	69,33%
	1	58,33%	41,18%	48,28%
	2	0,00%	0,00%	0,00%
Naive Bayes	-2	20,00%	50,00%	28,57%
	-1	20,00%	6,67%	10,00%
	0	63,89%	71,88%	67,65%
	1	61,54%	47,06%	53,33%
	2	33,33%	100,00%	50,00%
Decision Tree	-2	66,67%	50,00%	57,14%
	-1	41,67%	33,33%	37,04%
	0	60,61%	62,50%	61,54%
	1	52,94%	52,94%	52,94%
	2	20,00%	50,00%	28,57%
Random Forest	-2	66,67%	50,00%	57,14%
	-1	60,00%	40,00%	48,00%
	0	69,44%	78,13%	73,53%
	1	52,94%	52,94%	52,94%
	2	0,00%	0,00%	0,00%
Logistic Regression	-2	42,86%	75,00%	54,55%
	-1	64,29%	60,00%	62,07%
	0	73,91%	53,13%	61,82%
	1	52,63%	58,82%	55,56%
	2	28,57%	100,00%	44,44%
Support Vector Classification	-2	100,00%	25,00%	40,00%
	-1	50,00%	40,00%	44,44%
	0	58,14%	78,13%	66,67%
	1	57,14%	47,06%	51,61%
	2	0,00%	0,00%	0,00%
Neural Network	-2	0,00%	0,00%	0,00%
	-1	42,86%	40,00%	41,38%
	0	63,89%	71,88%	67,65%
	1	65,00%	76,47%	70,27%
	2	0,00%	0,00%	0,00%

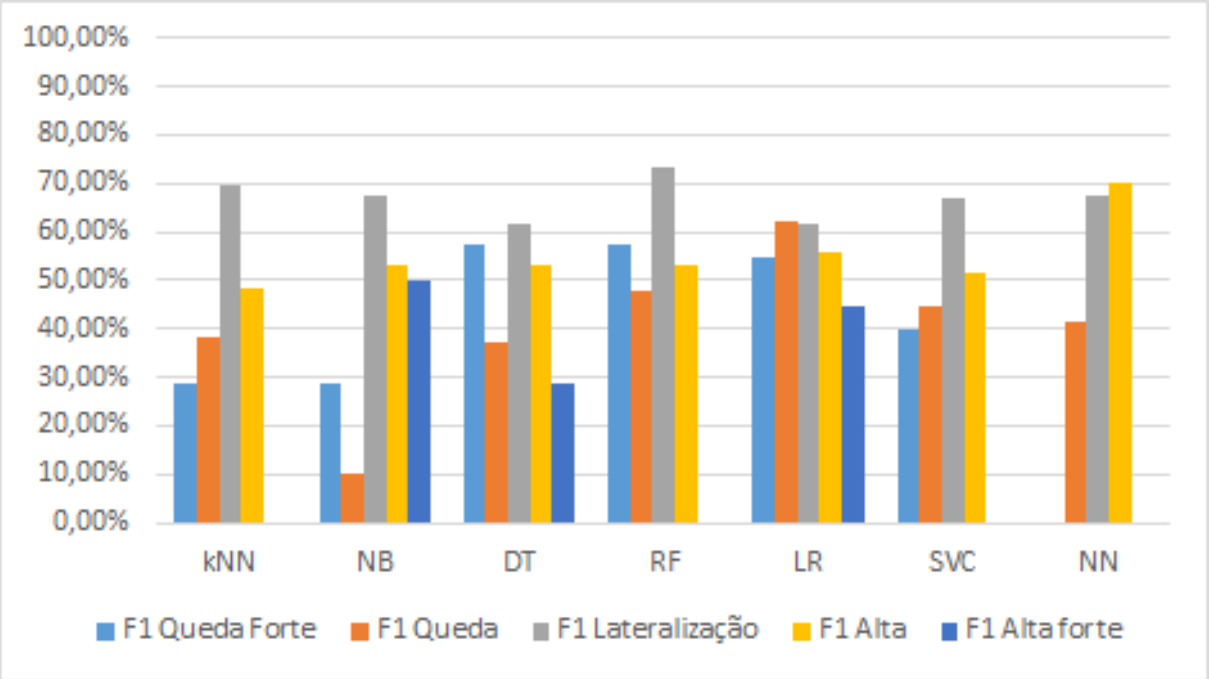

Figura 26 – Acurácia da classificação com 5 classes na base restrita

O gráfico 26 deixa claro que o modelo restrito com menos variáveis tem muito mais dificuldade para identificar as classes extremais. Isso levanta uma hipótese de que em cenários muito complexos, essas variáveis de baixa correlação que foram removidas, de fato tenham um papel relevante em capturar nuances referentes as não linearidades do modelo completo.

Apesar de nos modelos considerando classificação binária e com 3 classes, se ter identificado que as variáveis de menor correlação são praticamente irrelevantes para o modelo, talvez em situações mais extremas, essa riqueza adicional no modelo, apesar de parecer pequena pode ser relevante. Nesse sentido, cabe extender as 2 abordagens em paralelo, enquanto o dataset temporal cresce.

5.3 Comparação entre cenários

Para comparar os diversos cenários apresentados, é proposta uma análise comparativa entre todos os métodos, em relação à acurácia alcançada nas etapas de treinamento e de teste.

Além disso, uma análise mais detalhada do comportamente do score-F1 tanto em relação as diferentes discretizações, quanto ao que se refere ao conjunto de variáveis explicativas utilizadas em cada cenário.

5.3.1 Acurácia

Para analisar a acurácia na etapa de treinamento é preciso comparar todos os métodos propostos com todas as aboradagens de discretização.

Tabela 17 – Comparação de acurácia na etapa de treinamento

Métodos	2CC	2CR	3CC	3CR	5CC	5CR
kNN	78,79%	69,70%	78,79%	78,79%	51,52%	51,52%
NB	75,76%	72,73%	84,85%	78,79%	57,58%	57,58%
DT	72,73%	75,76%	60,61%	69,70%	51,52%	63,64%
RF	72,73%	72,73%	78,79%	69,70%	54,55%	54,55%
LR	75,76%	72,73%	81,82%	78,79%	63,64%	57,58%
SVC	72,73%	75,76%	81,82%	75,76%	69,70%	60,61%
NN	72,73%	75,76%	81,82%	75,76%	63,64%	57,58%

A tabela 17 mostra que houve considerável homogenidade nas medidas, apesar da diversidade de métodos e cenários apresentados. A figura 27 corrobora a idéia de que há uma maior dificuldade em classificar os modelos mais complexos, contendo 5 classes.

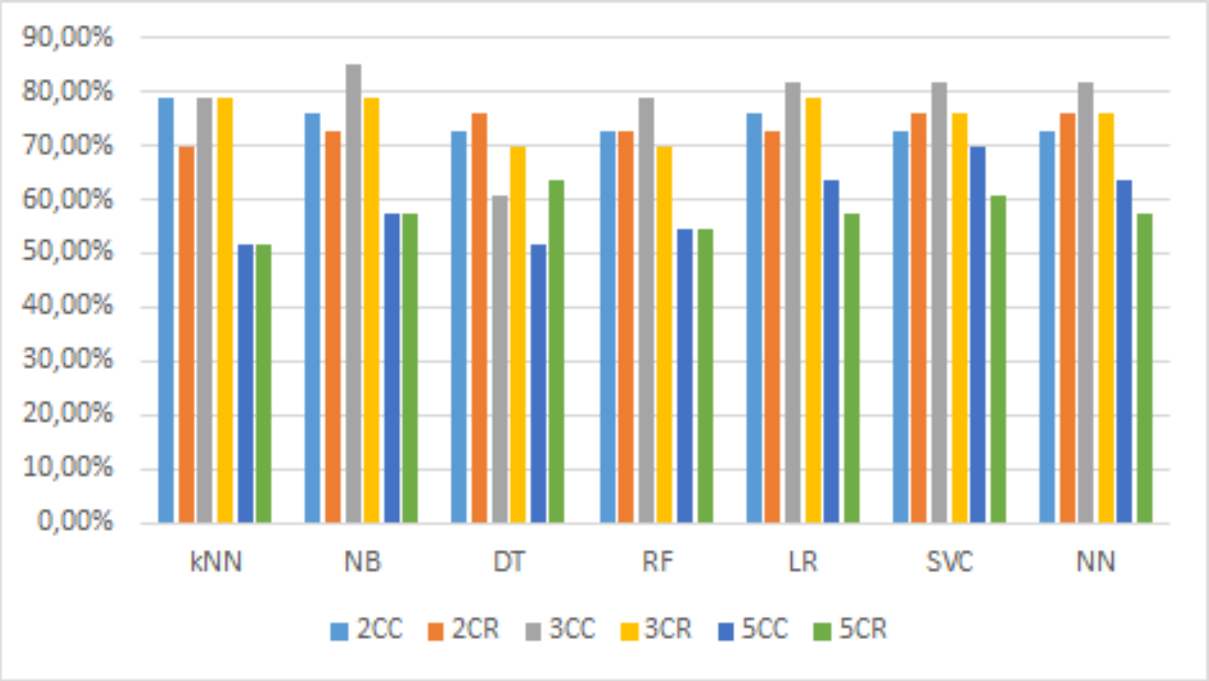

Figura 27 – Comparação de acurácia na etapa de treinamento

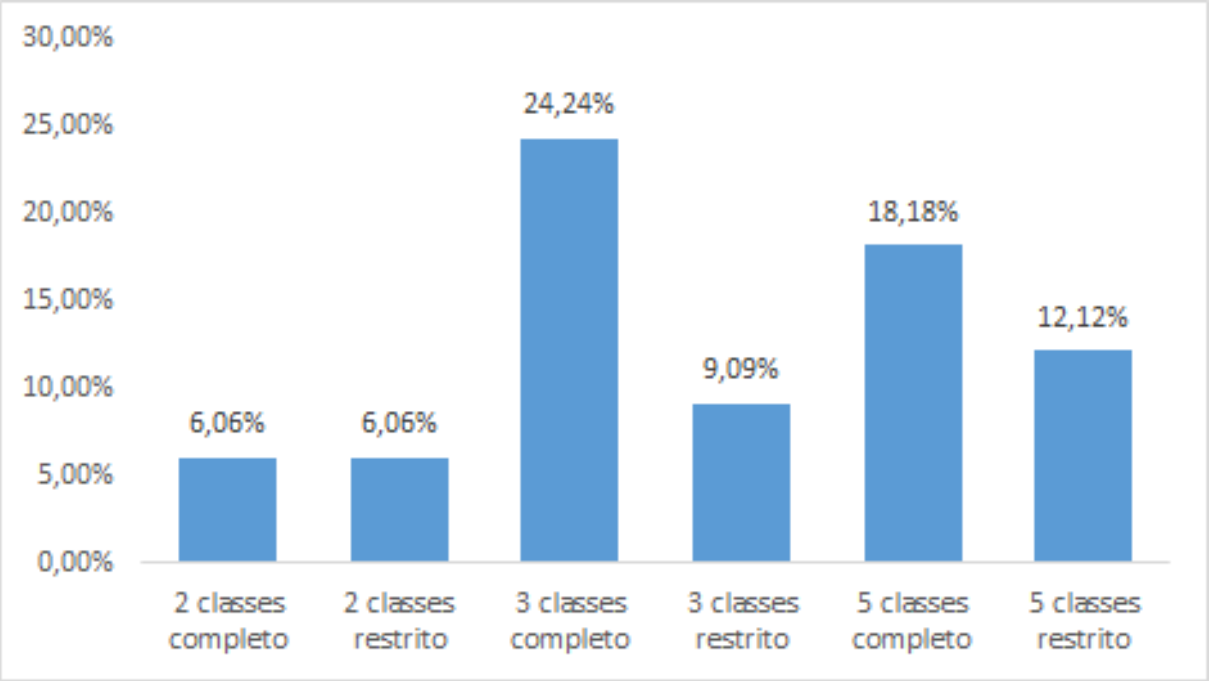

Figura 28 – Variação de acurácia entre os métodos na etapa de treinamento

Analisando a figura 28, podemos perceber um fato interessante. Para as discretizações com 3 e 5 classes, reduzir o número de variáveis explicativas trouxe mais homogenidade entre os métodos.

Analogamente, a tabela 18 mostra que houve considerável homogenidade nas medidas, apesar da diversidade de métodos e cenários apresentados. Houve uma ligeira melhora no cenário considerando classificação binária, provavelmente decorrente de viés como discutido na seção anterior.

Tabela 18 – Comparação de acurácia na etapa de teste

Métodos	2CC	2CR	3CC	3CR	5CC	5CR
kNN	90,00%	90,00%	81,43%	81,43%	55,71%	55,71%
NB	87,14%	87,14%	78,57%	80,00%	50,00%	51,43%
DT	80,00%	88,57%	72,86%	75,71%	48,57%	52,86%
RF	88,57%	90,00%	82,86%	75,71%	65,71%	60,00%
LR	84,29%	87,14%	68,57%	75,71%	57,14%	58,57%
SVC	88,57%	90,00%	81,43%	84,29%	54,29%	57,14%
NN	85,71%	85,71%	84,29%	82,86%	58,57%	60,00%

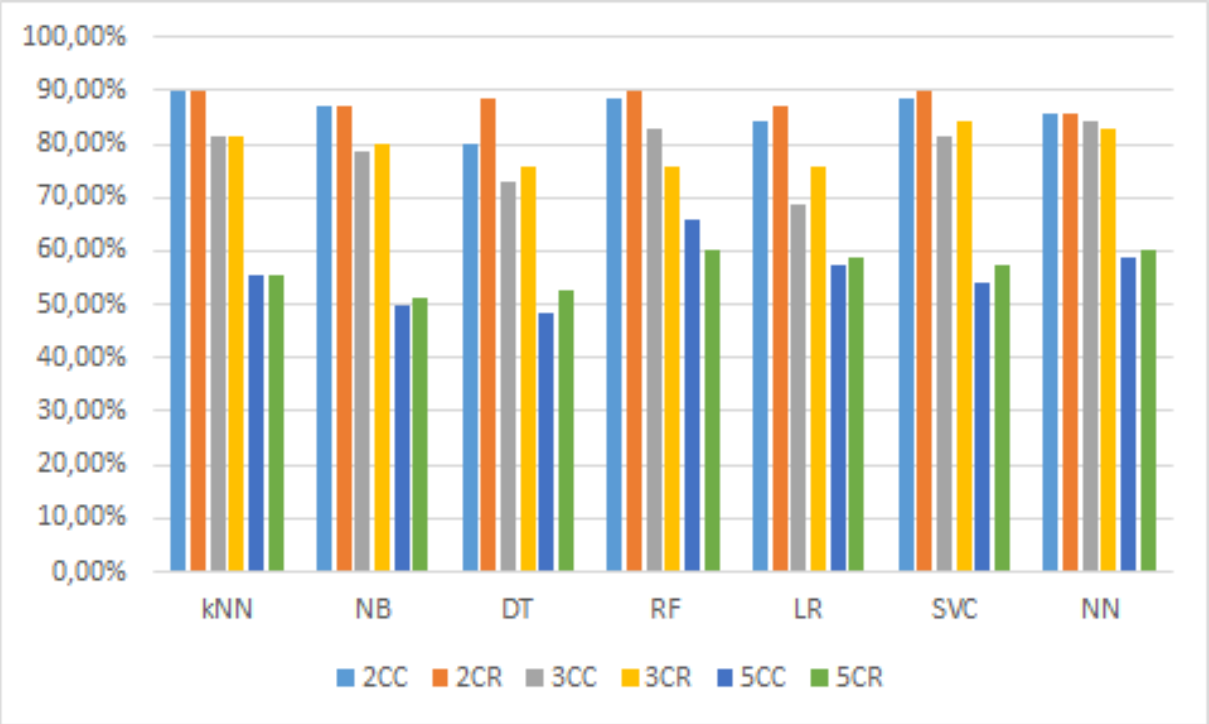

Figura 29 – Comparação de acurácia na etapa de treinamento

Assim como no treinamento, a figura 29 reforça o conceito da dificuldade na modelagem mais complexa também para a etapa de teste.

Por outro lado, a figura 28 deixa claro que reduzir o número de variáveis explicativas trouxe mais homogenidade entre os métodos, independente do cenário.

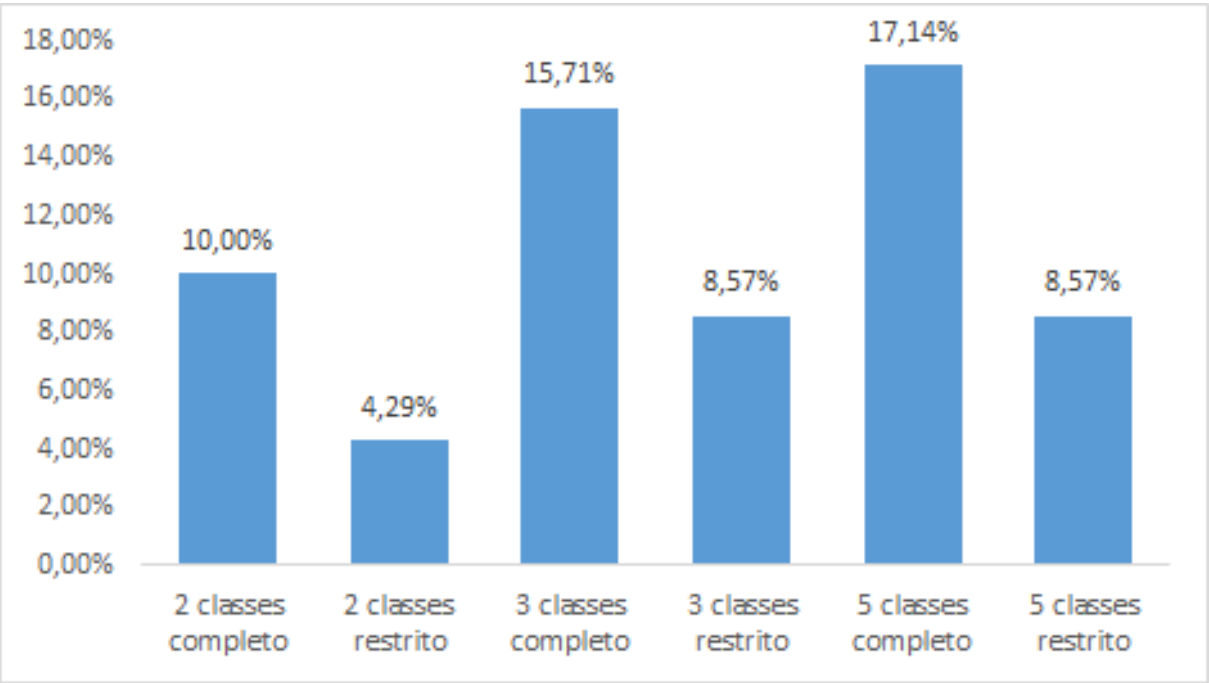

Figura 30 – Variação de acurácia entre os métodos na etapa de treinamento

5.3.2 Score-F1

Como discutido na seção anterior, a qualidade de cada método em relação a cada cenário é relativa, a depender do critério. No entanto, é importante comparar o efeito que cada cenário teve no desempenho de cada método. Dessa forma, será apresentado a melhora percentual quando aplicada a base que considera um número reduzido de variáveis explicactivas.

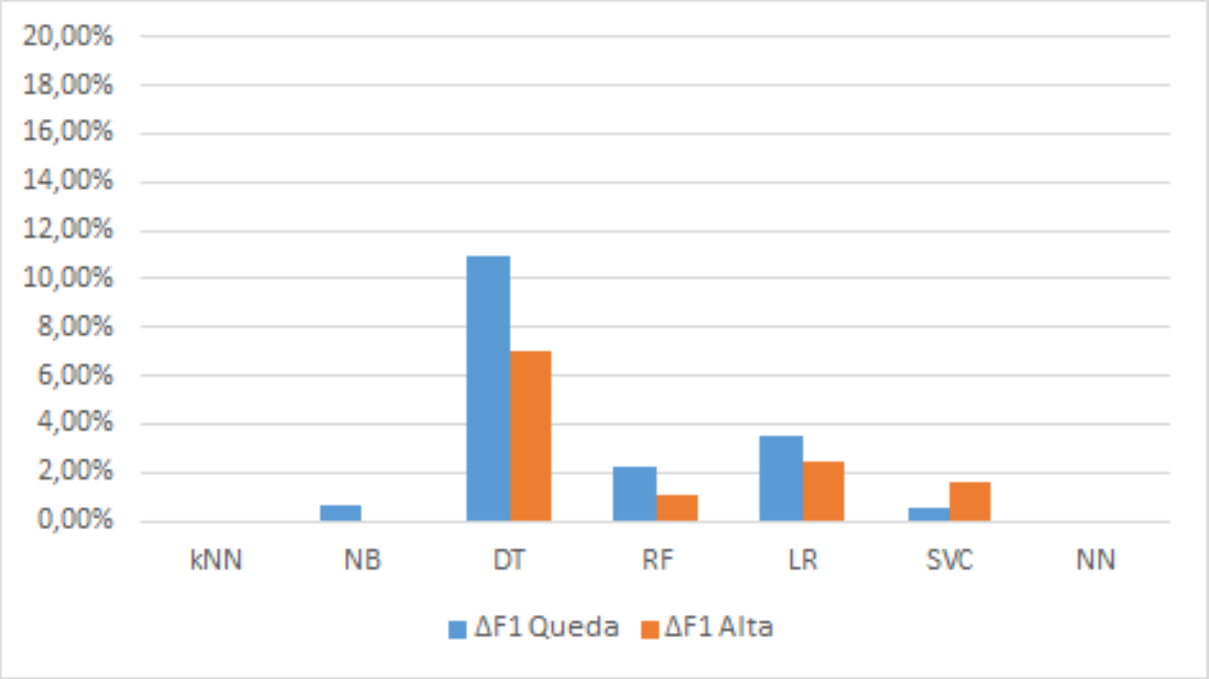

Figura 31 – Variação de Score-F1 entre os métodos com discretização binária

Considerando a discretização binária, podemos verificar pela figura 31 que praticamente todos os métodos tiveram um desempenho parecido quando aplicada a restrição de variáveis explicativas.

Apesar da árvore de decisão ter apresentado uma melhora mais significativa, apesar de significativo, o valor médio da ordem de 9% não chega a ser absurdo. Como a árvore de decisão é uma técnica cujos resultados podem ser interpretados (sob certas circunstâncias), cabe uma análise mais minuciosa desse resultado.

Avaliando o impacto da restrição de variáveis com a discretização de 3 classes, a figura 32 mostra que houve uma variação não desprezível, mas ainda dentro de uma variação média compatível com o caso anterior. Como houveram casos em que houve melhora e outros em que houve piora, não é possível afirmar que a restrição se justifica.

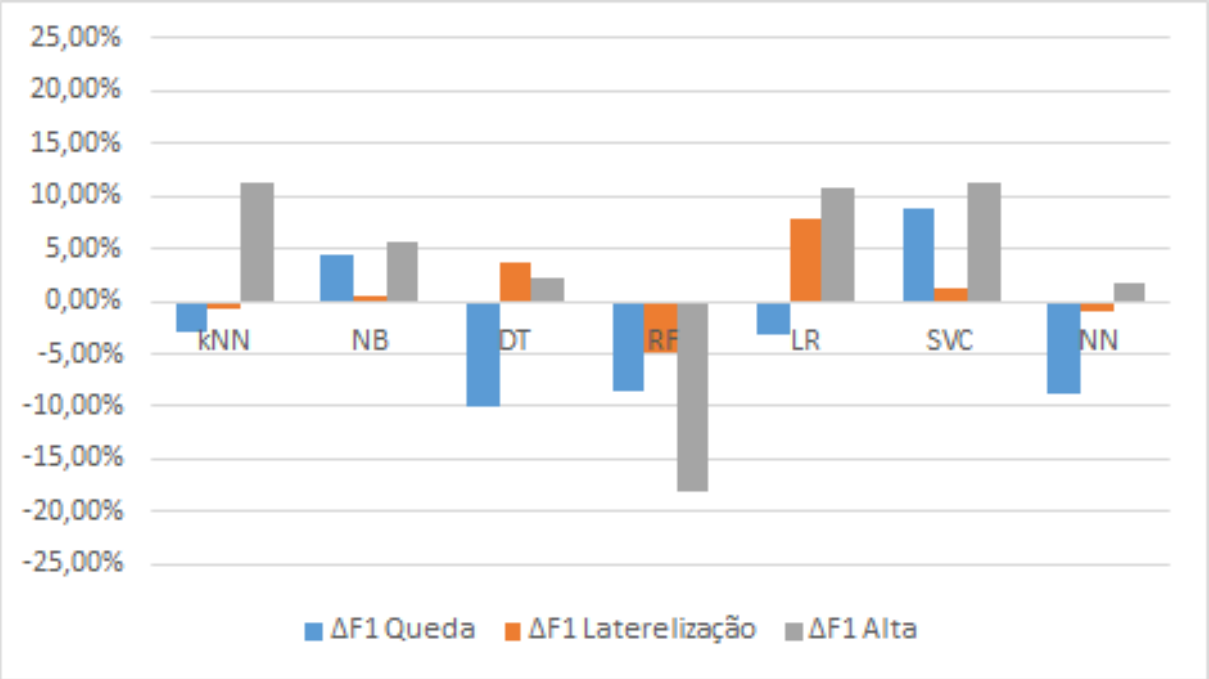

Figura 32 – Variação de Score-F1 entre os métodos com discretização considerando 3 classes

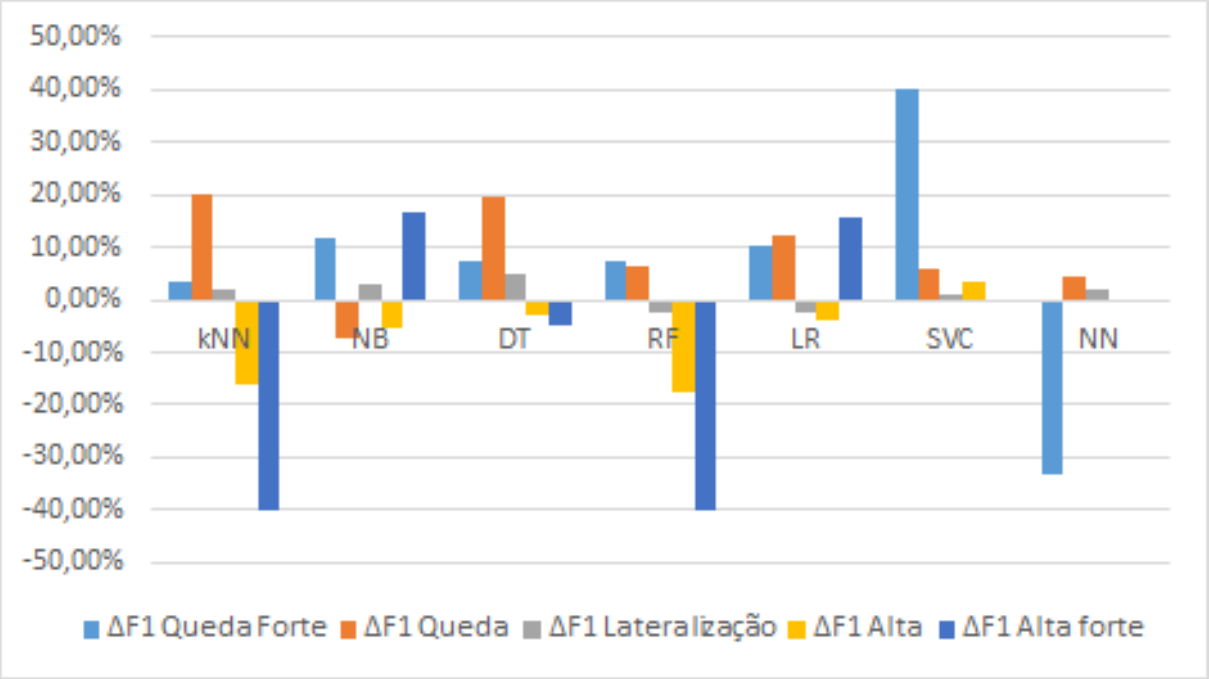

Figura 33 – Variação de Score-F1 entre os métodos com discretização considerando 5 classes

Finalmente, analisando a figura 33, percebemos uma influência relevante, mas não necessariamente positiva. Por um lado, as classes de queda tem uma ligeira melhora (sendo

bem significativa com relação ao SVC), mas por outro lado, as classes relativas a alta tem uma piora gigantesca.

No entanto, é importante ressaltar que SVC e Neural Network não convergiram para algumas classes, além do que a magnitude da piora em alguns casos beira os 40%. Essa alta variância mostra que a limitação está muito mais ligada a erros de modelagem do que a capacidade preditiva dos métodos.

Tabela 19 – Comparação de Score-F1 em diferentes cenários

classes	2 classes		3 classes		5 classes	
	absoluto	relativo	absoluto	relativo	absoluto	relativo
-2					46,53%	6,65%
-1			-20,14%	-2,88%	61,26%	8,75%
0	17,85%	2,55%	6,51%	0,93%	7,67%	1,10%
1	11,82%	1,69%	24,55%	3,51%	-42,50%	-6,07%
2					-52,22%	-7,46%
Total	29,67%	4,24%	10,92%	1,56%	20,74%	2,96%

Na tentativa de resumir qual foi o ganho absoluto de cada abordagem, apesar da alta variância, a tabela 19 combina os ganhos líquidos absolutos de cada classe em relação a cada cenário. Como são considerados 7 métodos, o ganho relatívo é justamente o ganho líquido médio entre os 7 modelos.

Apesar da melhora significativa no caso binário, já foi discutido que é uma abordagem muito grosseira que não apresenta um interesse prático direto. Por outro lado, a abordagem considerando 5 classes, além de apresentar um desempenho muito caótico, ainda teve problemas de convergência, sugerindo que a abodagem não é adequada.

Por tudo que foi exposto, o cenário que considera 3 classes teve um ganho efetivo de desempenho quando a base foi restrita a menos variáveis significativas. No entanto, a melhora relativa foi muito pequena, sendo pouco significativa. Portanto, pode-se optar por remove-las, mas também é válido investigar outras transformações, pois podem tornar os fenômenos mais compatíveis.

6 Conclusão

As abordagens apresentadas mostraram-se promissoras no sentido de conseguirem caracterizar de forma satisfatória as variações do PIB. No entanto, no decorrer do desenvolvimento, foram percebidas alguns equívocos que devem ser corrigidos com prioridade:

- Normalização das variáveis
- Classes de discretização
- Conjunto de treinamento/teste
- Setup dos hiperparâmetros

Apesar de muitas variáveis terem apresentado empírico de baixa correlação, conceitualmente, essa correlação parece não ser desprezível. Portanto, é preciso realizar um estudo mais profundo sobre como esses indicadores são utilizados na prática econômica, para manter alta correlação sem perder a interpretabilidade.

Outro problema possível tem haver com o conceito de "**one hot encoding**", ou seja, a metodologia empregada para categorização das classes. A escolha de rótulos negativos, apesar de intuitiva, pode ter implicações erráticas do ponto de vista numérico. Portanto, os resultados devem ser repetidos utilizando variáveis "Dummy"não-negativas.

A divisão do conjuntos de treinamento possuí alguns problems de concepção, pois o cross validation foi implementado de maneira a ignorar a hierarquia temporal dos dados. Para corigir esse problema, é proposto a aplicação do conceito conhecido como "Forward Chaining". Além disso, é preciso adequar essa implementação para corrigir o problema de "**look-ahead bias**", tanto no conjunto de treinamento quanto no conjunto de teste.

Finalmente, depois de implementar todas essas correções, é preciso reavaliar todos os resultados obtidos até então, para só então fazer o ajuste fino dos hiperparâmetros de todos os métodos sugeridos.

Por enquanto o estudo comparativo de cenários ainda é demasiadamente complexo e demorado, já que a implementação codificada até o momento possui baixo grau de automação. Existem limitações conceituais tanto na etapa de aquisição/armazenamento dos dados, quanto na exportação dos resultados consolidados, sendo necessário utilizar outra ferramenta externa ao Python para tal.

É fundamental estudar mais profundamente a ferramenta, para que a produtividade aumente e as análises possam ficar mais profundas. A estruturação do database utilizando SQL, o encapsulamento de tarefas por meio de funções implícitas e o dominio sobre as ferramentas de tratamento de variáveis e plots, referem-se aos objetivos mais importantes a se desenvolver para aumentar as perspectivas do projeto dentro do tempo restante.

6.1 Trabalhos futuros

Além da abordagem baseada em classificação, espera-se expandir o trabalho para uma abordagem baseada em regressão. Um estudo comparativo entre técnicas econométricas e modelos de aprendizagem estatística para conseguir quantificar os movimentos previstos pela etapa de classificação.

Há também a idéia de aplicar conceitos de aprendizagem não-supervisionada para caracterizar os movimentos de recessão, uma vez que não foi encontrado um indicador específico para isso nas bases nacionais.

Referências

ADRIAN, S. C. V.; EDUCATION, P. *The Long Range Impact of the Recession on Families*. 2010. Citado na página 32.

ADVISORS, C. C. of E. *Assessing the State of the Economy in Real Time Using Headline Economic Indicators*. 2017. Citado na página 32.

AGARWALA, S. K. *Principles of Economics*. 2. ed. [S.l.]: Excel Books, 2009. ISBN 9788174466921. Citado na página 30.

ALMEIDA, B.; NEVES, R.; HORTA, N. Combining support vector machine with genetic algorithms to optimize investments in forex markets with high leverage. *Applied Soft Computing Journal*, v. 64, n. 1, p. 596–613, March 2018. ISSN 1568-4946. Citado na página 20.

ALPAYDIN, E. *Introduction to Machine Learning*. 3. ed. Cambridge, MA: MIT Press, 2014. (Adaptive Computation and Machine Learning). ISBN 978-0-262-02818-9. Citado na página 26.

BERGE, T. J. *Predicting recessions with leading indicators: model averaging and selection over the business cycle*. [S.l.], 2013. Citado 3 vezes nas páginas 21, 22 e 33.

BURNS, A. F.; MITCHELL, W. C. *Measuring Business Cycles*. National Bureau of Economic Research, Inc, 1946. Disponível em: ⟨https://EconPapers.repec.org/RePEc:nbr:nberbk:burn46-1⟩. Citado 2 vezes nas páginas 17 e 31.

CANELAS, A.; NEVES, R.; HORTA, N. A sax-ga approach to evolve investment strategies on financial markets based on pattern discovery techniques. *Expert Syst. Appl.*, Pergamon Press, Inc., USA, v. 40, n. 5, p. 1579–1590, abr. 2013. ISSN 0957-4174. Disponível em: ⟨https://doi.org/10.1016/j.eswa.2012.09.002⟩. Citado na página 20.

CFNAI. *Chicago Fed National Activity Index (CFNAI) - Federal Reserve Bank of Chicago*. 2020. Disponível em: ⟨https://www.chicagofed.org/publications/cfnai/index⟩. Citado na página 18.

CROARKIN, C.; TOBIAS, P. *NIST/SEMATECH e-Handbook of Statistical Methods*. [s.n.], 2012. Disponível em: ⟨http://www.itl.nist.gov/div898/handbook/⟩. Citado na página 24.

DESIKAN, P.; SRIVASTAVA, J. Time series analysis and forecasting methods for temporal mining of interlinked documents. In: . [S.l.: s.n.], 2005. Citado na página 25.

ESTRELLA, A.; MISHKIN, F. S. *Predicting U.S. Recessions: Financial Variables as Leading Indicators*. [S.l.], 1995. (Working Paper Series, 5379). Disponível em: ⟨http://www.nber.org/papers/w5379⟩. Citado 4 vezes nas páginas 17, 21, 23 e 33.

GORGULHO, A.; NEVES, R. F.; HORTA, N. Applying a ga kernel on optimizing technical analysis rules for stock picking and portfolio composition. *Expert Syst. Appl.*, v. 38, n. 11, p. 14072–14085, 2011. Disponível em: ⟨http://dblp.uni-trier.de/db/journals/eswa/eswa38.html#GorgulhoNH11⟩. Citado na página 20.

HEIJ, C.; BOER, P. D.; FRANSES, P. H.; KLOEK, T.; DIJK, H. K. Econometric methods with applications in business and economics. In: . [S.l.: s.n.], 2004. Citado na página 27.

IMF. *World Economic Outlook, October 2020: Cyclical Upswing, Structural Change*. 2020. Citado na página 17.

KAUPPI, H.; SAIKKONEN, P. Predicting u.s. recessions with dynamic binary response models. *The Review of Economics and Statistics*, v. 90, p. 777–791, 02 2008. Citado 3 vezes nas páginas 18, 22 e 33.

KISINBAY, T.; BABA, C. *Predicting Recessions; A New Approach for Identifying Leading Indicators and Forecast Combinations*. [S.l.], 2011. Disponível em: ⟨https://ideas.repec.org/p/imf/imfwpa/2011-235.html⟩. Citado 4 vezes nas páginas 21, 22, 23 e 33.

LIU, W.; MOENCH, E. What predicts us recessions? *International Journal of Forecasting*, v. 32, n. 4, p. 1138–1150, 2016. Disponível em: ⟨https://EconPapers.repec.org/RePEc:eee:intfor:v:32:y:2016:i:4:p:1138-1150⟩. Citado 3 vezes nas páginas 22, 23 e 33.

NBER. *The NBER Business Cycle Dating Committee*. 2010. Disponível em: ⟨https://www.nber.org/cycles/recessions.html⟩. Citado na página 31.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, v. 135, n. 3, p. 370–384, 1972. Disponível em: ⟨https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2344614⟩. Citado na página 27.

OECD Economic Outlook, Volume 2018 Issue 1. [S.l.]: OECD Publishing, 2018. ISBN 92-64-30873-3. Citado na página 30.

O'NEILL, B. L.; XIAO, J. Financial behaviors before and after the financial crisis: Evidence from an online survey. *Macroeconomics: Prices*, 2012. Citado na página 32.

RITTENBERG, L.; TREGARTHEN, T. *Principles of Macroeconomics*. 1. ed. [S.l.]: Flat World Knowledge, 2009. Citado na página 30.

ROSSER J. BARKLEY, J.; ROSSER, M. V. *Comparative Economics in a Transforming World Economy, third edition*. The MIT Press, 2018. v. 1. (MIT Press Books, 0262037335). ISBN ARRAY(0x4fe595c8). Disponível em: ⟨https://ideas.repec.org/b/mtp/titles/0262037335.html⟩. Citado na página 30.

RUDEBUSCH, G.; WILLIAMS, J. Forecasting recessions: The puzzle of the enduring power of the yield curve. *Journal of Business & Economic Statistics*, v. 27, n. 4, p. 492–503, 2009. Disponível em: ⟨https://EconPapers.repec.org/RePEc:bes:jnlbes:v:27:i:4:y:2009:p:492-503⟩. Citado 2 vezes nas páginas 18 e 33.

SILVA, A.; NEVES, R.; HORTA, N. A hybrid approach to portfolio composition based on fundamental and technical indicators. *Expert Syst. Appl.*, Pergamon Press, Inc., USA, v. 42, n. 4, p. 2036–2048, mar. 2015. ISSN 0957-4174. Disponível em: ⟨https://doi.org/10.1016/j.eswa.2014.09.050⟩. Citado na página 20.

SILVIA, J. E. *Can Machine Learning Improve Recession Prediction? Big Data Applications in Economics: Part III | Economics Group Special Commentary*. [S.l.], 2018. Citado 3 vezes nas páginas 21, 22 e 23.

STOCK, J. H.; WATSON, M. W. Forecasting inflation. *Journal of Monetary Economics*, v. 44, n. 2, p. 293–335, 1999. ISSN 0304-3932. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0304393299000276⟩. Citado na página 18.

TAY, F. E. H.; CAO, L. Application of support vector machines in financial time series forecasting. *Omega*, v. 29, n. 4, p. 309–317, August 2001. Disponível em: ⟨https://ideas.repec.org/a/eee/jomega/v29y2001i4p309-317.html⟩. Citado na página 24.