MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
GOIANO – CÂMPUS RIO VERDE

MÁRIO DOUGLAS ALVES CABRAL

ANÁLISE DE SÉRIES NUMÉRICAS UTILIZANDO APRENDIZADO NÃO SUPERVISIONADO

RIO VERDE – GO
2019

MÁRIO DOUGLAS ALVES CABRAL

ANÁLISE DE SÉRIES NUMÉRICAS UTILIZANDO APRENDIZADO NÃO SUPERVISIONADO

Trabalho de Conclusão de Curso apresentado ao Instituto Federal de Educação, Ciência e Tecnologia Goiano – Câmpus Rio Verde ligado ao Ministério da Educação, como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: MSc. Adriano Soares de Oliveira Bailão
Instituto Federal de Educação, Ciência e Tecnologia Goiano – Câmpus Rio Verde

RIO VERDE – GO
2019

Sistema desenvolvido pelo ICMC/USP
Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas - Instituto Federal Goiano

Cabral, Mário Douglas Alves
C117a Análise de Séries Numéricas utilizando Aprendizado Não Supervisionado / Mário Douglas Alves Cabral;orientador Adriano Soares de Oliveira Bailão. -- Rio Verde, 2019.
71 p.

Monografia (em Ciência da Computação) -- Instituto Federal Goiano, Campus Rio Verde, 2019.

1. Aprendizado de máquina . 2. Séries numéricas. 3. Classificador. 4. Modelo de classificação. I. Bailão, Adriano Soares de Oliveira , orient. II. Título.

Responsável: Johnathan Pereira Alves Diniz - Bibliotecário-Documentalista CRB-1 n°2376

INSTITUTO FEDERAL
Goiano

Repositório Institucional do IF Goiano - RIIF Goiano
Sistema Integrado de Bibliotecas

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610/98, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano, a disponibilizar gratuitamente o documento no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

Identificação da Produção Técnico-Científica

[ ] Tese
[ ] Dissertação
[ ] Monografia – Especialização
[x] TCC - Graduação
[ ] Produto Técnico e Educacional - Tipo: _____
[ ] Artigo Científico
[ ] Capítulo de Livro
[ ] Livro
[ ] Trabalho Apresentado em Evento

Nome Completo do Autor: Mário Douglas Alves Cabral
Matrícula: 2016102192010412
Título do Trabalho: Análise de Séries Numéricas utilizando Aprendizado Não Supervisionado

Restrições de Acesso ao Documento

Documento confidencial: [x] Não [ ] Sim, justifique: _____

Informe a data que poderá ser disponibilizado no RIIF Goiano: 19/02/20
O documento está sujeito a registro de patente? [ ] Sim [x] Não
O documento pode vir a ser publicado como livro? [ ] Sim [x] Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a autor/a declara que:

1. o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
2. obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
3. cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Rio Verde, 18/02/2020
Local Data

Mário Douglas Alves Cabral
Assinatura do Autor e/ou Detentor dos Direitos Autorais

Ciente e de acordo:

_____
Assinatura do(a) orientador(a)

INSTITUTO FEDERAL GOIANO – CÂMPUS RIO VERDE
DIRETORIA DE ENSINO
GERÊNCIA DE ENSINO DE GRADUAÇÃO

ATA DE DEFESA DO TRABALHO DE CURSO (TC)

ANO	SEMESTRE
2019	02

No dia 16 do mês de dezembro de 2019, às 11 horas e 55 minutos, reuniu-se a banca examinadora composta pelos docentes Adriano Soares Bailão, Douglas Cedrim Oliveira e Fábio Montanha Ramos, para examinar o Trabalho de Curso (TC) intitulado Análise de Séries Numéricas Utilizando Aprendizado não Supervisionado do(a) acadêmico(a) Mário Douglas Alves Cabral, Matrícula nº _____ do curso de Ciência da Computação do IF Goiano – Câmpus Rio Verde. Após a apresentação oral do TC, houve arguição do candidato pelos membros da banca examinadora. Após tal etapa, a banca examinadora decidiu pela APROVAÇÃO do(a) acadêmico(a). Ao final da sessão pública de defesa foi lavrada a presente ata, que segue datada e assinada pelos examinadores.

Rio Verde, 16 de dezembro de 2019

Adriano S. C. Bailão
Nome:
Orientador(a)

Fábio Montanha Ramos
Nome:
Membro

Douglas Cedrim Oliveira
Nome:
Membro

Observação:
( ) O(a) acadêmico(a) não compareceu à defesa do TC.

Dedico este trabalho aos meus pais, que sempre me apoiaram nesta jornada de estudos.

AGRADECIMENTOS

A Deus, que tornou possível tudo aquilo que me parecia impossível.

Aos meus pais, por sempre incentivar meus estudos e todo apoio financeiro que me deram, durante toda essa jornada.

A todas as amizades que adquiri ao longo desses anos, e por toda ajuda que tive, em momentos difíceis.

Ao meu orientador prof. Msc. Adriano Soares de Oliveira Bailão, pela paciência, cooperação, apoio, e todos os ensinamentos e experiências que me passou durante todo o desenvolvimento deste projeto.

A todos aqueles que, direta ou indiretamente, contribuíram ao longo dessa caminhada.

RESUMO

CABRAL, Mário Douglas Alves Cabral. Análise de Séries Numéricas utilizando Aprendizado Não Supervisionado. 2019. 71 f. Trabalho de Conclusão de Curso – , Instituto Federal de Educação, Ciência e Tecnologia Goiano – Câmpus Rio Verde. Rio Verde – GO, 2019.

Análise de série numéricas é uma tarefa que traz uma grande demanda de tempo e custo e, portanto, torna-se necessário ferramentas que agilizem e diminuam o custo dessas análises. O respectivo trabalho apresenta a concepção de um modelo de classificação para o conjunto de série numéricas, possibilitando-se assim a automatização de análise de série numérica, principalmente em séries que possuam grande quantidade de elementos, pois uma das fases do modelo culmina na compressão da série. Verificou-se que, o classificador possui baixo desempenho em séries numéricas, como possuem poucos elementos, porém constatou-se que o desempenho foi alto em série numérica, com muitos elementos. A partir desses resultados podemos concluir que, embora o desempenho do classificador seja baixa em séries pequenas, no contexto de várias áreas geram dados no formato de séries numéricas, o modelo de classificação traria grandes benefícios, como uma ferramenta de apoio de decisão, em variadas áreas.

Palavras-chave: Aprendizado de máquina. Séries numéricas. Classificador. Modelo de classificação.

ABSTRACT

CABRAL, Mário Douglas Alves Cabral. Analysis of Numerical Series Using Unsupervised Learning. 2019. 71 f. Trabalho de Conclusão de Curso – , Instituto Federal de Educação, Ciência e Tecnologia Goiano – Câmpus Rio Verde. Rio Verde – GO, 2019.

Numerical series analysis is a task that brings a great demand of time and cost and, therefore, it becomes necessary tools that speed up and reduce the cost of these analyzes. The respective work presents the design of a classification model for the set of numerical series, thus enabling the automation of numerical series analysis, especially in series that have a large number of elements, as one of the phases of the model culminates in the compression of series. It was found that the classifier has low performance in numerical series, as they have few elements, however it was found that the performance was high in numerical series, with many elements. From these results we can conclude that, although the classifier's performance is low in small series, in the context of several areas they generate data in the form of numerical series, the classification model would bring great benefits, as a decision support tool, in several areas.

Keywords: Machine learning. Numerical series. Classifier. Classification model.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Ao longo dos anos, várias tecnologias vêm sendo desenvolvidas para coleta de dados, muitos desses estão em formato de série numérica. Uma importante tarefa para que trabalha com séries numéricas é a sua análise, um trabalho que demanda tempo para ser feito, e sendo feito por mão-de-obra humana pode acarretar erros em sua análise, o que em algumas áreas pode trazer um grande risco, como por exemplo, análise de tráfego aéreo.

Nesse contexto, o aprendizado de máquina introduziria um grande auxílio para análise destas séries numéricas, com o intuito da automatização dessa tarefa, visto que não foi encontrado nenhum modelo de classificação de séries numéricas na literatura.

A área análise ou ciência de dados dispõe de uma série de conhecimentos, como a compreensão do domínio do problema e conhecimento estatístico. Entretanto, as ferramentas de análise de dados disponíveis, são direcionadas a domínios e tipos de dados específicos. Assim as abordagens de análises de dados baseadas em medição da informação, têm possibilitado menos interferência do analista no processo de análise, a partir do tratamento genérico das fontes de dados.

O *framework* como *DAMICORE (Data Mining Of Code Repositories)* introduzido por Sanches, Cardoso e Delbem (2011), utiliza aproximações chamadas de *compactadores*, para a medição de informações em objetos de fontes de dados não estruturados. *Compactadores* utilizam técnicas adaptativas e estatísticas de compressão *sem perdas*. Embora, os *Frameworks* baseados em *compactadores* possuam alta capacidade de generalização para o tratamento das fontes de dados, em aplicações de Reconhecimento de Padrões (RP), o modelo de representação dos objetos de dados com base neste tipo de aproximação, torna a abordagem dependente do algoritmo que o compactador encapsula, levando assim a implicações como por exemplo, ambiguidade na representação das amostras, falta de precisão em problemas de classificação e resultados indeterminísticos. Uma forma de contornar o problema é desmembrando o algoritmo implementado pelo Perfil compactador em duas fases, uma etapa probabilística encarregada da identificação e contabilização das ocorrências dos componentes elementares de uma série numérica, e a outra etapa é responsável por sua codificação.

Portanto esse trabalho propõe o desenvolvimento de um classificador de séries numéricas, por meio de extensão do fluxo de operação do *Framework DAMICORE* de forma a direcionar o *bias* de uma aplicação de RP a partir de uma fase chamada *Granulação* e outra fase chamada *Codificação* além da extensão do modelo para a aplicação de técnicas de compressão *com perdas*. Como resultado, processos de RP como formação de agrupamentos e classificação, podem incorporar em seus modelos a análise de fontes de dados estruturados e semi estruturados, ganho qualitativo na formação e visualização dos agrupamentos e melhora da delimitação das fronteiras (limites de decisão) entre grupos de objetos.

O objetivo geral desta pesquisa é a detecção de padrões em Séries Numéricas, juntamente com a definição do método para compressão numérica, evitando assim a maldição de dimensionalidade. Como objetivos específicos: definir o modelo de representação do conjunto de séries numéricas e vetores numéricos, oferecer um modelo de classificação para um conjunto de séries numéricas e vetores numéricos, visualizar as correlações com padrões encontrados.

1.1 Estrutura do trabalho

No capítulo 2 são apresentados os principais conceitos utilizados nesse trabalho, no capítulo 3 são definidos os materiais, assim como foi realizado a sua coleta e o desenvolvimento dos métodos utilizados no trabalho, no capítulo 4 são expostos os resultados dos testes realizados com o classificador assim como a discussão dos resultados, no capítulo 5 apresenta-se a conclusão geral do trabalho e sugestão para trabalhos futuros.

2 Revisão Bibliográfica

Na atualidade, várias áreas de conhecimento vêm ao longo dos anos acumulando dados, com um imenso volume de informações, onde muitas delas estão em formato de séries sequenciais, justamente por serem dados colhidos ao longo de vários anos, caracterizando-se assim uma série numérica. Neste capítulo estão apresentados os principais conceitos utilizados nesse trabalho.

2.1 Série númerica

Segundo o dicionário Michaelis, série é uma sequência de acontecimentos que se sucedem ininterruptamente ou a pequenos intervalos, em nosso dia a dia encontramos diversos conjuntos de elementos que seguem um certa ordem, por exemplo, o Brasil é penta campeão mundial, esses títulos seguem uma sequência cronológica que é: 1958, 1962, 1970, 1994 e 2002. Para Ramos (2011), dentro da Matemática, o estudo de sequência numérica abrange o conjunto de números reais, dispostos em certa ordem .

Segundo Ercole (2010), "uma sequência numérica - ou simplesmente sequência - é uma lista infinita e ordenada de números reais $a_1, a_2, a_3, \ldots$".

Diante destas considerações ressalva-se que a Análise de Série Numérica é uma tarefa interessante, que vem sendo aplicada potencialmente em diversas áreas, onde elas geram dados de séries numéricas, tais como, Economia, Medicina, Epidemiologia, Meteorologia, dentre outros, o que acumulou uma extensa base de dados com grande volume de informação, tornando-se difícil a interpretação e processamento de toda essa informação por Seres Humanos. De acordo com Zalewski (2015) neste cenário se torna inviável o processo de análise manual, o que torna importante o desenvolvimento de processos, através de técnicas computacionais dentro destes já desenvolvidos, um deles é o aprendizado de máquina, com intuito de reduzir a intervenção humana, bem como a dependência por especialistas do domínio.

2.2 Aprendizado de máquina

Um **agente inteligente** é uma entidade autônoma que, percebe e age em um ambiente, um agente com aprendizagem pode melhorar o seu desempenho, por meio do estudo diligente de suas próprias experiências. Existem vários tipos de aprendizagem, mas para este trabalho utilizou-se de Aprendizagem **não supervisionada**.

Em Aprendizagem não supervisionada, não há rotulação ou *feedback* nos exemplos de entrada, ou seja, não há um professor ou especialista para correção, o agente aprende padrões na entrada. Comumente é utilizado para detecção de grupos de exemplos de entrada potencialmente úteis, denominado agrupamentos.

De acordo com Russel e Norvig (2004), em **Aprendizagem Supervisionada**, os exemplos são rotulados por um especialista, onde o agente observa exemplos de entrada e saída, a entrada seriam as percepções, ou seja, os dados propriamente ditos, e assim a saída é fornecida por um instrutor, por exemplo, fornecida uma série de conjuntos de vinhos, um especialista rotularia esses vinhos de acordo com seus tipos.

Monard e Baranauskas (2003) definiram que no **Aprendizado Supervisionado**, um algoritmo de aprendizado (**indutor**), tem por objetivo maior, extrair um bom classificador a partir de um conjunto de exemplos rotulados, também chamado de Conjunto de Treinamento. A partir da saída do indutor é possível predizer corretamente os rótulos de um novo conjunto de exemplos. Com isso o classificador pode ser avaliado por sua precisão, compreensibilidade, grau de interesse de aprendizado, requisitos de armazenamento, grau de compactação, etc. Um **exemplo**, na literatura também conhecido como caso, registro ou dado, é uma tupla ou vetor de valores de atributos, nele é descrito o objeto de interesse, como os valores de temperatura de uma cidade durante os doze meses que compõem um ano. Um **atributo** descreve alguma característica de um exemplo, há dois tipos de atributos: Nominal (sem ordem entre os valores) e Contínuo (existe uma ordem linear nos valores).

2.2.1 Escolha da métrica

Em Aprendizado de Máquina, existem vários métodos para ensinar uma máquina a reconhecer padrões de uma série numérica, que faz uso da escolha de uma métrica, para determinar o quão parecido são os objetos analisados.

Segundo Cesar (2016) "A escolha da métrica é um dos passos mais sensíveis ao desenvolver um algoritmo efetivo, dado que ela define como os elementos são similares/dissimilares entre si e irá impactar em como elementos são considerados "próximos" ou "distantes" ao obter agrupamentos e/ou classificar instâncias."

A elaboração do classificador neste trabalho, foi usado como base do *framework* *DAMICORE*, que utiliza a distância de compressão como métrica. A proposta será a modificação do fluxo de operação do *Framework*.

O fluxo normal do *DAMICORE* culmina em:

1. Obter matriz de distância entre as amostras de um conjunto de dados, com a métrica NCD com o compressor PPMd (*Prediction by partial matching*)
2. Criação de árvore binária através da matriz de distância, por meio do Neighbor Joining
3. Geração do agrupamento das amostras, com uso do Fast Newman

Com alteração do fluxo que concede:

1. Obter matriz de distância entre as amostras de um conjunto de dados, com a métrica NCD na compactação se adicionaria:
   a) Compressão com perdas, para generalizar os valores na granulação

b) Granulação, contagem de ocorrências de elementos da série de mesmo valor ou próximos
c) Codificação pela compressão sem perca do granulos

2. Criação de árvore binária através da matriz de distância, por meio do Neighbor Joining

2.2.1.1 Normalized Compression Distance

O NCD (*Normalized Compression Distance*) é um tipo de Métrica indicada a vários tipo de dados. Possui o diferencial de não depender que os dados estejam estruturados, mas apenas em sua codificação binária, assim esses bits passam para um compressor, que obtém uma medida do conteúdo da informação. No NCD, a ideia de similaridade de objetos vem do conceito de Entropia, que reflete o tanto de informação que é preciso para predizer a resposta, como exemplo X e Y. Com o X tendo a entropia expressiva em relação a Y, torna-se melhor descrever Y apenas referenciando Y. As relações são descritas pela origem de entropia a partir de um compressor. (TORRES, 2018)

O cálculo de distância de compressão de objeto é baseada na complexidade de *Kolmogorov* $K(x)$, onde o $K(x)$ de um objeto x, é o menor comprimento da descrição deste objeto x em uma linguagem universal, isto é, $K(x) \leq |x|$.

É computacionalmente difícil realizar o cálculo da complexidade de *Kolmogorov*, o NCD realiza uma aproximação deste cálculo em tempo de computação viável, com a seguinte equação:

$$NCD(x,y) = \frac{C(xy) - min(Cx), C(y))}{max(C(x), C(y))}$$

Onde C(x) é dado como a distância de compressão do objeto x, e xy é um objeto criado a partir da concatenação de x com y.

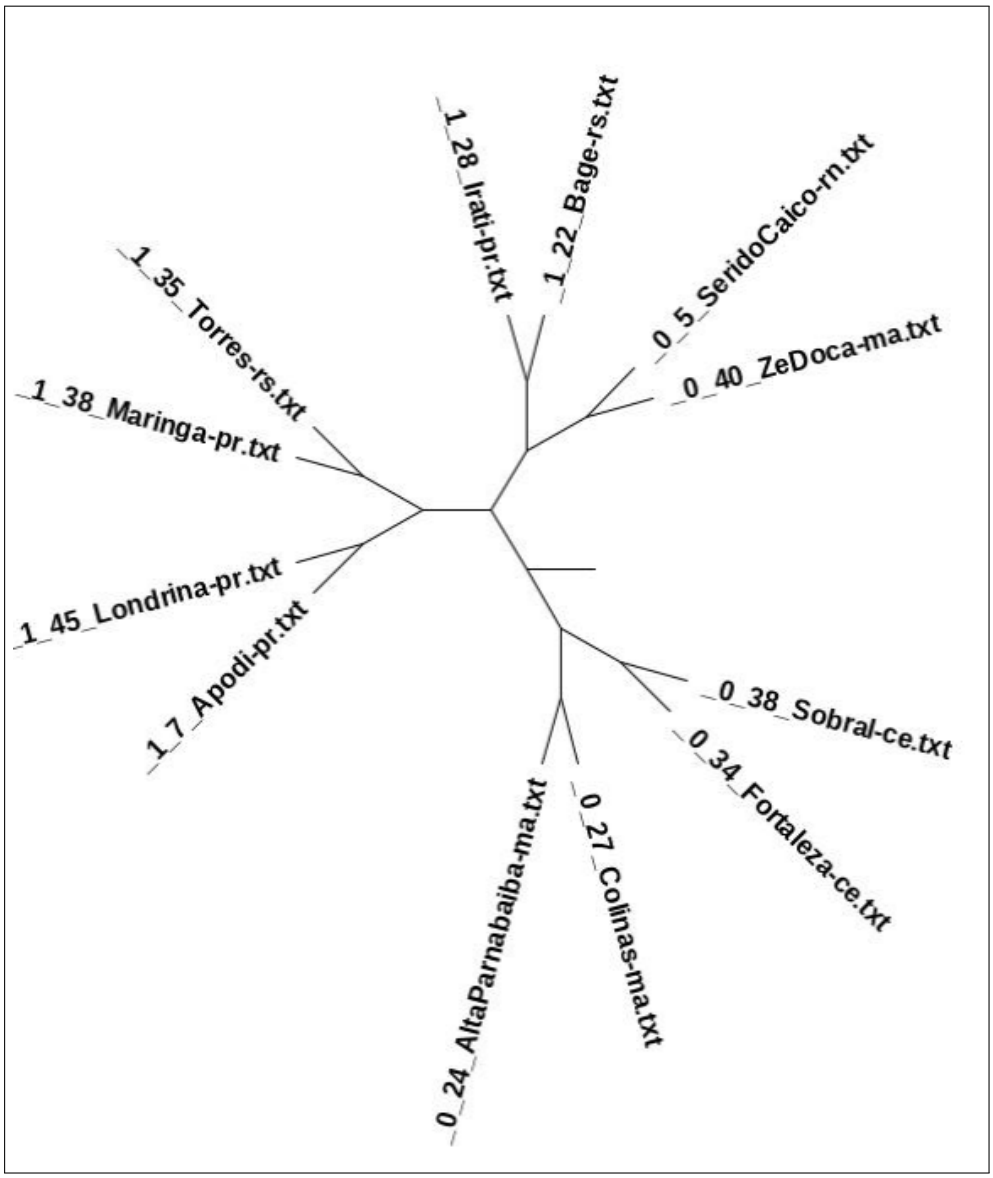
2.2.1.2 Neighbor Joining

Nas palavras de Cesar (2016) o método de *Neighbor Joining* (NJ), Junção de vizinhos em sua tradução para português, foi consolidada na biologia para uma eficiente identificação de filogenias entre espécies. É utilizada uma matriz de distância que, pode ser vista como um grafo completo, sendo os objetos como nós e as distâncias representando as arestas. É realizada uma simplificação, para resumir a informação completa e adquirir uma rede complexa, assim destacando a informação potencial nas relações. Dentro os métodos utilizados, os mais comuns são:

- **Limiarização:** Remove todas as arestas cuja distância está acima de um limiar θ, resultando-se assim em um grafo desconexo.
- **k-NN (*k-nearest neighbors algorithm*):** Para cada nó, se mantém as arestas para seus k vizinhos mais próximos, resultando-se em um grafo conexo.

Como saída do algoritmo NJ, têm-se uma árvore filogenética, que possui a evolução mínima dos objetos, na Figura 1 é exibido um exemplo de árvore filogenética.

Figura 1 – Exemplo de árvore filogenética



Fonte: O autor.

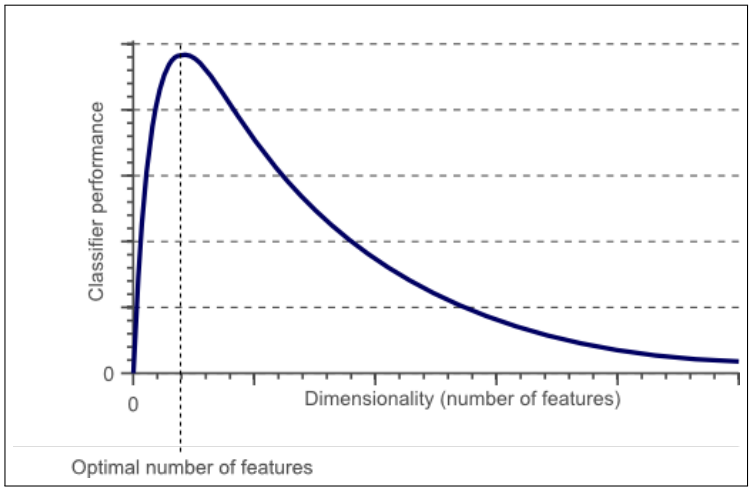
2.3 Maldição da dimensionalidade

Vários Algoritmos de classificação obtém um bom desempenho com objetos de poucas dimensões (atributos), mas com o aumento da quantidade de dimensões os objetos se tornam mais difíceis de serem diferenciados entre em si, tornando mais difícil etapas como a de classificação, o que leva a um problema chamado de maldição da dimensionalidade. Por isso Silva (2002) recorda que com o aumento na dimensionalidade dos dados, resulta

inicialmente em um incremento na acurácia do resultado do agente. Este fato seria de se esperar, pois com o aumento do número de bandas, ocorre um incremento na informação disponível. A partir de um certo ponto, entretanto, e utilizando-se as mesmas amostras de treinamento (para fins de estimação dos parâmetros do classificador), a acurácia começa a decrescer, devido o aumento da dimensionalidade dos dados. Este efeito conhecido como o fenômeno de *Hughes* ou a maldição da dimensionalidade (*the curse of dimensionality*), deve-se ao aumento do número de parâmetros a serem estimados, especialmente na matriz covariância.

Na Figura 2 é exibido um gráfico exibindo a performance do classificador, ao longo da adição de mais objetos, exemplifica bem como a maldição de dimensionalidade influencia o classificador.

Figura 2 – Performance vs dimensionalidade



Fonte: (RAJ, 2019).

Para contornar da maldição da dimensionalidade, usa-se a Compressão de dados. A Compressão de dados fornece métodos, conceitos e procedimentos para reduzir o número de bits a serem utilizados para armazenar ou transmitir informações. Para esta finalidade, utiliza-se de uma grande variedade de técnicas tanto de Software quanto de Hardware. Estas técnicas podem adotar os seguintes critérios: Compressão Lógica que, por meio de raciocínio lógico, substitui a informação por uma informação equivalente e Compressão Física que age diretamente sobre os dados (FILHO, 1994).

A compressão lógica explora o significado relativo existente entre os dados que necessitam ser comprimidos, ou seja, substitui a informação por uma informação equivalente. A compressão física explora a probabilidade da ocorrências de caracteres e de grupos de caracteres diferem uns dos outros, reduz os dados codificando os caracteres de maior ocorrência, por esses ocuparem maior número de bits.

Neste trabalho foi escolhido a compressão com perdas, para melhorar a eficácia e taxa de compressão, mas conservando-se uma compressão de dados mais próximo possível dos dados originais.

A Maldição da Dimensionalidade é um fator desafiador na modelagem Matemática visto que, para um hiperplano cartesiano com d dimensões de entrada onde cada dimensão de entrada é particionada em s células, o número total de células seria de s^d. Como consequência disso, a criação de modelos destes dados necessita considerar espaços de busca, inerentemente esparsos. Desta forma, os cientistas constantemente têm se deparado com a necessidade de encontrar estruturas significativas ocultas, de baixa dimensão, dentro de dados de alta dimensão, sendo tal técnica denominada de redução de dimensionalidade dos dados (RDD). Analogamente, o cérebro humano se confronta com o mesmo problema em suas percepções diárias, extraindo de forma eficiente, um pequeno número de estímulos relevantes a partir de aproximadamente 30.000 fibras nervosas sensoriais. Dada à capacidade limitada do cérebro humano de lidar com a complexidade, esta RDD consiste em um fator chave para permitir a generalização de conceitos, transformando as experiências diárias, em conhecimento e ideias. Adicionalmente, a quantidade de exemplos necessários para adaptar um modelo multivariado cresce exponencialmente, em relação à quantidade de características que representam cada amostra (CAMARGO, 2010) .

3 Metodologia

Nesse capítulo descreveu-se os métodos utilizados na coleta de dados, tais informações foram utilizados para a entrada do modelo de inteligência que foi desenvolvido, e assim para a automatização da análise de série numérica, diminuiu-se a intervenção humana.

3.1 Materiais

A base de dados utilizada neste trabalho, foi construída através da coleta de dados em várias base de dados já existentes, foram coletadas 9861 amostras em formato de série numérica, de diversas áreas. O primeiro banco de dados utilizados foi o BDMEP (Banco de Dados Meteorológicos para Ensino e Pesquisa) que segundo INMET (2019) é uma espécie de banco de dados, cuja finalidade está em apoiar as atividades de ensino, pesquisa e outras aplicações em Meteorologia, Hidrologia, recursos Hídricos, Saúde Pública, Meio Ambiente, dentre outros. Os dados são provenientes de estações meteorológicas convencionais da rede de estações do INMET (Instituto Nacional de Meteorologia) com milhões de informações, referentes as medições diárias. Todos os dados desta base de dados estão no formato de série temporal.

O segundo banco de dados utilizado, foi o UCI (*University of California, Irvine*) *Machine Learning Repository*, esta base de dados é altamente indicado para utilização em algoritmos de Inteligência Artificial contendo vários tipos de dados pertencentes a variadas áreas. De acordo com Dua e Graff (2017) ”uma coleção de bancos de dados, teorias de domínios e geradores de dados que são usados pela comunidade de aprendizado de máquina, para análise empírica de algoritmos de aprendizado de máquina”.

Algumas destas amostras vieram com ‘lixo’, dentro de algumas das bases de dados escolhidas, ocorreu-se de certas amostras terem valores ausentes, sendo denotados na amostra por ’?’, então foi necessário fazer a ‘limpeza’ destes dados, onde estes dados com esse ’lixo’ foram ignorados, também foi ignorados atributos nominais de algumas amostras visto que se visa trabalhar somente com dados numéricos. Algumas destas amostras de dados foram armazenados em planilhas eletrônicas, em seu banco de dados original e em outras bases de dados, todas as amostras de dados foram armazenadas em um único arquivo de texto, sendo necessário separá-los, e assim cada amostra foi armazenada em um arquivo de texto, arquivos que foram entrada para o classificador, e por conseguinte, estes testes serão aqui descrito como experimento. Inicialmente foi focado experimentos com séries numéricas, mas para extensão desta pesquisa, foram adicionados experimentos utilizando dados no formato de vetores numéricos.

3.1.1 Experimento 1 - Análise de dados climáticos

Dados retirados do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP). Os dados climáticos possuem as seguintes características: precipitação ocorrida nas últimas vinte e quatro horas; temperatura compensada; temperatura máxima; temperatura mínima; umidade relativa do ar; pressão atmosférica média; insolação; direção e velocidade do vento máxima e média; evaporação do piche; evapotranspiração potencial e real BH; nebulosidade média. Os dados em questão são referentes ao ano de 2018, sendo cada série temporal referindo a uma cidade e tendo 12 elementos com valor decimal e positivo, que se referem a cada mês deste ano, sendo assim uma série temporal. O experimento terá o objetivo de predizer as cidades que possuem o clima similar, portanto cidades com clima quente, seriam agrupadas com outras cidades de clima quente. Um exemplo de amostra deste experimento, é a série temporal de precipitação total da cidade de Vitória-ES: 55.1, 201.5, 174.6, 371.6, 200.8, 148.3, 9.4, 92.1, 39.7, 179.8, 144.9, 169.4. Os dados coletados para esse experimento, encontram-se descritos na Tabela 1 a seguir.

Tabela 1 – Experimento 1 - Dados das amostras climáticas.

Variáveis Climáticas	Total de amostras
Direção do vento predominante	119
Evapotranspiração real BH	122
Evapotranspiração potencial BH	122
Evaporação do piche	53
Insolação total	104
Nebulosidade média	140
Precipitação total	173
Pressão atmosférica média	92
Temperatura compensada média	106
Temperatura máxima média	118
Temperatura mínima média	133
Umidade relativa média	116
Velocidade do vento máxima	117
Velocidade do vento média	116
Total	1631

Fonte: O autor.

3.1.2 Experimento 2 - PIB nominal

Dados retirados do Fundo Monetário Internacional (FMI). Neste experimento, coletou-se os dados referentes ao período compreendido entre 2000 a 2009 com valores

de PIB nominal de 181 países, a série temporal contém 10 elementos de valor inteiro e positivo, onde cada série representa um país e cada elemento referente a um ano, como no exemplo de amostra do país Cazaquistão: 18292, 22153, 24637, 30822, 43152, 57125, 81003, 103142, 135229, 115308. Desta forma, o objetivo do experimento será predizer os países que possuem o valor de PIB similares, divididos em duas classes: **Alto** e **Baixo**.

3.1.3 Experimento 3 - Vinhos (*Wine Data Set*)

Dados retirados do *UCI Machine Learning Repository*, neste experimento foi utilizado vetores numéricos. Para o experimento 3, foram coletados dados resultantes de uma análise química de vinhos cultivados na mesma região da Itália, estes vinhos foram separados em 3 tipos, onde cada vinho possui 13 atributos de valor decimal e positivo, como é exibido na amostra de um dos vinhos: 14.23, 1.71, 2.43, 15.6, 127, 2.8, 3.06, 0.28, 2.29, 5.64, 1.04, 3.92, 1065. O objetivo será agrupar os vinhos de acordo com seu tipo: **Tipo 1**, **Tipo 2** ou **Tipo 3**, ou seja, vinhos do **Tipo 1** agrupados com vinho de **Tipo 1**. Os atributos do vinho são:

1. Álcool
2. Ácido málico
3. Cinzas
4. Alcalinidade das cinzas
5. Magnésio
6. Fenóis totais
7. Flavonoides
8. Fenóis não flavonoides
9. Proantocianinas
10. Intensidade da cor
11. Matiz
12. OD280/OD315 de vinhos diluídos
13. Prolina

Figura 3 – Informações originais da base de dados vinhos

Data Set Characteristics:	Multivariate	**Number of Instances:**	178	**Area:**	Physical
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	13	**Date Donated**	1991-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	1191460

Fonte: (DUA; GRAFF, 2017).

Na Figura 3 é mostrado as informações sobre a base dados original. Após a

'limpeza' dos dados foi obtido 178 amostras no total, sendo que a classe vinho **Tipo 1** possui 59 amostras, a classe vinho **Tipo 2** possui 71 amostras e a classe vinho **Tipo 3** possui 48 amostras.

3.1.4 Experimento 4 - Concha orgânica (*Abalone Data Set*)

Dados retirados do *UCI Machine Learning Repository*, neste experimento foi utilizado vetores numéricos. Neste experimento, foram utilizados dados de medições físicas de um abalone (concha orgânica), o objetivo culminou em predizer se o abalone é **Masculino**, **Feminino** ou **Infantil**, cada concha possui 9 atributos de valor decimal e inteiro, como no exemplo de uma dessas conchas: M, 0.455, 0.365, 0.095, 0.514, 0.2245, 0.101, 0.15, 15. Esse conjunto de dados possui os seguintes atributos:

1. Sexo
2. Comprimento
3. Diâmetro
4. Altura
5. Peso total
6. Peso com casca
7. Peso das vísceras
8. Peso da concha
9. Anéis

O atributo Sexo foi ignorado por ser nominal (Masculino ou feminino), pois trabalha-se assim apenas conjuntos de dados numéricos, onde tal atributo pode ser utilizado como rótulo para identificação da classe em um classificador (Aprendizado de Máquina Supervisionado).

Figura 4 – Informações originais da base de dados *abalone* (concha)

Data Set Characteristics:	Multivariate	**Number of Instances:**	4177	**Area:**	Life
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	8	**Date Donated**	1995-12-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	802120

Fonte: (DUA; GRAFF, 2017).

Na Figura 4 exibe as informações sobre a base dados original. Após a 'limpeza' dos dados foi obtido 4177 amostras no total, sendo que a classe concha **Masculino** possui 1528 amostras, a classe concha **Feminino** possui 1307 amostras e a classe concha **Infantil** possui 1342 amostras.

3.1.5 Experimento 5 - Iris (*Iris Data Set*)

Dados retirados do *UCI Machine Learning Repository*, neste experimento foi utilizado vetores numéricos. Este conjunto de dados é um dos mais conhecidos na literatura de reconhecimento de padrões. Dados que foram coletados para quantização morfológica de três espécies de flor de iris, foram coletadas da Península de Gaspé, segundo Anderson (1936) "todas do mesmo pasto e colhidos no mesmo dia e medidos ao mesmo tempo pela mesma pessoa com o mesmo aparelho". O conjunto contém 3 classes com 50 amostras de cada classe, cada flor de iris possui 4 atributos de valor decimal e positivo, assim como no exemplo de uma das amostras de flor: 5.1, 3.5, 1.4, 0.2. O objetivo é agrupar cada flor em uma das classes: **Versicolor**, **Virginica** e **Setosa**. Este conjunto de dados possui os seguintes atributos:

1. comprimento da sépala em cm
2. largura da sépala em cm
3. comprimento da pétala em cm
4. largura da pétala em cm

Figura 5 – Informações originais da base de dados iris

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	2719958

Fonte: (DUA; GRAFF, 2017).

Na Figura 5 mostra as informações sobre a base dados original. Após a 'limpeza' dos dados foi obtido 150 amostras no total, sendo que a classe iris **Setosa** possui 50 amostras, a classe iris **Versicolor** possui 50 amostras e a classe iris **Virginica** possui 50 amostras.

3.1.6 Experimento 6 - Câncer de mama (*Breast Cancer Wisconsin (Original) Data Set*)

Dados retirados do *UCI Machine Learning Repository*, neste experimento foi utilizado vetores numéricos. O experimento tem como objetivo agrupar os câncer em dois grupos: **Maligno** e **Benigno**, embora essa categorização não seja mais utilizada na medicina, por motivo de estudo foi preservado a classificação antiga. De acordo com Dua e Graff (2017) "As amostras chegavam periodicamente quando o Dr. Wolberg relata seus casos clínicos. O banco de dados, portanto, reflete esse agrupamento cronológico dos dados.". Cada câncer possui 11 atributos de valor inteiro e positivo, como no exemplo de

uma das amostras: 1000025, 5, 1, 1, 1, 2, 1, 3, 1, 1, 2. Este conjunto de dados possui os seguintes atributos:

1. Número do código de amostra: número de identificação
2. Espessura do grupo: 1 - 10
3. Uniformidade do tamanho da célula: 1 - 10
4. Uniformidade da forma da célula: 1 - 10
5. Adesão marginal: 1 - 10
6. Tamanho de célula epitelial única: 1 - 10
7. Núcleos desencapados: 1 - 10
8. Cromatina Branda: 1 - 10
9. Núcleos normais: 1 - 10
10. Mitoses: 1 - 10
11. Classe: (2 para benignos, 4 para malignos)

O atributo 'número do código de amostra' foi ignorado, pois possui um valor muito elevado, enquanto os outros atributos possui um valor muito baixo, que pode levar a uma perda na qualidade dos resultados. Na Figura 6 exibe as informações sobre a base dados original.

Após a 'limpeza' dos dados foi obtido 683 amostras no total, sendo que a classe de câncer **Maligno** possui 239 amostras e a classe câncer **Benigno** possui 444 amostras.

Figura 6 – Informações originais da base de dados câncer de mama

Data Set Characteristics:	Multivariate	**Number of Instances:**	699	**Area:**	Life
Attribute Characteristics:	Integer	**Number of Attributes:**	10	**Date Donated**	1992-07-15
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	503801

Fonte: (DUA; GRAFF, 2017).

3.1.7 Experimento 7 - Prognóstico câncer de mama (*Breast Cancer Wisconsin (Prognostic)*)

Dados retirados do *UCI Machine Learning Repository*, neste experimento foi utilizado vetores numéricos. O experimento tem como objetivo agrupar os câncer em dois grupos: **Recorrente** e **Não recorrente**. Segundo Dua e Graff (2017) "Estes são pacientes consecutivos atendidos pelo Dr. Wolberg desde 1984 e incluem apenas os casos que exibem câncer de mama invasivo e nenhuma evidência de metástases distantes no momento do diagnóstico.". Uma amostra de câncer possui 34 atributos de valor decimal e positivo, como nessa amostra: 119513, N, 31, 18.02, 27.6, 117.5, 1013, 0.09489, 0.1036, 0.1086, 0.07055,

0.1865, 0.06333, 0.6249, 1.89, 3.972, 71.55, 0.004433, 0.01421, 0.03233, 0.009854, 0.01694, 0.003495, 21.63, 37.08, 139.7, 1436, 0.1195, 0.1926, 0.314, 0.117, 0.2677, 0.08113, 5, 5. Este conjunto de dados possui os seguintes atributos:

1. Número de identificação
2. Resultado (R = recorrente, N = não recorrente)
3. Tempo (tempo de recorrência se campo 2 = R, tempo livre de doença se campo 2 = N)
4. Dez recursos com valor real são calculados para cada núcleo celular:
   - A raio (média das distâncias do centro aos pontos do perímetro)
   - B textura (desvio padrão dos valores da escala de cinza)
   - C perímetro
   - D área
   - E suavidade (variação local no comprimento do raio)
   - F compactação (perímetro^2 / área - 1,0)
   - G concavidade (severidade das porções côncavas do contorno)
   - H pontos côncavos (número de partes côncavas do contorno)
   - I simetria
   - J dimensão fractal (aproximação da costa - 1)

O atributo 'número do código de amostra' foi ignorado, pois possui um valor muito elevado enquanto os outros atributos possui um valor muito baixo, que pode levar um perda na qualidade dos resultados, assim como na experiência anterior, o atributo 'resultado' também foi ignorado por ser nominal.

Figura 7 – Informações originais da base de dados prognóstico câncer de mama

Data Set Characteristics:	Multivariate	**Number of Instances:**	198	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	34	**Date Donated**	1995-12-01
Associated Tasks:	Classification, Regression	**Missing Values?**	Yes	**Number of Web Hits:**	190489

Fonte: (DUA; GRAFF, 2017).

Na Figura 7 é exibido as informações sobre a base dados original. Após a 'limpeza' dos dados foi obtido 194 amostras no total, sendo que a classe de câncer **Recorrente** possui 46 amostras e a classe câncer **Não recorrente** possui 148 amostras.

3.1.8 Experimento 8 - Diagnóstico câncer de mama (*Breast Cancer Wisconsin (Diagnostic)*)

Dados retirados do *UCI Machine Learning Repository*, neste experimento foi utilizado vetores numéricos. O experimento tem como objetivo agrupar os câncer em dois

grupos: **Maligno** e **Benigno**, termos de categorização não são utilizados atualmente na medicina, mas por motivo de estudo foi preservado os termos antigos. Segundo Dua e Graff (2017) "Os recursos são calculados a partir de uma imagem digitalizada de um aspirado por agulha fina (PAAF) de uma massa mamária. Eles descrevem características dos núcleos celulares presentes na imagem.". Cada câncer possui 31 atributos, como mostrado nessa amostra: 842302, M, 17.99, 10.38, 122.8, 1001, 0.1184, 0.2776, 0.3001, 0.1471, 0.2419, 0.07871, 1.095, 0.9053, 8.589, 153.4, 0.006399, 0.04904, 0.05373, 0.01587, 0.03003, 0.006193, 25.38, 17.33, 184.6, 2019, 0.1622, 0.6656, 0.7119, 0.2654, 0.4601, 0.1189. Este conjunto de dados possui os seguintes atributos:

1. Número de identificação
2. Diagnóstico (M = maligno, B = benigno)
3. Tempo (tempo de recorrência se campo 2 = R, tempo livre de doença se campo 2 = N)
4. Dez recursos com valor real são calculados para cada núcleo celular:
   - A raio (média das distâncias do centro aos pontos do perímetro)
   - B textura (desvio padrão dos valores da escala de cinza)
   - C perímetro
   - D área
   - E suavidade (variação local no comprimento do raio)
   - F compactação (perímetro^2 / área - 1,0)
   - G concavidade (severidade das porções côncavas do contorno)
   - H pontos côncavos (número de partes côncavas do contorno)
   - I simetria
   - J dimensão fractal (aproximação da costa - 1)

O atributo 'diagnóstico' foi ignorado por ser nominal, pois trabalha-se assim apenas conjuntos de dados numéricos, esse atributo pode ser utilizado como rótulo para identificação da classe em um classificador, assim como nos 2 experimentos anteriores o atributo 'número do código de amostra' por possuir um valor muito grande.

Figura 8 – Informações originais da base de dados diagnóstico câncer de mama

Data Set Characteristics:	Multivariate	**Number of Instances:**	569	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	32	**Date Donated**	1995-11-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	979523

Fonte: (DUA; GRAFF, 2017).

Na Figura 8, consta as informações sobre a base dados original. Após a 'limpeza' dos dados foi obtido 569 amostras no total, sendo que a classe de câncer **Maligno** possui

212 amostras e a classe câncer **Benigno** possui 357 amostras.

3.1.9 Experimento 9 - Escala de balança (Balance Scale Data Set)

Dados retirados do *UCI Machine Learning Repository*, neste experimento foi utilizado vetores numéricos. O experimento tem como objetivo, agrupar as escalas de balança em três grupos: **B**, **L** e **R**. De acordo com Dua e Graff (2017) "Esse conjunto de dados foi gerado para modelar resultados experimentais psicológicos. Cada exemplo é classificado como tendo a ponta da balança à direita, a ponta à esquerda ou equilibrada.". Cada amostra representa uma balança com 5 atributos, como nesse exemplo: B, 1, 1, 1, 1. Este conjunto de dados possui os seguintes atributos:

1. Nome da classe: (L, B, R)
2. Peso esquerdo: (1, 2, 3, 4, 5)
3. Distância esquerda: (1, 2, 3, 4, 5)
4. Peso certo: (1, 2, 3, 4, 5)
5. Distância certa: (1, 2, 3, 4, 5)

O atributo nome da classe foi ignorado por ser nominal, pois será trabalhado somente conjuntos de dados numéricos, esse atributo pode ser utilizado como rótulo para identificação da classe em um classificador.

Figura 9 – Informações originais da base de dados escala de balança

Data Set Characteristics:	Multivariate	**Number of Instances:**	625	**Area:**	Social
Attribute Characteristics:	Categorical	**Number of Attributes:**	4	**Date Donated**	1994-04-22
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	223625

Fonte: (DUA; GRAFF, 2017).

Na Figura 9 mostra as informações sobre a base dados original. Após a 'limpeza' dos dados foi obtido 625 amostras no total, sendo que a classe balança **B** possui 49 amostras, a classe balança **L** possui 288 amostras e a classe balança **R** possui 50 amostras.

3.1.10 Experimento 10 - Escolha de método contraceptivo (*Contraceptive Method Choice Data Set*)

Dados retirados do *UCI Machine Learning Repository*, neste experimento foi utilizado vetores numéricos. O experimento tem como objetivo agrupar as escolhas de método contraceptivo em três grupos: **Sem uso**, **Longo prazo**, **Curto prazo**. Segundo Dua e Graff (2017) "O conjunto de dados é um subconjunto da Pesquisa Nacional de Prevalência Contraceptiva da Indonésia em 1987. As amostras são mulheres casadas que

não estavam grávidas, ou não sabem se estavam no momento da entrevista.". Uma amostra representa uma mulher com 10 atributos, como é exibido neste exemplo: 24, 2, 3, 3, 1, 1, 2, 3, 0, 1. Este conjunto de dados possui os seguintes atributos:

1. Idade da esposa (numérica)
2. Educação da esposa (categórica) 1 = baixa, 2, 3, 4 = alta
3. Educação do marido (categórica) 1 = baixa, 2, 3, 4 = alta
4. Número de filhos já nascidos (numéricos)
5. Religião da esposa (binária) 0 = Não-Islã, 1 = Islã
6. A esposa está trabalhando agora? (binário) 0 = Sim, 1 = Não
7. Ocupação do marido (categórica) 1, 2, 3, 4
8. Índice de padrão de vida (categórico) 1 = baixo, 2, 3, 4 = alto
9. Exposição na mídia (binária) 0 = Bom, 1 = Não Bom
10. Método contraceptivo usado (atributo de classe) 1 = Sem uso, 2 = Longo prazo, 3 = Curto prazo

Tabela 2 – Experimento 10 - Dados das amostras escolha de método contraceptivo.

Classe	Total de amostras
Sem uso	629
Longo prazo	333
Curto prazo	511
Total	1473

Fonte: O autor.

Figura 10 – Informações originais da base de dados escolha de método contraceptivo

Data Set Characteristics:	Multivariate	**Number of Instances:**	1473	**Area:**	Life
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	9	**Date Donated**	1997-07-07
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	167565

Fonte: (DUA; GRAFF, 2017).

O atributo método contraceptivo usado foi ignorado, pois foi utilizado aprendizado de máquina não supervisionado, tal atributo pode ser utilizado como rótulo de identificação da classe, em um classificador supervisionado. Na Tabela 2 está exposto a quantidade de amostras em cada classe, após a 'limpeza' e na Figura 10 exibe as informações sobre a base dados original.

3.2 Métodos

A área de estudo dessa pesquisa é a análise de séries numéricas, para o início do desenvolvimento de qualquer análise primariamente estão sendo coletados diversos dados que estejam em formatos de séries numéricas. Com essa coleta de dados realizada, será feita o modelo de representação do conjunto de séries numéricas, com esse modelo se iniciará a concepção de um modelo de classificação para os conjuntos de séries numéricas já coletados. A criação de um modelo de conhecimento através de aprendizado não supervisionado, pode servir como auxílio para um classificador, na tarefa de predizer se uma série numérica como pertencente a um rótulo (se duas séries possuem semelhanças entre elas).

3.2.1 Modelo de inteligência

A concepção deste classificador utilizará um modelo de inteligência criado por aprendizado de máquina não supervisionado, irá se basear na metodologia *DAMICORE*, será realizado mudança no fluxo desta metodologia, na fase anterior ao cálculo da matriz de distâncias com o NCD, será introduzida uma fase chamada granulação com o papel de identificar e contar os elementos de uma série numérica, realizar-se-á inicialmente uma compressão com perdas.

Após a fase de cálculo da matriz de distâncias, será gerado uma árvore filogenética com o algoritmo *Neighbor Joining*. O classificador não supervisionado foi desenvolvido na linguagem de programação Python na versão 3.

3.2.1.1 Inicialização do classificador e compressão

O classificador se alimenta com fontes de dados do diretório contendo os arquivos com as séries numéricas. Com estes dados, o algoritmo inicia a aplicação de algum tipo de técnica de compressão de dados, para criar uma generalização maior para a fase de granulação.

Para compressão da série, desenvolve-se um modelo Matemático, para auxiliar na implementação do método de compressão para séries numéricas, visto que durante as pesquisas, não encontrou-se o modelo para isso. No algoritmo, desenvolveu-se o código na *Python*,na versão 3.

3.2.1.2 Compressão da série numérica

Para a compressão da série numérica foi desenvolvido o algoritmo com os seguintes passos:

1. Abstração do ponto da série numérica
2. Marcar ponto inicial e o final da série
3. Calcular a inclinação entre os dois pontos

4. Calcular a distância entre cada ponto da série original a uma reta que inicia no ponto inicial (reta de suporte)
5. Selecionar o ponto com maior distância a esta reta de suporte e marcar como usado, então está sendo adicionado a série comprimida
6. Os passos anteriores são repetidos no trecho entre o ponto inicial (referência a esquerda) e o ponto final (referência a direita)
7. Esses passos são efetuados até que o número de pontos para a série comprimida seja alcançado

Para abstrair um ponto da série numérica foi definido o seguinte o objeto:

Seja o objeto *Ponto* (Ponto) que possui as seguintes propriedades:

- v: **valor** - valor contido em um ponto na série
- d: **distância** - distância entre o número e a reta teórica que liga os *pivots* da esquerda e da direita
- u: **usado** - se o número já foi inserido na série
- i: **índice** - índice que o valor ocupava na série original

O objeto *Ponto* é iniciado por $Ponto(v, i)$, é marcado como não usado, a distância será atualizada no decorrer do algoritmo. A função que inicializa o objeto é detalhada no Apêndice A.

O Apêndice B tem-se o algoritmo que calcula a inclinação entre dois pontos no plano cartesiano, ou seja, calcula a tangente.

No Apêndice C é exibido o algoritmo que calcula a distância entre o valor e a reta suporte traçada entre dois pontos, que são os *pivots* esquerdo e direito, a lista *serieOriginalPontos* é global em todo algoritmo.

O Apêndice D é detalhado o algoritmo que seleciona o ponto mais distante da reta de suporte. Como as distâncias são atualizadas apenas em um trecho da série, a busca é global, ou seja, em toda série.

Quando um *pivot* é escolhido, é necessário atualizar a distância no trecho entre os dois pontos de referência anteriores. No Apêndice E é detalhado o algoritmo que retorna o ponto a esquerda e no Apêndice F é detalhado o algoritmo que retorna o ponto a direita.

O Apêndice G exibe o algoritmo que extrai os pontos marcados para duas listas independentes contendo as informações da série comprimida, uma lista conterá os índices da série original e a outra os valores de da série.

O Apêndice H detalha o algoritmo que marca o início e o final da série, após calcula a inclinação entre os dois pontos e calcula a distância entre cada ponto da série original a uma reta que inicia no ponto inicial (reta de suporte). Em seguida o algoritmo seleciona o ponto com maior distância a esta reta de suporte e marca-o como usado (será adicionado a série comprimida). Os passos anteriores são repetidos entre o ponto inicial e final, será recalculado a inclinação, distância e etc, até que o número de pontos para série comprimida seja alcançado.

No n-ésimo ponto o *pivot* pode estar no meio da série e haver outros pontos que já foram *pivots* e que, portanto, fazem parte da série comprimida. Esses pontos são as referências à direita e à esquerda para calcular a inclinação, distância e etc e só atualizada em único trecho de código.

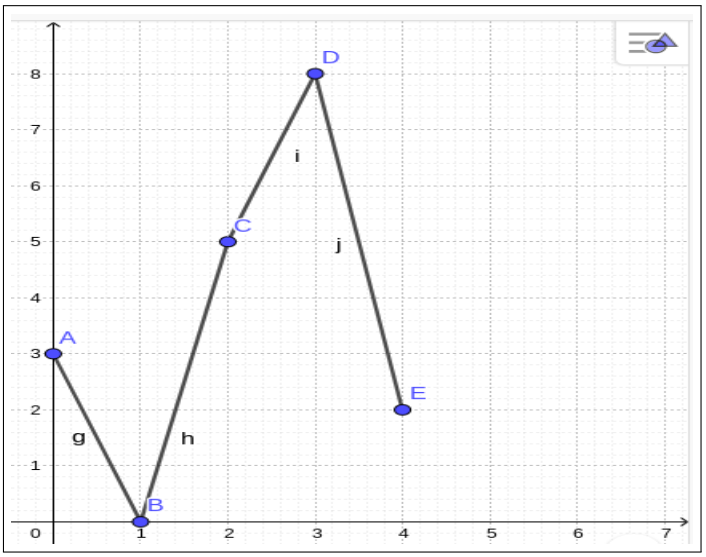
No Apêndice I é detalhado o algoritmo que preenche a lista global *serieOriginal-Pontos* com cada elemento da série original abstraído para um objeto Ponto e chama a função comprimir que retorna a série comprimida.

Para exemplificar o algoritmo foi utilizado o *software* geogebra para elaboração de gráficos, para melhor visualização de alguns passos do algoritmo. Foi utilizado como entrada a série: $3, 0, 5, 8, 2$ e fator de redução (FR) 3. Na Figura 11 é exibido o gráfico gerado pela série original.

Em seguida é realizado o cálculo de inclinação entre dois pontos no plano que no caso do fator de redução 3 é realizado somente uma vez, na Figura 12 é exibido o gráfico com o calculo de inclinação e a reta de suporte f tracejada para cálculo da distância.

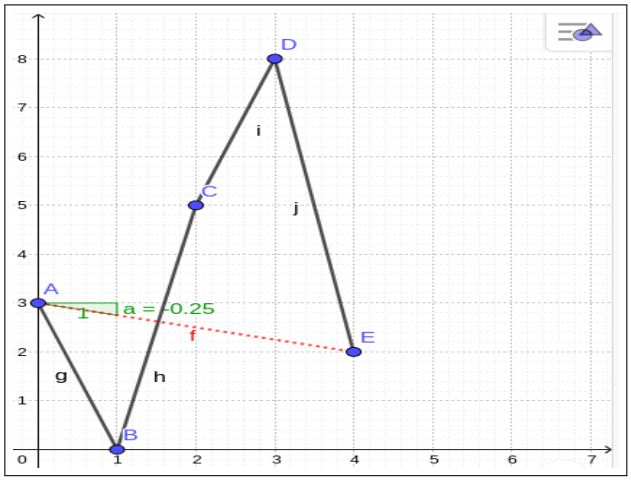
Após o cálculo da distância da primeira iteração do algoritmo são resultados: a série de pontos = [(d: 0.0, i: 0, v: 1, v: 3), (d: 2.75, i: 1, u: 0, v: 0), (d: 2,5, i: 2, u: 0, v: 5), (d: 5,75, i: 3, u:0, v: 8), (d: 0, i: 4, u:1, v:2)], foi selecionado **pivot** = 3, a referência a direita é 3, referência a esquerda é 0, como o primeiro elemento e o último são marcados e o elemento 3 será marcado também, como o fator de redução é 3, o algoritmo encerra e temos a seguinte série exibida na Figura 13.

Figura 11 – Gráfico gerado pela série original



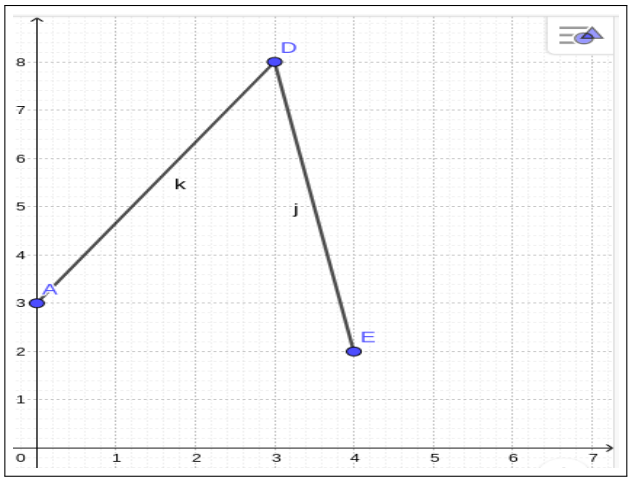
Fonte: O autor.

Figura 12 – Cálculo da inclinação e reta suporte



Fonte: O autor.

Figura 13 – Gráfico gerado pela série comprimida com FR = 3

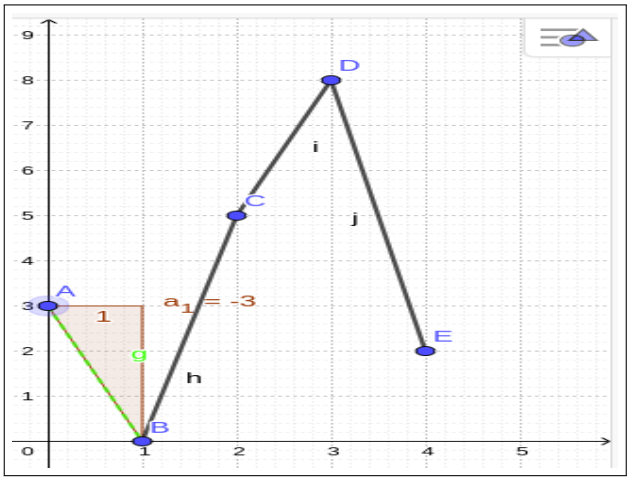


Fonte: O autor.

Agora com essa mesmo parâmetros utilizados anteriormente mas mudando o fator de redução para 4, na segunda iteração o novo **pivot** selecionado é 1, a referência da direita é 3, a referência da esquerda é 3 e a serie de pontos = [(d: 0.0, i: 0, v: 1, v: 3), (d: 0.0, i: 1, u: 1, v: 0), (d: 1.0, i: 2, u: 0, v: 5), (d: 5,75, i: 3, u:0, v: 0), (d: 0, i: 4, u:1, v:2)], como o calculo da distância teve que ser refeito foi traçado outra reta de suporte g, conforme exibido na Figura 14.

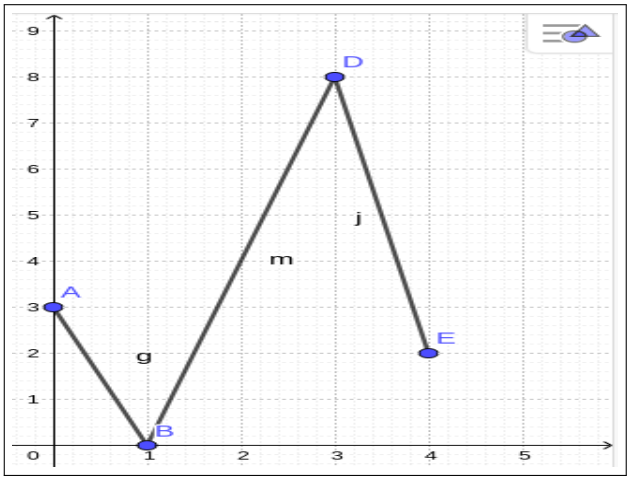
O elemento 1 é adicionada na série comprimida e atingimos a quantidade de elementos da série comprimida e então o algoritmo é encerrado é temos a série comprimida como exibido na Figura 15.

Figura 14 – Recálculo da inclinação e nova reta suporte



Fonte: O autor.

Figura 15 – Gráfico gerado pelo série comprimida com FR = 4



Fonte: O autor.

3.2.1.3 Granulação e codificação

Com a compressão da série executada, será iniciada a fase de granulação, objetivando-se contar a ocorrência de certos valores da série, com a quantização de valores feita pela compressão poderá gerar uma capacidade de generalização maior. O algoritmo de *Huffman* realiza a fase de codificação, passando-se para a fase de cálculo da matriz de distâncias com algoritmo NCD.

3.2.1.4 Matriz de distâncias e árvore filogenética

Com todas as fases anteriores finalizadas, será calculado a matriz de distâncias através do algoritmo NCD, essa matriz será utilizada no algoritmo *Neighbor Joining*, para a geração da Árvore Filogenética, representando-se assim os agrupamentos das séries.

3.3 Resultado esperados

O resultado esperado culmina na criação de um modelo de representação do conjunto de séries numéricas, assim como a concepção de um modelo de classificação para um conjunto de séries numéricas.

Com esse modelo espera-se o aumento da precisão em análise de série numérica, que também contribui para a academia com esse modelo desenvolvido.

4 Resultados e discussões

Nesta seção expõe-se os resultados das experiências realizadas durante as pesquisas, todas as experiências foram realizadas utilizando-se dois valores de fator de redução de dimensionalidade (FR).

Para todas as experiência foi realizado um método para avaliação do classificador por meio de uma constante de normalização G, para que assim possa ter um meio de medir a perfomance do classificador.

Para as figuras com a representação da série numérica e a Árvore Filogenética, escolheu-se as fonte de dados mais pequenas, pois como possui menos dados, para assim que a figura gerada possua uma melhor visualização.

4.1 Constante de normalização

Para realizar avaliação do classificador foi utilizada a constante de normalização G que é definida na equação 1 e 2.

$$G = \frac{\sum_{i=1}^{n_a-1} a_i}{n_v - 1 - (n_c - 1)} \tag{1}$$

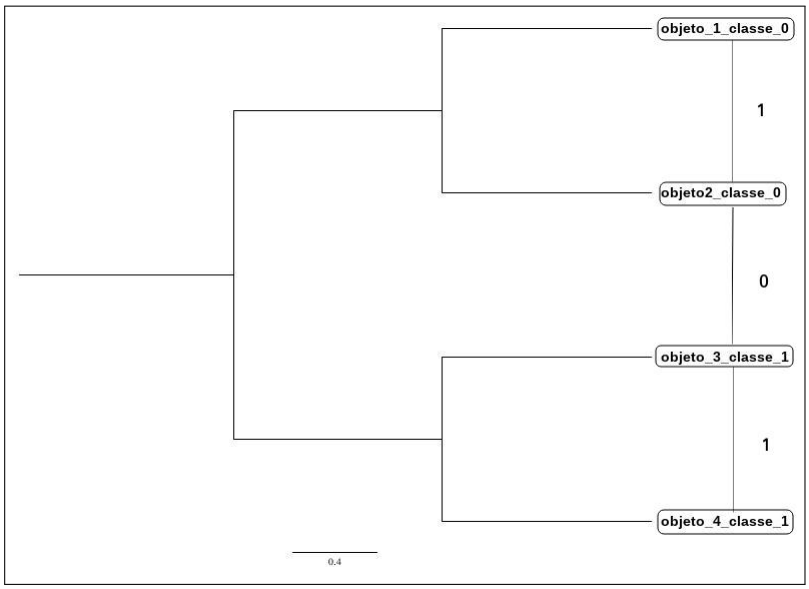
$$G = \frac{\sum_{i=1}^{n_a-1} a_i}{n_v - n_c} \tag{2}$$

Onde:

- n_v: número de vértice
- n_c: número de classes
- a_i: valor da aresta
- n_a: total de arestas

Para exemplificar, na Figura 16 é exibida uma árvore onde os nós folhas foram transformadas em vértices e em cada aresta temos o valor de 0 (os nós não são do mesmo pai) ou 1 (os nós são do mesmo pai), seguindo a equação 2, é realizado o somatório dos valores da arestas $\sum_{i=1}^{3} a_i = 1 + 0 + 1 = 2$, aplicando o resultado do somatório e os valores $n_c = 2$ e $n_v = 4$, $G = \frac{2}{4-2} = 1$.

Figura 16 – Árvore transformada em um grafo



Fonte: O autor.

4.2 Resultado 1 - Dados climáticos

Dentro dos dados climáticos foram utilizados os dados de precipitação total e temperatura média total durante os 12 meses que compõe o ano de 2018.

4.2.1 Fonte de dados - Precipitação total

O objetivo foi agrupar cidades com clima similar em relação a quantidade de precipitação, divididas em duas classes: chuvoso e não chuvoso. Os dados se referem à precipitação total durante os 12 meses do ano de 2018, ou seja, uma série numérica contendo 12 elementos.

Nos testes realizados, observou-se que com o FR = 3, obteve-se melhor resultado em dois testes, 3 testes mantiveram valor da avaliação com a alteração do valor de FR e somente um teste obteve um melhor com o FR = 4. Observando-se a média dos valores das avaliações, os melhores resultados utilizou-se FR = 3.

Realizou-se no total 7 testes com esses dados, a saída obtida com a fonte1B foi a seguinte árvore filogenética exibida na Figura 17, os resultados dos outros 6 testes é exibido na Tabela 3.

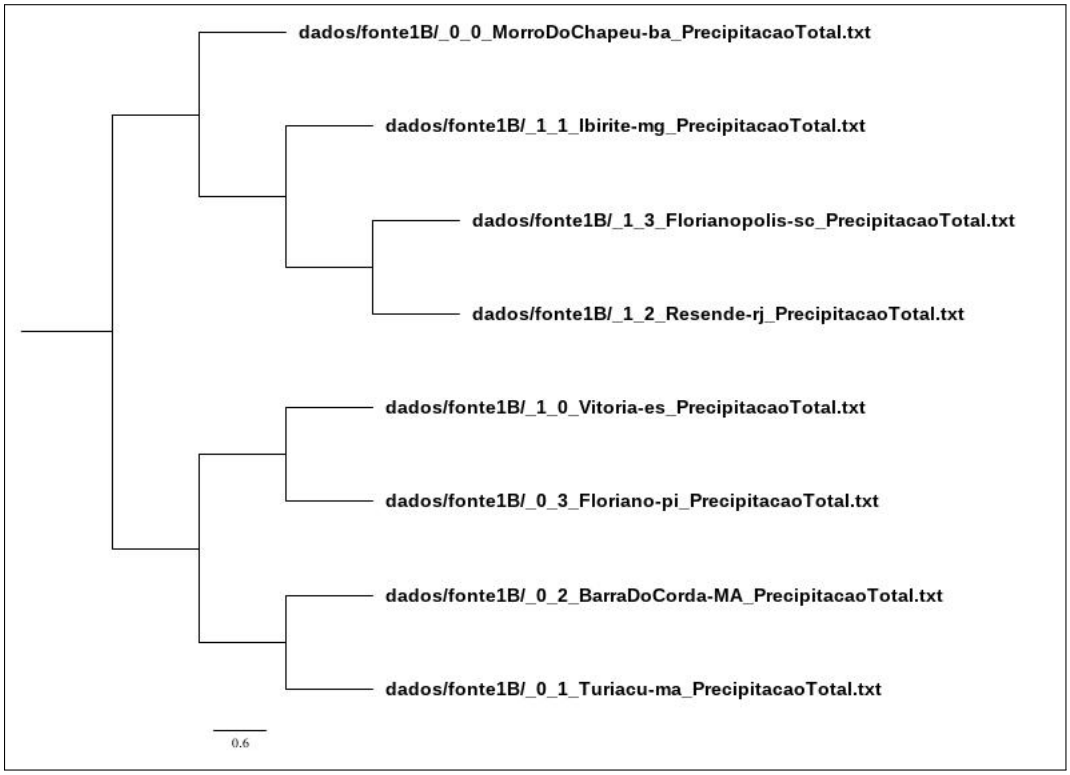
Na Figura 18 exibem-se as as representação das séries comprimidas, com essa representação é possibilitado comprovar visualmente quais séries possuem similaridade entre si.

Tabela 3 – Fonte de dados - Precipitação total.

Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)	Avaliação (FR = 3)
fonte1A	4	1.0	1.0
fonte1B	8	1.0	1.0
fonte1C	12	0.8	0.8
fonte1D	16	0.5714	0.8571
fonte1E	20	0.5556	0.7778
fonte1F	40	0.5263	0.6316
fonte1G	60	0.6207	0.5517
Média Avaliação		0.7249	0.8026

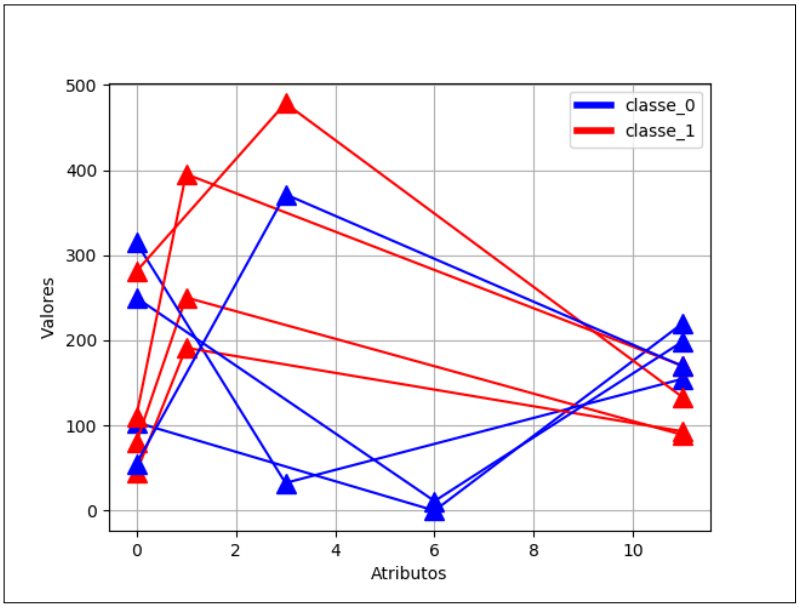
Fonte: O autor.

Figura 17 – Árvore filogenética - FR = 3 (fonte1B)



Fonte: O autor.

Figura 18 – Série comprimida - FR = 3 (fonte1B)



Fonte: O autor.

4.2.2 Fonte de dados - Temperatura média total

O objetivo foi agrupar cidades com clima similar em relação a quantidade de precipitação, sendo duas classes: quente e frio. Os dados se referem à temperatura máxima média durante os 12 meses do ano de 2018, ou seja, uma série numérica contendo 12 elementos.

Tabela 4 – Fonte de dados - Temperatura média total.

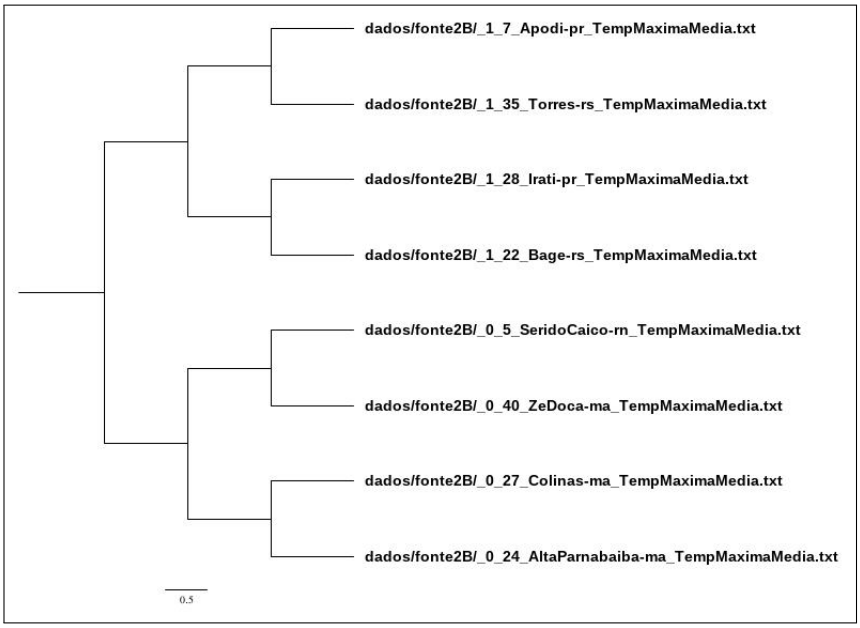
Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)	Avaliação (FR = 3)
fonte2A	4	1.0	1.0
fonte2B	8	1.0	1.0
fonte2C	12	1.0	1.0
fonte2D	16	1.0	1.0
fonte2E	20	1.0	1.0
fonte2F	40	0.8947	1.0
fonte2G	60	0.8276	0.9655
Média Avaliação		0.9603	0.995

Fonte: O autor.

Realizou-se assim 7 testes, onde na Figura 19 é exibido a árvore filogenética gerada pelo teste realizado com a fonte2B, na Tabela 4 são exibidos os outros 6 resultados.

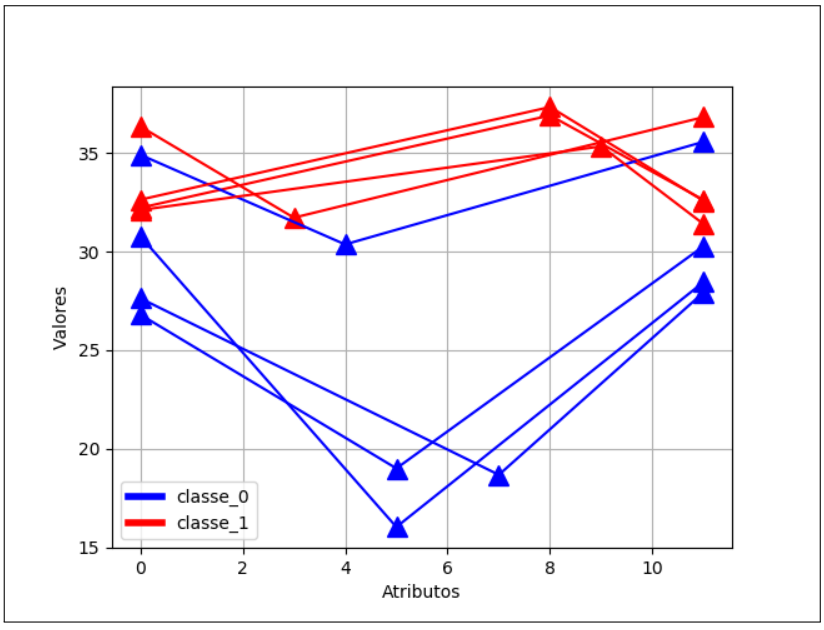
Na Figura 20 é exibido as representações das séries comprimidas da fonte2B, com essas representações é possibilitado comprovar visualmente quais séries possuem similaridade entre si.

Figura 19 – Árvore filogenética - FR = 3 (fonte2B)



Fonte: O autor.

Figura 20 – Série comprimida - FR = 3 (fonte2B)



Fonte: O autor.

Neste resultado onde se obteve um dos melhores resultados na avaliação, fica bem evidente na representação como as classes foram bem agrupadas, as cidades com temperatura mais alta foram separadas das cidades com temperatura mais baixa. Desta forma com o FR = 3, houve uma melhora em todos os resultados da avaliação. Com a média dos valores da avaliação, observou-se que os testes com FR = 3, obtiveram os melhores resultados.

4.3 Resultado 2 - PIB nominal

O objetivo culminou em agrupar países com valor do PIB nominal similar, sendo duas classes: maiores que 9999 e menores que 9999. Os dados se referem ao valor do PIB durante o período 2000-2009 (10 anos), ou seja, uma série numérica contendo 10 elementos.

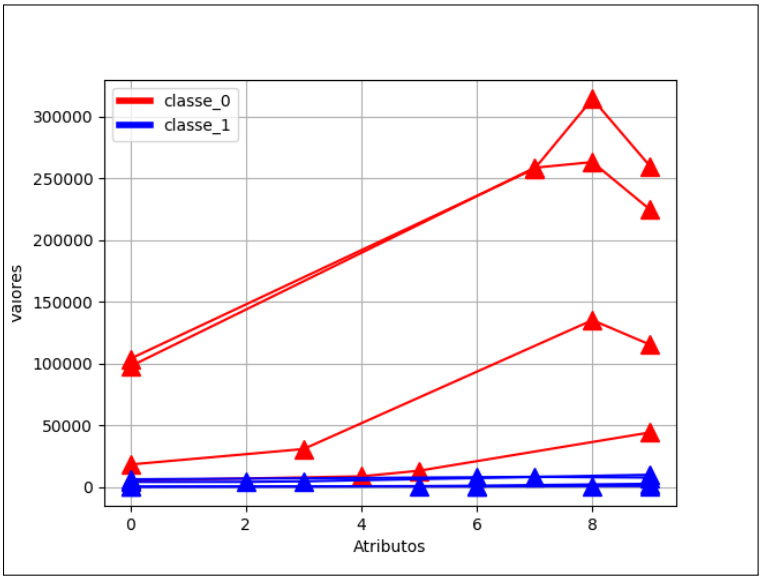
Realizou-se 5 testes com os dados do PIB nominal, onde na Figura 22 mostra-se a árvore filogenética retornada pelo algoritmo, na Tabela 5 é exposto os outros 4 testes realizados.

Na Figura 21 é exibido a representação das séries comprimidas, onde frente a essa representação, é possibilitado comprovar visualmente quais séries tem similaridade entre si.

Dentro dos 5 casos de testes, 3 mantiveram o mesmo valor de avaliação, com a alteração do FR. Um dos testes teve melhor avaliação com FR = 3, enquanto o outro teve melhor avaliação com FR = 4.

Com as média dos valores de avaliação, observa-se que os melhores resultados foi utilizando FR = 4.

Figura 21 – Série comprimida - FR = 4 (fonte3B)



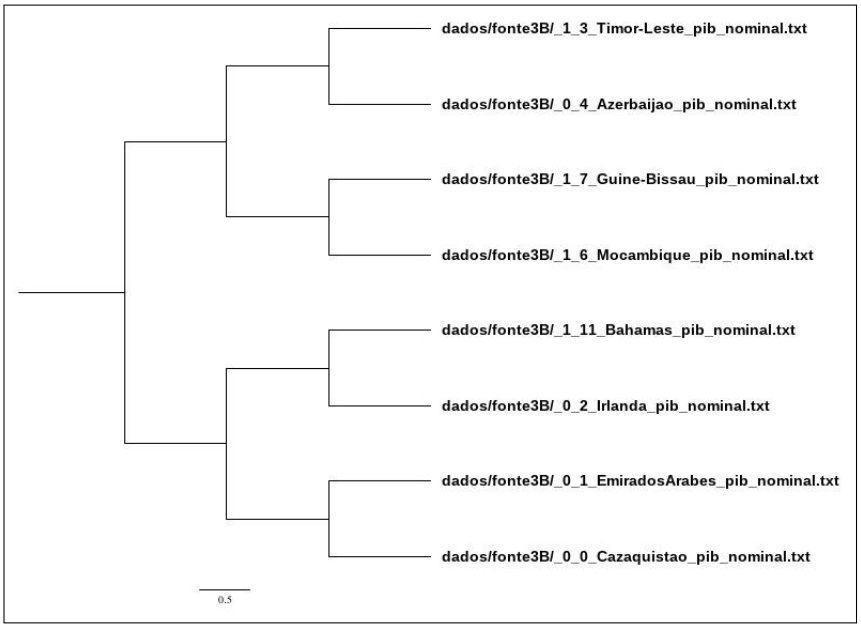
Fonte: O autor.

Tabela 5 – Fonte de dados - PIB nominal.

Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)	Avaliação (FR = 3)
fonte3A	4	1.0	1.0
fonte3B	8	0.6667	0.6667
fonte3C	12	0.8	0.8
fonte3D	16	0.7143	0.5714
fonte3E	20	0.6667	0.7778
Média Avaliação		0.7695	0.7632

Fonte: O autor.

Figura 22 – Árvore filogenética - FR = 4 (fonte3B)



Fonte: O autor.

4.4 Resultado 3 - Vinhos

O objetivo foi de agrupar os tipos de vinhos segundo suas características químicas, cujos vinhos possuem 3 classes, e assim cada série representa um vinho e possui 13 elementos (atributos).

Foram 8 testes com esses dados, o resultado do teste com a fonte5A é exibido na Figura 23, os restantes dos testes são detalhados na Tabela 6.

Nos testes dessa fonte, os dados são bem próximos, devido os valores dos conjuntos numéricos serem muito próximos e ainda assim obteve resultados bons, tanto com valor de

FR = 3 quanto o FR = 4, mas com FR menor obteve melhor desempenho no resultado da avaliação.

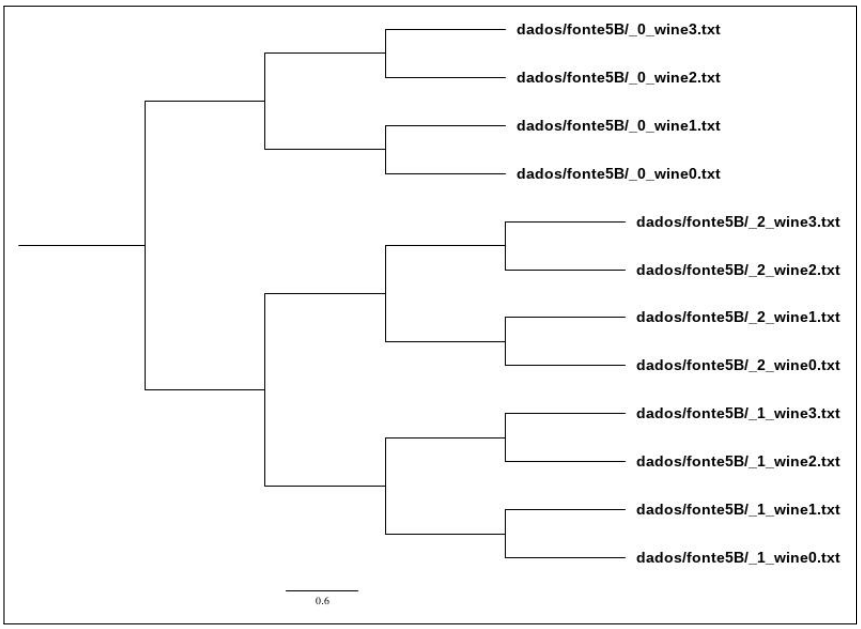
Pode ser observado que na série original exibida na Figura 25 e na série comprimida exibida na Figura 24, os dados ainda são bem próximos no início, mas no final a diferença entre as classes ficam evidentes, ou seja, mesmo com a compressão, a série não afetou o resultado. Através dos valores da média de avaliações é observado que os melhores resultados foi o que se utilizou FR = 3.

Tabela 6 – Fonte de dados - Vinhos.

Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)	Avaliação (FR = 3)
fonte5A	6	1.0	1.0
fonte5B	12	1.0	1.0
fonte5C	18	1.0	1.0
fonte5D	24	0.9524	1.0
fonte5E	30	0.963	1.0
fonte5F	60	0.8421	0.8246
fonte5G	90	0.7931	0.8966
fonte5H	178	0.7886	0.8743
Média Avaliação		0.9174	0.9494

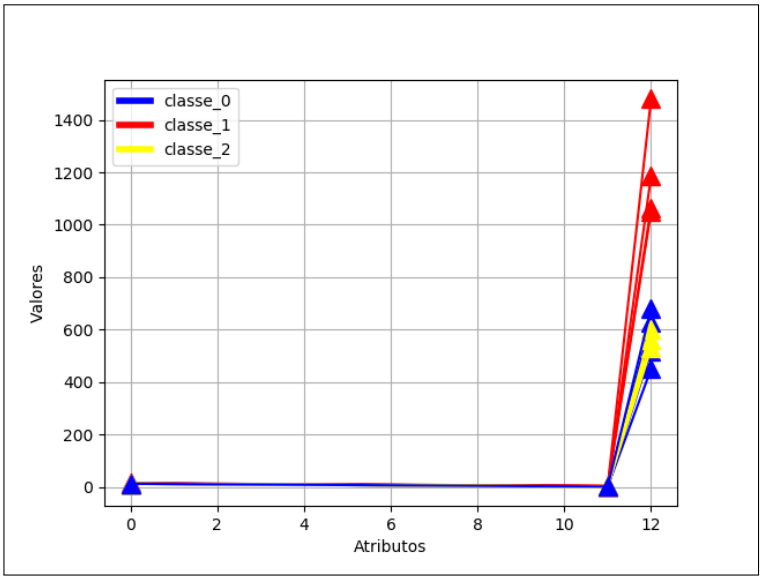
Fonte: O autor.

Figura 23 – Árvore filogenética - FR = 3 (fonte5B)



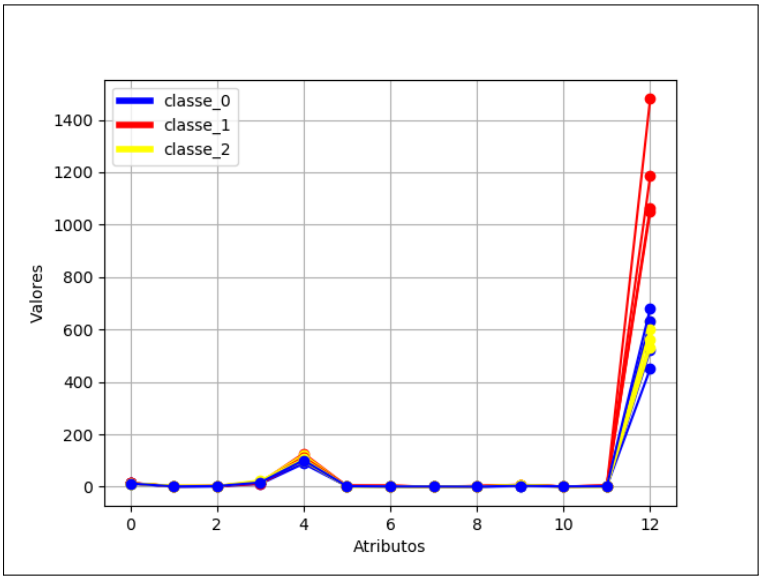
Fonte: O autor.

Figura 24 – Série comprimida - FR = 3 (fonte5B)



Fonte: O autor.

Figura 25 – Série Original (fonte5B)



Fonte: O autor.

4.5 Resultado 4 - Concha orgânica

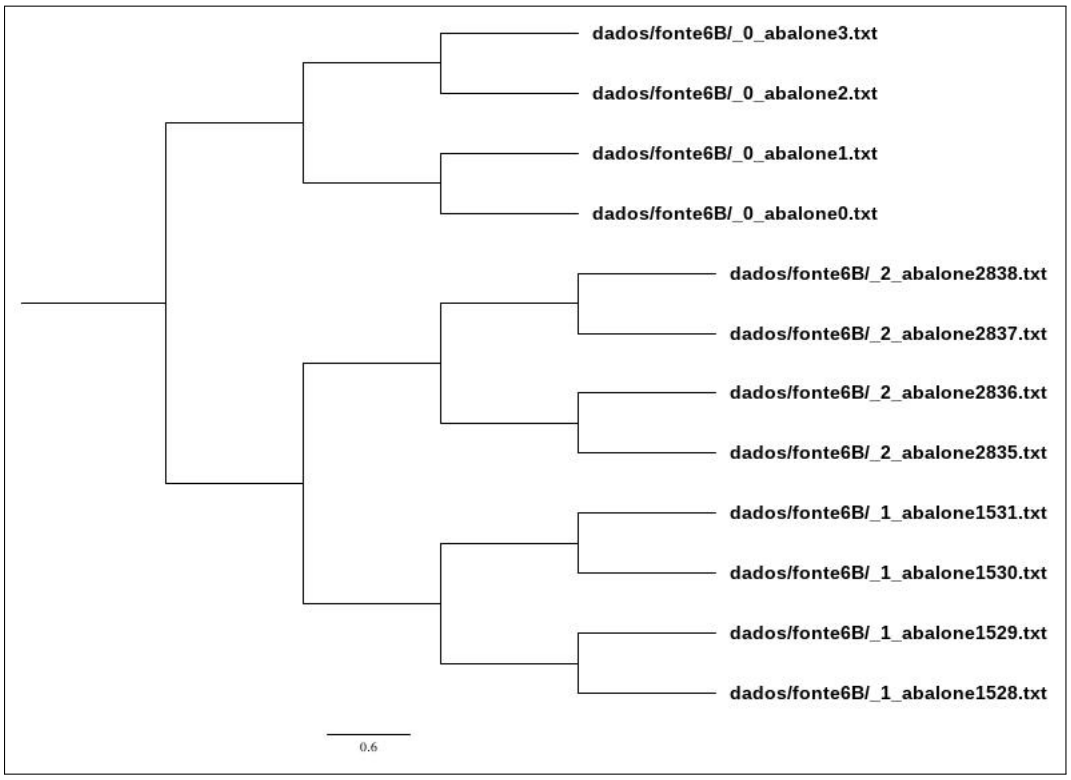
O objetivo foi o de agrupar as conchas segundo suas características físicas, onde identificou-se 3 classes de conchas, onde cada concha representa uma série numérica e cada uma possui 8 elementos. Com esses dados foram realizados 7 testes, detalhados na Tabela 7, na Figura 26 é exibido o a árvore filogenética da fonte6B.

Tabela 7 – Fonte de dados - Conchas.

Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)	Avaliação (FR = 3)
fonte6A	6	1.0	1.0
fonte6B	12	1.0	1.0
fonte6C	18	0.9333	0.8
fonte6D	24	0.8571	0.9048
fonte6E	30	0.8519	0.7037
fonte6F	60	0.8246	0.8070
fonte6G	90	0.9195	0.7931
Média Avaliação		0.9123	0.8584

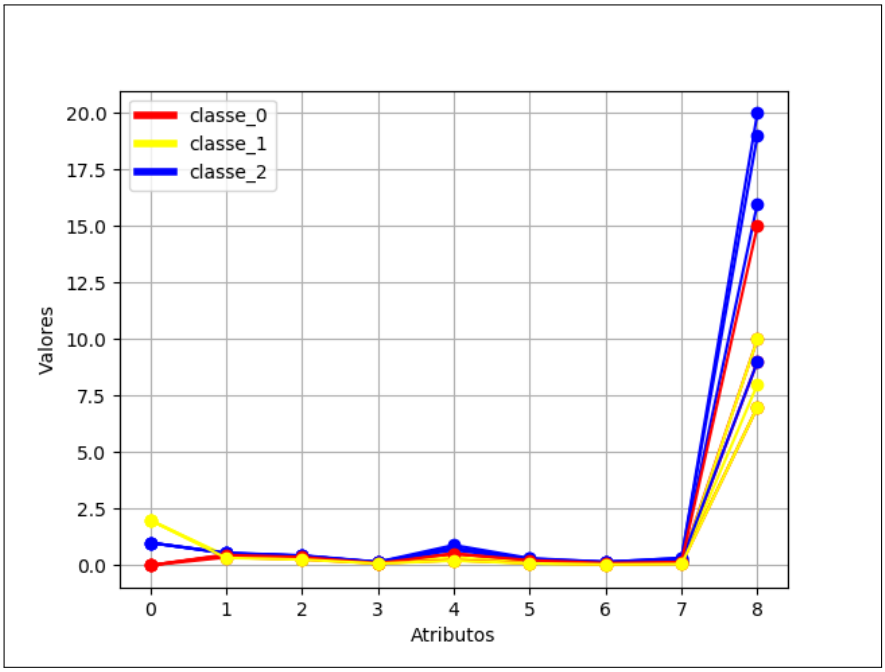
Fonte: O autor.

Figura 26 – Árvore filogenética - FR = 4 (fonte6B)



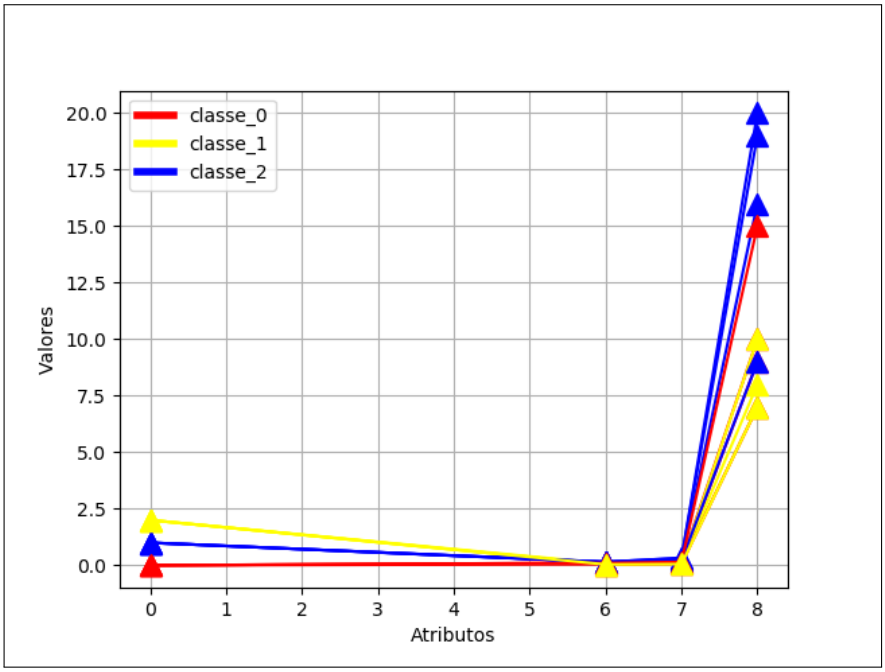
Fonte: O autor.

Figura 27 – Série Original (fonte6B)



Fonte: O autor.

Figura 28 – Série Comprimida - FR = 4 (fonte6B)



Fonte: O autor.

Estes são outros dados que se mostram bem próximos dos outros, na série original

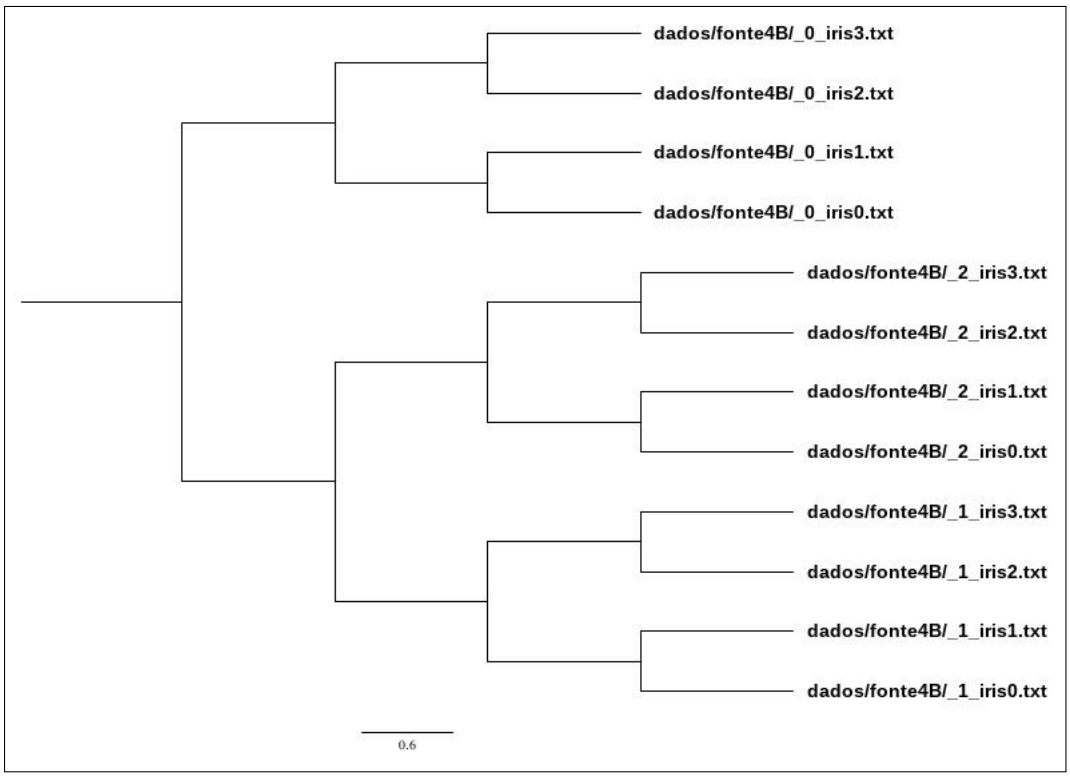
os dados eram bem próximos como observado na Figura 28 e na Figura 27, após a compressão os dados ficaram mais afastados, o que gerou bons resultados na avaliação. Dentro os 7 casos de testes, somente 2 mantiveram os valores da avaliação iguais, 1 teste obteve melhor resultado com FR = 4, enquanto os outros 4 testes tiveram melhor resultado com FR = 3. Com a média dos valores de avaliações, conclui-se que os melhores resultados foram obtidos com valor de FR = 4.

4.6 Resultado 5 - Iris

Esta experiência teve por objetivo, agrupar em 3 classes de flores segundo as características físicas da pétala e sépala, onde cada flor representada por uma série numérica, continham 4 elementos.

Na Tabela 8 são detalhados os 8 testes realizados, na Figura 29 é exibido a árvore filogenética gerada da fonte4B, na Figura 30 é exibida a série comprimida para verificação da similaridade.

Figura 29 – Árvore filogenética - FR = 4 (fonte4B)



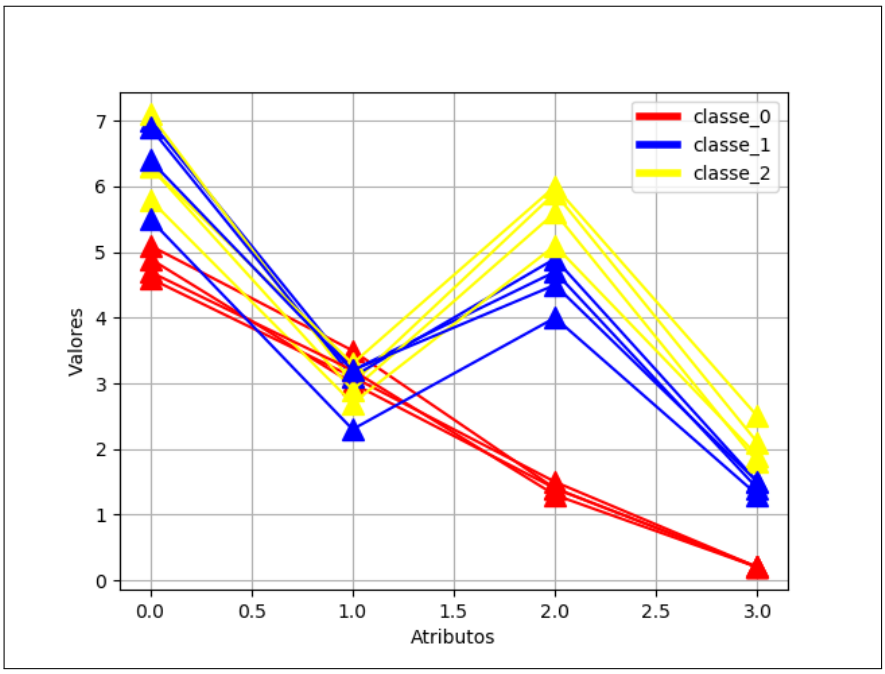
Fonte: O autor.

Tabela 8 – Fonte de dados - Iris.

Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)	Avaliação (FR = 3)
fonte4A	6	1.0	1.0
fonte4B	12	1.0	1.0
fonte4C	18	1.0	0.7333
fonte4D	24	0.9524	0.8095
fonte4E	30	0.9259	0.7778
fonte4F	60	1.0	0.8596
fonte4G	90	0.977	0.8966
fonte4H	150	0.966	0.8844
Média Avaliação		0.9777	0.8702

Fonte: O autor.

Figura 30 – Série comprimida - FR = 4 (fonte4B)



Fonte: O autor.

Nesta base de dados clássica, observou-se que os melhores resultados foi obtido que com o FR = 4, todos os testes teve um aumento na avaliação em comparação com o FR = 2, essa diferença pode ser observado nos valores de média de cada avaliação. Portanto conclui-se que dados onde possui menos atributos, se obtém melhores resultados utilizando-se valor de FR = 3.

4.7 Resultado 6 - Câncer de mama

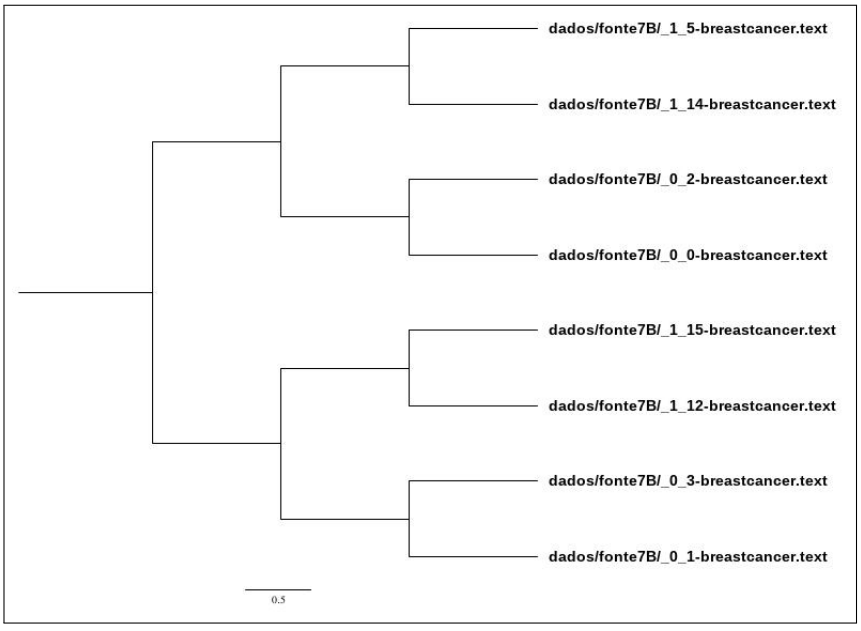
A experiência teve por objetivo agrupar cânceres em 2 classes segundo as características da célula cancerígena, onde cada câncer sendo representado por uma série numérica contendo 10 elementos. Os detalhes dos 7 testes realizados são exibidos na Tabela 9, a árvore filogenética da fonte7G é exibida na Figura 31 e a representação de suas série numérica é exibida na Figura 32.

Tabela 9 – Fonte de dados sem atributo numero de amostra - Câncer de mama.

Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)	Avaliação (FR = 3)
fonte7A	4	1.0	1.0
fonte7B	8	0.6667	0.6667
fonte7C	12	0.8	1.0
fonte7D	16	0.8571	1.0
fonte7E	20	0.7778	0.8889
fonte7F	40	0.7368	0.8421
fonte7G	60	0.931	0.931
Média Avaliação		0.8242	0.9041

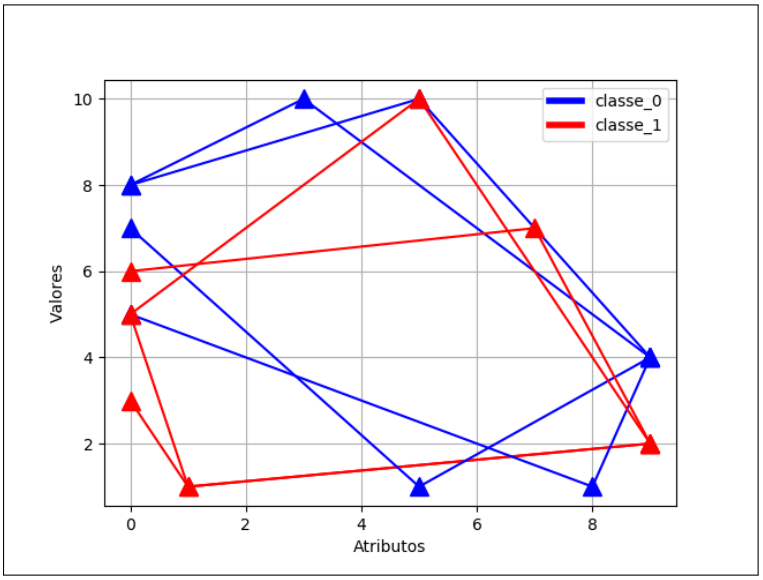
Fonte: O autor.

Figura 31 – Árvore filogenética - FR = 3 (fonte7B)



Fonte: O autor.

Figura 32 – Série numérica - FR = 3 (fonte7B)



Fonte: O autor.

Tabela 10 – Fonte de dados com atributos numero do código de amostra - Câncer de mama.

Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)
fonte7A	4	1.0
fonte7B	8	1.0
fonte7C	12	1.0
fonte7D	16	1.0
fonte7E	20	1.0
fonte7F	24	1.0
fonte7G	28	1.0
fonte7H	32	1.0
fonte7I	36	1.0
fonte7J	40	1.0
fonte7K	80	1.0
fonte7L	683	1.0

Fonte: O autor.

Dentro os 7 casos testes, 3 testes mantiveram o valor igual na avaliação e 4 teste obtiveram melhor desempenho com mudança do valor FR. Com os valores da média de cada avaliação foi obtido que os melhores resultados foram utilizando FR = 3. Como

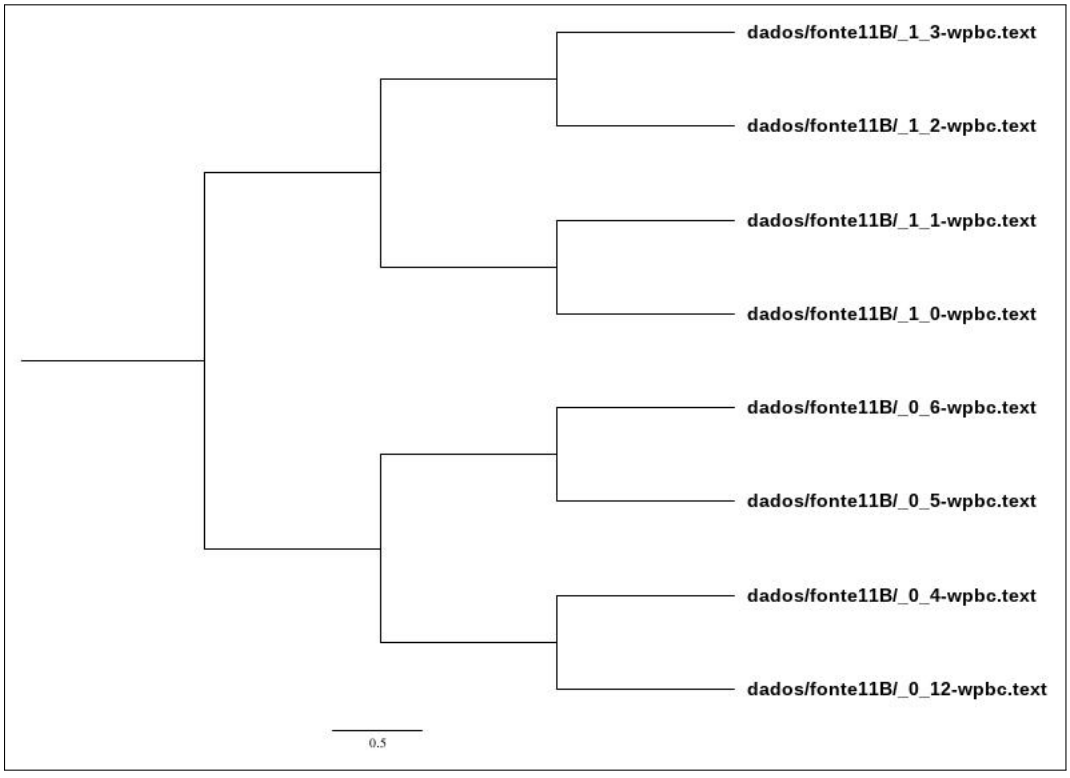
descreveu-se na Experiência 6 - Câncer de mama (Breast *Cancer Wisconsin (Original) Data Set*) foi ignorado um dos atributos, pois possui um valor muito elevado em relação aos outros, nos testes realizados com esse atributo inserido em todas as fontes o resultado da avaliação foi igual 1.0, como é mostrado na Tabela 10.

4.8 Resultado 7 - Prognóstico câncer de mama

A experiência teve por objetivo, agrupar cânceres em duas classes segundo as características do prognóstico do câncer, onde cada câncer foi representado por uma série numérica, contendo-se 34 elementos.

Para demonstração da saída do algoritmo de um dos testes na Figura 33 exibe a árvore filogenética da fonte11B, a representação comprimida das séries numéricas da fonte11B é mostrada na Figura 34 para melhor visualização da similaridade entre as classes, além deste teste foram realizados 6 testes que são detalhadas na Tabela 11.

Figura 33 – Árvore filogenética - FR = 3 (fonte11B)



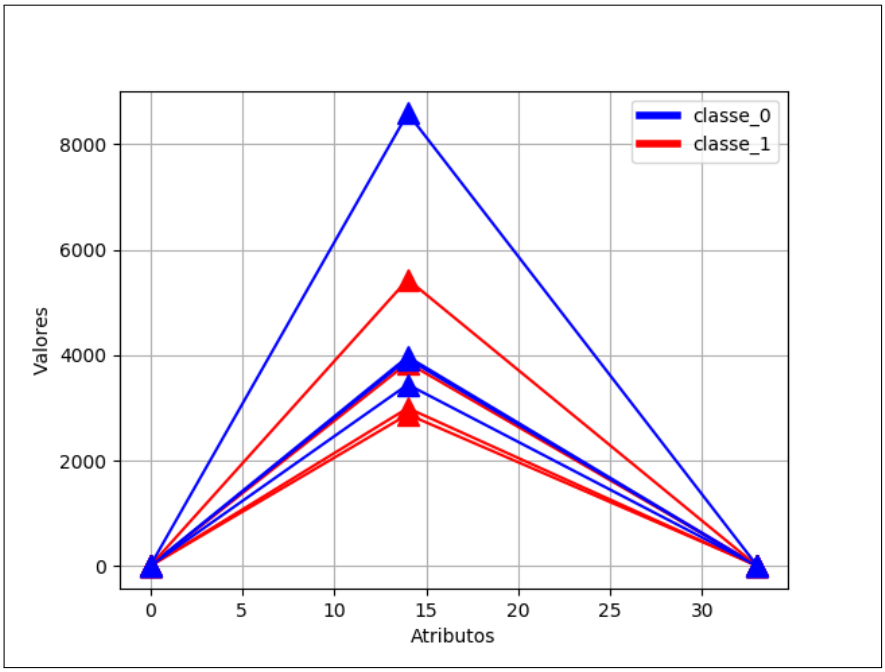
Fonte: O autor.

Tabela 11 – Fonte de dados - Prognóstico câncer de mama.

Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)	Avaliação (FR = 3)
fonte11A	4	1.0	1.0
fonte11B	8	0.6667	1.0
fonte11C	12	0.8	1.0
fonte11D	16	0.7143	1.0
fonte11E	20	0.8889	1.0
fonte11F	40	0.6842	1.0
fonte11G	60	0.6552	0.9655
Média Avaliação		0.7728	0.995

Fonte: O autor.

Figura 34 – Série Comprimida - FR = 3 (fonte11B)



Fonte: O autor.

Das experiências realizadas, essa foi uma das fontes de dados que obteve melhor resultado com FR = 3, a diferença é grande em alguns casos de testes com FR = 3, por exemplo, a fonte11F obteve a avaliação de 0.6667 com FR = 4 e obteve avaliação de 1 com FR = 3.

4.9 Resultado 8 - Diagnóstico câncer de mama

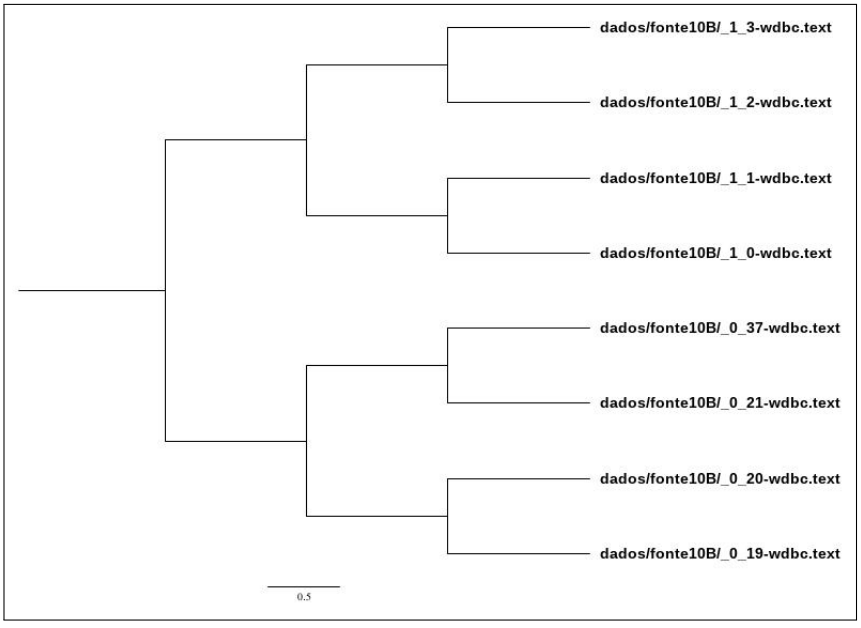
A experiência teve por objetivo, agrupar cânceres em duas classes segundo as características do diagnóstico do câncer, onde cada câncer foi representado por uma série numérica, contendo 31 elementos. Foram realizados 7 testes que são detalhados na Tabela 12, a árvore filogenética gerado de um dos testes é exibida na Figura 35 e a representação das séries comprimidas é exibida na Figura 36.

Tabela 12 – Fonte de dados - Diagnóstico câncer de mama.

Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)	Avaliação (FR = 3)
fonte10A	4	1.0	1.0
fonte10B	8	1.0	1.0
fonte10C	12	0.8	1.0
fonte10D	16	0.8571	1.0
fonte10E	20	0.8889	1.0
fonte10F	40	0.6842	0.9474
fonte10G	60	0.6207	1.0
Média Avaliação		0.8358	0,9925

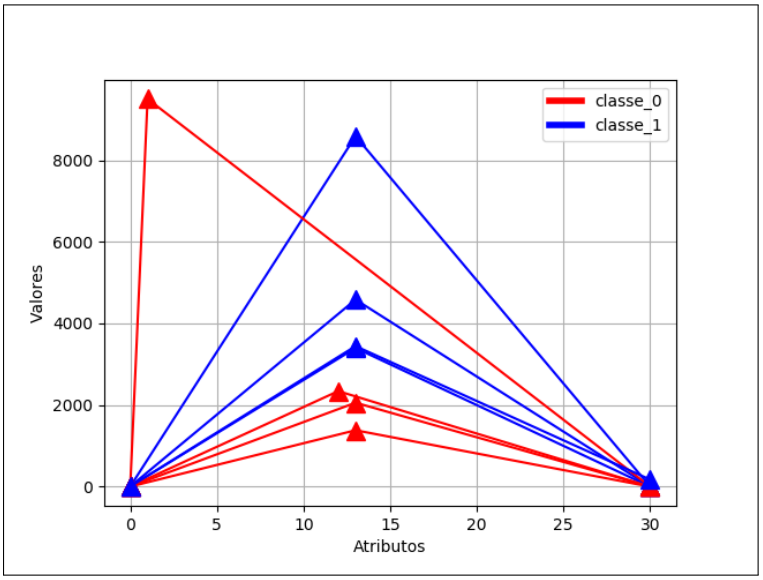
Fonte: O autor.

Figura 35 – Árvore filogenética - FR = 3 (fonte10B)



Fonte: O autor.

Figura 36 – Série comprimida - FR = 3 (fonte10B)



Fonte: O autor.

Outra experiência que teve os melhores resultados utilizando FR = 3, também obteve alguns de casos de testes onde a diferença foi grande, como por exemplo a fonte10F teve avaliação de 0.6207 com FR = 3 e passando-se o FR para 4 obteve uma avaliação de 1.

4.10 Resultado 9 - Escala de balança

Experiência resultou em 3 agrupamentos das balanças, segundo a ponta da balança, onde cada balança representada por uma série numérica que contém 5 elementos. Na Tabela 13 é detalhado os resultados dos 7 testes realizados, a árvore filogenética gerada de um dos teste é exibida na Figura 37, e assim a representação da série numérica é exibida na Figura 38.

Como pode ser constatado nos casos de testes na Tabela 13 om FR = 3, o único resultado que obteve 1 na avaliação foi a fonteA, pode ser observado também que no teste com FR = 4, obteve-se 2 casos de testes com avaliação 1, na série comprimida exibida na Figura 38, observa-se que os dados dessa série são muitos próximos.

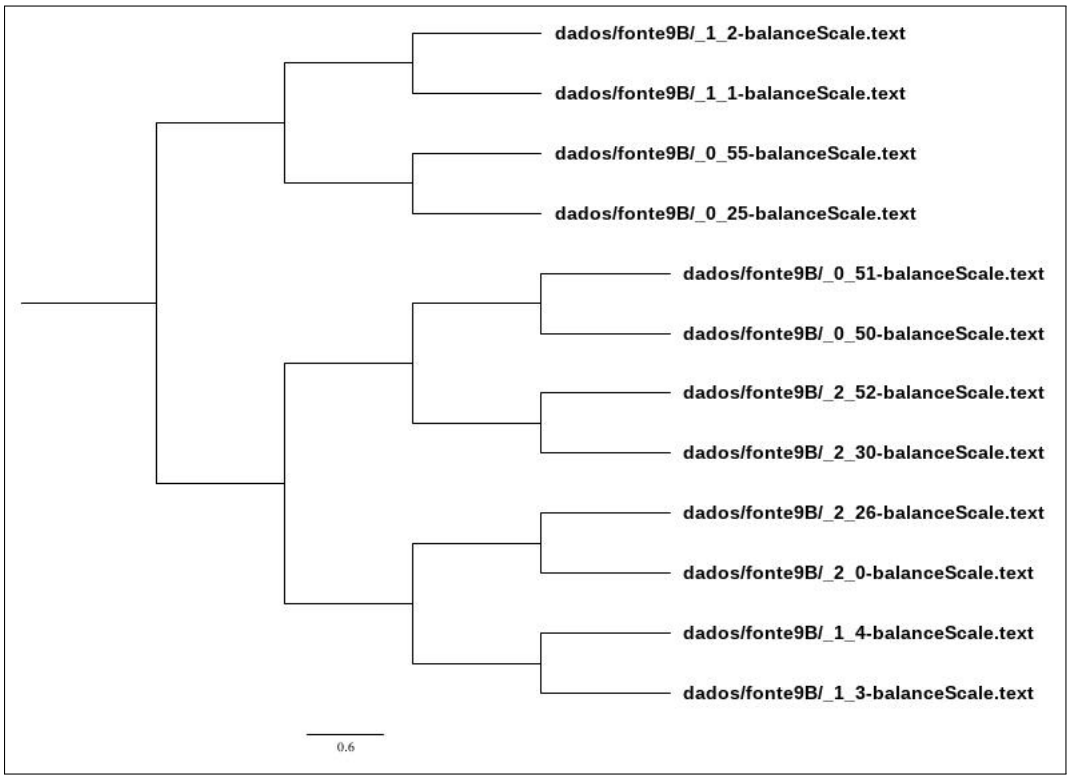
Portanto conclui-se que os dados que possuem poucos atributos e seus elementos são muito próximos entre si, verificando-se que se obtém melhor resultado, diminuindo no máximo o FR. Com realização da média dos valores das avaliações, compreende-se que os melhores resultados foram utilizando-se FR = 4.

Tabela 13 – Fonte de dados - Escala de balança.

Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)	Avaliação (FR = 3)
fonte9A	4	1.0	1.0
fonte9B	8	1.0	0.7778
fonte9C	12	0.9333	0.7333
fonte9D	16	0.9525	0.8571
fonte9E	20	0.8519	0.9259
fonte9F	40	0.8596	0.8947
fonte9G	60	0.8966	0.8276
Média Avaliação		0.9277	0.8595

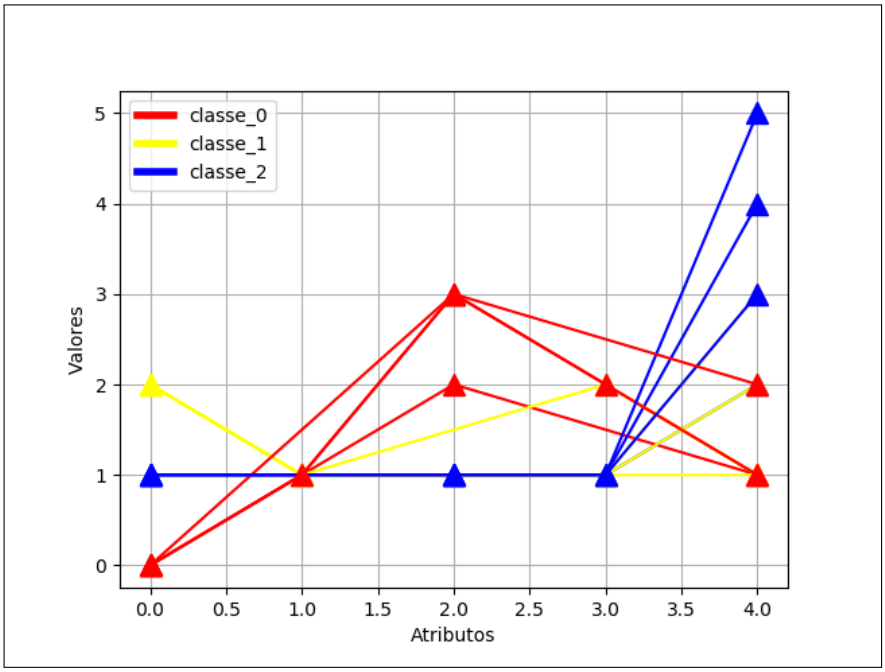
Fonte: O autor.

Figura 37 – Árvore filogenética - FR = 4 (fonte9B)



Fonte: O autor.

Figura 38 – Série comprimida - FR = 4 (fonte9B)



Fonte: O autor.

4.11 Resultado 10 - Escolha de método contraceptivo

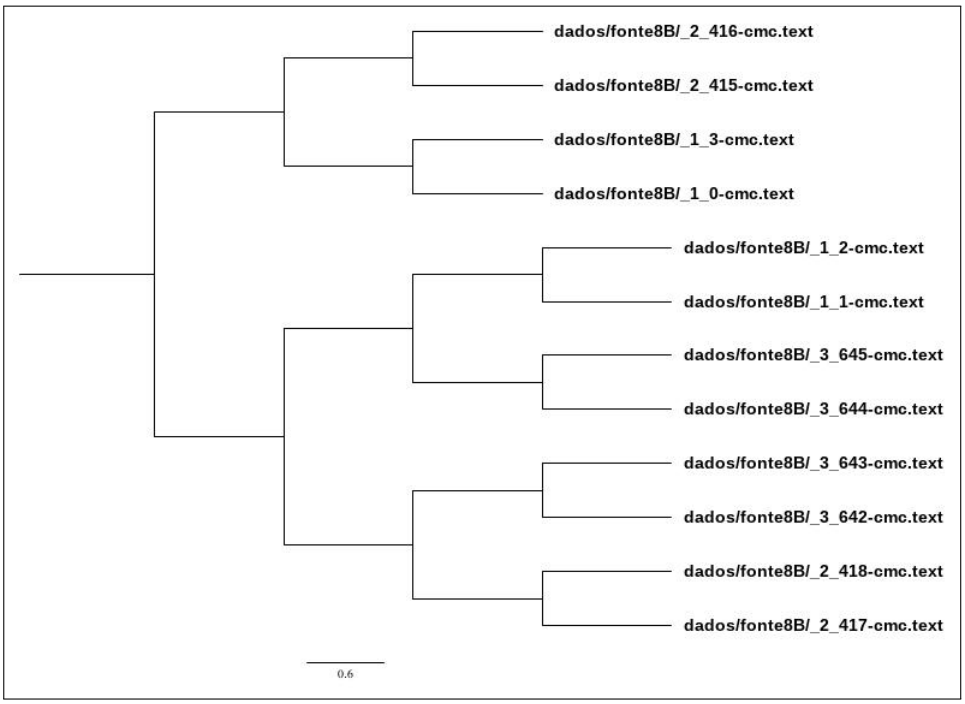
A experiência teve por objetivo, de agrupar mulheres de acordo com sua escolha de método contraceptivo, desta forma cada mulher teve 10 atributos. Foram realizados no total 7 testes utilizando-se essa fonte de dados, na Tabela 14 é detalhado o resultados dos testes realizados , um dos resultados é exibido na Figura 39 a árvore filogenética gerada pela fonte8B e a representação de suas séries comprimidas é exibida na Figura 40.

Tabela 14 – Fonte de dados - Escolha de método contraceptivo.

Fonte de dados	Núm. de instâncias	Avaliação (FR = 4)	Avaliação (FR = 3)
fonte8A	4	1.0	1.0
fonte8B	8	1.0	1.0
fonte8C	12	0.7333	0.8
fonte8D	16	0.7619	0.6667
fonte8E	20	0.5926	0.6667
fonte8F	40	0.6316	0.6140
fonte8G	60	0.6552	0.6782
Média Avaliação		0.7678	0.7751

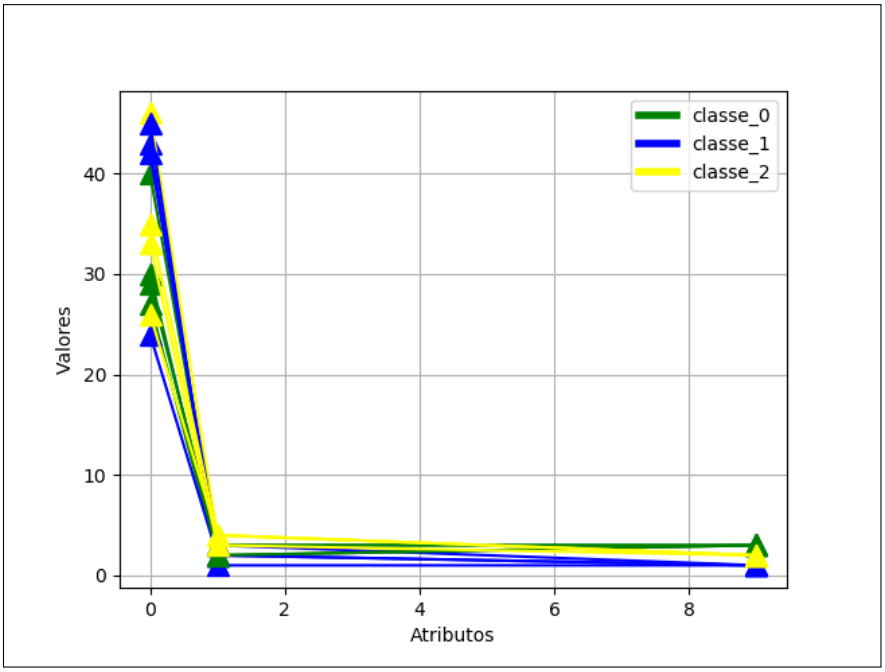
Fonte: O autor.

Figura 39 – Árvore filogenética - FR = 3 (fonte8B)



Fonte: O autor.

Figura 40 – Série comprimida - FR = 3 (fonte8B)



Fonte: O autor.

Este foi uma experiência onde os elementos da série são bem próximos, como pode ser observado na Figura 40, que não teve uma grande alteração no valor da avaliação, mesmo com a alteração do fator de redução. Realizando-se as médias dos valores das avaliações, constatou-se que os melhores resultados foram obtidos utilizando-se FR = 3.

4.12 Discussão dos resultados

Durante a realização das 10 experiências, observou-se vários resultados de avaliação, nos casos de testes em cada experiência, a realização da média para casos de testes com FR igual 3 e 4 foi um fator determinante, para se obter em quais casos de testes tiveram o melhor resultado. Das experiências realizadas, conclui-se que em séries onde os dados são muito próximos entre em si ou possuem poucos atributos, é melhor a utilização de um FR = 4, e enquanto em séries, esses casos não ocorrem e principalmente se possuem várias atributos, se obtém um resultado muito melhor com FR = 3.

5 CONCLUSÃO

Dada a importância da tarefa de análise de séries numéricas, tarefa que demanda tempo e custo alto para ser desenvolvido, o modelo de classificação de séries numéricas trazem muita praticidade, como uma ferramenta de apoio de decisão.

O desenvolvimento da presente pesquisa, possibilitou a concepção de um modelo de classificação, para um conjunto de séries numéricas, bem como a definição do modelo de representação do conjunto de séries numéricas. Assim um conjunto de dados que é encontrado em variadas áreas e, portanto, existe uma infinidade de aplicações que esse modelo de classificação pode ser utilizado. Com os testes realizados, verificou-se que mesmo com método de compressão numérica, algumas bases de dados tiveram diferentes resultados com o aumento e diminuição do fator de redução de dimensionalidade, verificou-se que a redução de dimensionalidade traz melhor desempenho, quando aplicado em séries que possuam uma alta quantidade de atributos.

5.1 Trabalhos Futuros

Como trabalho futuro propõe-se o desenvolvimento de um classificador que utilizará a árvore filogenética como modelo de conhecimento, ele receberá uma entrada de um série numérica, para inserir um rótulo a série de entrada. Os rótulos serão colocados previamente. O objetivo culmina em encontrar em local da árvore filogenética que a série de entrada se encaixa, a ideia do algoritmo pode ser baseada na lógica do KNN (*K-nearest neighbors*) para observação dos vizinhos mais próximos.

Referências Bibliográficas

ANDERSON, E. The species problem in iris. *Annals of the Missouri Botanical Garden*, Missouri Botanical Garden Press, v. 23, n. 3, p. 457–509, 1936. ISSN 00266493. Disponível em: <http://www.jstor.org/stable/2394164>. Citado na página 13.

CAMARGO, S. d. S. Um modelo neural de aprimoramento progressivo para redução de dimensionalidade. 2010. Citado na página 8.

CESAR, B. K. M. Estudo e extensão da metodologia DAMICORE para tarefas de classificação. p. 119, 2016. Citado 2 vezes nas páginas 4 e 5.

DUA, D.; GRAFF, C. *UCI Machine Learning Repository*. 2017. Disponível em: <http://archive.ics.uci.edu/ml>. Citado 9 vezes nas páginas 9, 11, 12, 13, 14, 15, 16, 17 e 18.

ERCOLE, G. *Cálculo V Séries numéricas*. [S.l.]: UFMG, 2010. 1–88 p. ISBN 9788570418449. Citado na página 3.

FILHO, J. de L. On estimation of a probability density function and mode. *Implementação Modular da Técnica de Compressão e Descompressão JPEG para imagens*, IFSC, 1994. Citado na página 7.

INMET. *BDMEP Banco de Dados Meteorológicos para Ensino e Pesquisa*. 2019. Disponível em: <http://www.inmet.gov.br>. Citado na página 9.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003. Citado na página 4.

RAJ, J. T. *A beginner's guide to dimensionality reduction in Machine Learning*. 2019. [Online; acessado 1-Julho-2019]. Disponível em: <https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e>. Citado na página 7.

RAMOS, D. de M. *Sequência Numérica*. 2011. Disponível em: <https://brasilescola.uol.com.br/matematica/sequencia-numerica.htm>. Acesso em: 30 de março de 2019. Citado na página 3.

RUSSEL, S.; NORVIG, P. *Inteligência artificial*. CAMPUS - RJ, 2004. ISBN 9788535211771. Disponível em: <https://books.google.com.br/books?id=wBMvAAAACAAJ>. Citado na página 4.

SANCHES, A.; CARDOSO, J. M.; DELBEM, A. C. Identifying merge-beneficial software kernels for hardware implementation. In: IEEE. *2011 International Conference on Reconfigurable Computing and FPGAs*. [S.l.], 2011. p. 74–79. Citado na página 1.

SILVA, M. E. d. Simulação de quase-monte carlo em finanças: quebrando a maldição da dimensionalidade. 2002. Citado na página 6.

TORRES, B. K. M. Utilizando árvores filogenéticas para a identificação de similaridades em pinturas digitalizadas. p. 76, 2018. Citado na página 5.

ZALEWSKI, W. Modelagem simbólica de padrões morfológicos para classificação de séries temporais. 2015. Citado na página 3.

Apêndices

APÊNDICE A – Algoritmo para inicialização do objeto Ponto

Algoritmo 1: Inicialização do objeto Ponto

Entrada: valor inteiro do elemento da série e valor inteiro do índice da série

função PONTO(*valor*, *indice*)

início

$v \leftarrow valor$

$d \leftarrow 0$

$u \leftarrow 1$

$i \leftarrow indice$

fim

end função

APÊNDICE B – Algoritmo para cálculo da inclinação entre dois pontos

Algoritmo 2: Calculo da inclinação entre dois pontos

Entrada: Ponto inicial - $Ponto(v, i)$, valor inteiro do índice inicial, Ponto final - $Ponto(v,i)$, valor inteiro do índice final

função INCLINACAO($serie$, $pontoInicial$, $indiceInicial$, $pontoFinal$, $indiceFinal$)

início

$tg \leftarrow \frac{\Delta y}{\Delta x} = \frac{pontoFinal.valor - pontoInicial.valor}{indiceFinal - indiceInicial}$

fim

end função

retorna tg

APÊNDICE C – Algoritmo para cálculo da distância entre o valor de um ponto e a reta suporte traçada

Algoritmo 3: Calculo da distancia entre o valor e a reta suporte traçada entre dois pontos

Entrada: Lista série de pontos - $Ponto(v,i)$, valor inteiro do índice esquerda, valor inteiro do índice direito
Saída: Nenhuma
função CALCULADISTANCIA($serie$, $indiceEsquerda$, $indiceDireita$)
início
  $i \leftarrow indiceEsquerda$
  $pontoInicial \leftarrow serie[indiceEsquerda]$
  $pontoFinal \leftarrow serie[indiceDireita]$
  $tg \leftarrow inclinacao(pontoInicial, pontoFinal)$
  enquanto $i < indiceDireita$ **faça**
    $y \leftarrow tg * (i - indiceEsquerda)$
    $serieOriginalPontos[i].d \leftarrow |serie[i].v - y$ $i \leftarrow i + 1$
  fim
fim
end função

APÊNDICE D – Algoritmo para selecionar o ponto que possui maior distância da reta de suporte

Algoritmo 4: Encontra o ponto mais distante da reta de suporte

Entrada: Lista série de pontos - *Ponto*(*v*,*i*)

Saída: Índice do ponto que possui maior distancia

função OBTERPIVOT(*serie*)

início

```
indice ← 0
maior ← −1
para cada ponto i ∈ serie faça
    se serie[i].u = 1 então
        CONTINUE
    fim
    se serie[i].d então
        indice ← i
        maior ← serie[i].d
    fim
fim
```

fim

end função

retorna *indice*

APÊNDICE E – Algoritmo para selecionar o ponto a esquerda

Algoritmo 5: Encontra o índice do ponto a esquerda que já foi marcado como usado

Entrada: Lista série de pontos - *Ponto*(*v*,*i*), valor inteiro do índice do **pivot**
Saída: Índice do ponto a esquerda
função obterReferenciaEsquerda(*serie*, *indicePivot*)
início
    indice ← *indicePivot* − 1
    para *i de indice até -1* **faça**
        se *serie*[*indice*].*u* = 1 **então**
            indice ← *i*
            INTERROMPA
        fim
    fim
fim
end função
retorna *indice*

APÊNDICE F – Algoritmo para selecionar o ponto a esquerda

Algoritmo 6: Encontra o índice do ponto a direita que já foi marcado como usado

Entrada: Lista serie de pontos - $Ponto(v,i)$, valor inteiro do índice do **pivot**
Saída: Índice do ponto a direita
função obterReferenciaDireita(*serie*, *indice***pivot**)
início
  $indice \leftarrow indicePivot - 1$
  para *i de indice até -1* **faça**
    se $serie[indice].u = 1$ **então**
      $indice \leftarrow i$
      INTERROMPA
    fim
  fim
fim
end função
retorna *indice*

APÊNDICE G – Algoritmo para extração de pontos

Algoritmo 7: Extrai pontos marcados

Entrada: Lista serie de pontos - $Ponto(v,i)$

Saída: Lista com índice da serie original e Lista com valores da série

função EXTRAILISTA(serie)

início

- $listaX \leftarrow \emptyset$
- $listaY \leftarrow \emptyset$
- **para** *cada* $i \in serie$ **faça**
  - **se** $serie[i].u = 1$ **então**
    - $listaX \cup serie[i].i$
    - $listaY \cup serie[i].v$
  - **fim**
- **fim**
- **retorna** $listaX, listaY$

fim

end função

APÊNDICE H – Algoritmo para comprimir uma série numérica

Algoritmo 8: Realiza compressão da série original

Entrada: Lista serie Ponto(v, i), valor inteiro fator de redução

Saída: Série comprimida

função COMPRIMIR($serie$, $fatorReducao$)

início

- **se** $fatorReducao <= 2$ **então**
  - **retorna**
- **fim**
- $pontoInicial \leftarrow 0$
- $pontoFinal \leftarrow tamanho(serie)$
- $serie[pontoInicial].u \leftarrow 1$
- $serie[pontoFinal].u \leftarrow 1$
- $pivot = -1$
- **para** x *de 0 até* $fatorReducao - 2$ **faça**
  - **se** $x = 0$ **então**
    - CALCULADISTANCIA($serie, pontoInicial, pontoFinal$)
  - **fim**
  - $pivot \leftarrow$ OBTERPIVOT($serie$)
  - $serieOriginalPontos[i].u = 1$
  - $esquerda \leftarrow$ OBTERREFERENCIAESQUERDA($serie, pivot$)
  - $direita \leftarrow$ OBTERREFERENCIADIREITA($serie, pivot$)
  - $pontoInicial \leftarrow esquerda\ pontoFinal \leftarrow direita$ **se** $x > 0$ **então**
    - CALCULADISTANCIA($serie, pontoInicial, pivot$)
    - CALCULADISTANCIA($serie, pivot, pontoFinal$)
  - **fim**
- **fim**
- $xy \leftarrow$ EXTRAILISTA(SERIE

fim

end função

retorna xy

APÊNDICE I – Algoritmo para abstração de uma série numérica

Algoritmo 9: Abstração da série original para uma lista de objetos *Ponto*(v,i)

Entrada: Série original, valor inteiro do fator de redução
função ABSTRAISERIE($serieOriginal$, $fatorReducao$)
início
  $i \leftarrow 0$ **para** *cada elemento* $e \in serieOriginal$ **faça**
    $ponto \leftarrow Ponto(e, i)$
    $serieOriginalPontos \cup pt$
    $i \leftarrow i + 1$
  fim
fim
end função
retorna COMPRIMIR($serieOriginalPontos$, $fatorReducao$