Universidade da Beira Interior

Ricardo Oliveira nr^{o} 21934, António Drogas nr^{o} 23319

I. Introdução

A. Objectivos

Este trabalho surge no âmbito da unidade curricular de Inteligência Artifícial e tem como principal objectivo desenvolver uma aplicação que permita uma análise automática de sinais cardiotocográficos. Considerando unicamente caracteristicas relevantes ao processo.

B. Organização do relatório

Antes demais, este relatório foi escrito em LaTeXe foi utilizado o *Gummi*, este é um editor de LaTeXdisponível para varias plataformas. Este documento está organizado em 4 capítulos principais:

Capítulo I: Introdução - onde o problema a tratar e os objectivos mais importantes a alcançar são referidos.

Capítulo II: Desenvolvimento - Neste capitulo são apresentadas algumas metodologias utilizadas na classificação, bem como uma descrição dos classificadores utilizados no projeto e os respetivos resultados.

Capítulo III: Resultados Obtidos do Algoritmo Genético - Neste capítulo são apresentados resultados comparativos de dois classificadores disponibilizados no WEKA, nomeadamente o K* e o BayesNet.

Capítulo IV: Reflexão Critíca e Problemas encontrados - onde é feita uma reflexão detalhada de cada momento do trabalho, a distribuição de tarefas, assim como as dificuldades encontradas ao longo do seu desenvolvimento.

Capítulo V: Conclusões e Trabalho Futuro - onde é referido o que foi implementado e também o que ficou por implementar.

II. Desenvolvimento

A. Ferramentas e Tecnologias Utilizadas

A aplicação foi totalmente criada na linguagem de programação C e foi compilada e testada no sistema operativo Linux Mint 17. Para o desenvolvimento foi necessário recorrer a um IDE (Integrated Development Enviroment), neste caso foi utilizado pelo grupo de trabalho o Eclipse 3.8.1.

B. Metodologias

1) Dataset: O dataset utilizado neste trabalho foi disponibilizado pelo docente. O nosso conjunto de dados é composto por 1500 instâncias. Sendo que cada instância é composta por 23 características sendo que a vigésima segunda é a classe e a viségima terceira o NSP(normal,suspected,pathologic).

2) Principais Tarefas de Pré-Processamento: Muitas vezes o conjunto de dados a que temos acesso pode ser inconsistente, redudante e imcompleto. O processo de apredizagem é díficil se o nosso dataset contém dados irrelevantes e ou redundantes. Será descrito nos tópicos que se seguem os passos necessários para obtermos um conjunto de dados optimizado.

- Limpeza dos Dados - Preenche os valores em falta, suaviza dados ruidosos, identifica ou remove "outliers" e resolve inconsistências.
- Integração - Dados de origens diferentes devem ser integrados.
- Transformação (normalização/integração)- Ainda em relação aos dados de entrada é útil realizar uma normalização dos padrões. As variáveis de entrada devem ser pre-processadas de modo que o seu valor médio calculado sobre todo o conjunto de treino seja próximo de zero, ou seja, pequeno quanto ao todo do conjunto. O propósito da normalização é minimizar problemas oriundos do uso de unidades e dispersões distintas entre as variáveis.Inicialmente o método de normalização utilizado na implementação do trabalho foi o Z-Score dado pela expressão $Z_i = \frac{(Xi - \bar{X}_i)}{\sigma_i}$. Posteriormente foi utilizado o método de normalização $min - max$ dado pela formula $x' = \frac{x - min}{max - min}$. No KNN com o data set normalizado desta forma obtivemos um incremento no numero de instancias classificadas corretamente.

- Redução - Esta estapa de pre-processamento tenta reduzir o volume de dados com pouca alteração no resultado final. No decorrer do trabalho ocorreram situações em que o desvio padrão de um atributo era igual a zero. Neste tipo de situação não é possível efectuar a normalização do atributo. Mas analisando com espírito crítico a situação conclui-se que se o desvio padrão de um determinado atributo é igual a zero, significa que o atributo se mantém constante em todas as intâncias das classes em análise e portanto pode ser considerado um aspecto irrelevante no processo de classificação.
- Discretização - Faz parte do processo de redução, mas tem papel importante, especialmente com dados numéricos.

C. K-Nearest Neighbors

A classificação recorrente ao KNN classifica instâncias baseadas na sua similaridade. É um dos algoritmos mais populares em reconhecimento de padrões. Um objecto é classificado segundo a maioria dos seus vizinhos. K é sempre positivo e um número ímpar por forma a evitar empates entre o número de vizinhos. Os vizinhos são escolhidos de um conjunto de dados do o qual é conhecida a classificação correcta. Todos os exemplos são memorizados e usados na fase de aprendizagem. Para obter os resultados deste algoritmo é necessário entender o seu funcionamento, que passa pela aplicação dos passos que se seguem:

1º Estabelecer K, número de vizinhos mais próximos
2º Usando uma medida de distância, calcular a distância entre o objecto X e todos as amostras contidas no nosso conjunto de treino.
3º As distâncias de todos os objectos do conjunto de treino ao objecto X são ordenadas de forma crescente e são determinados os objectos com distância de K mínima.
4º É escolhida a classe com maior número de ocorrências dentro da distância K.

O valor estabelecido para o numero de vizinhos mais próximos pode alterar os valores resultantes deste algoritmo.

Efeito da escolha de K

O KNN é um classificador que possui apenas um parâmetro livre (o número de K-vizinhos) que é controlado pelo utilizador com o objectivo de melhorar a classificação. Mudando o valor de k obtemos espaços de resultados diferentes, sendo que quando k é grande produz fronteiras mais suaves e "ponderadas". Mas quando k é demasiado grande, a nossa classificação será sempre igual à classe com maior frequência relativa. K terá de ser sempre um número impar por forma a evitar empates entre as classe. No caso de mesmo com K ímpar não ser possível fazer um predição, o valor de k é aumentado em pelo menos 2 unidades. Com este aumento são considerados novos vizinhos que levarão ao desempate e a uma posterior predição.

Fronteiras definidas pelo KNN

- K Grande
  Fronteiras suaves, "ponderadas",
  Estimador razoável da densidade de probabilidade.
- K Pequeno
  Fronteiras mais rugosas, sensíveis a outliers,
  Mau estimador de densidade de probabilidade.

Medidas de Distâncias

- euclidiana
- hamming
- minkowski
- mahalanobis

Problemas com KNN

- Exige muita memória para guardar o conjunto de treino.
- Exige muito tempo na fase de classificação.
- São muito sensíveis a outliers (valores atípicos).
- São muito sensíveis à função de distância escolhida.

Resultados da Análise Automática de Sinais Cardiotocográficos

No âmbito do projeto desta unidade curricular foi utilizado um data set que continha 1500 instâncias. Para utilização correta do KNN, este data set é dividido em 2 conjuntos de forma aleatória, sendo que cada vez que este for executado os nossos conjuntos contenham dados diferentes.
Foi estabelecido que 70% do data set seria utilizado como conjunto de treino e os restantes 30% como conjunto de teste.
Este data set continha 23 caracteristicas para estudo, sendo duas destas as classes que iremos atribuir como resultado de classificação de cada instância. Este será o nosso ground-truth.
Os resultados apresentados a seguir são a média de acerto dos vários testes executados ao classificador KNN alterando algumas parámetros tais como a alteração do K e o uso de diferentes tipos de normalização, nomeadamente a normalização Z-score e $min - max$.

K	% de acerto Z-score	% de erro Z-score	% de acerto $min - max$	% de Erro $min - max$	% de acerto WEKA	% de erro WEKA
3	$73,77\%$	$26,23\%$	$85,11\%$	$14,89\%$	$98,89\%$	$1,11\%$
5	$84,89\%$	$15,11\%$	$87,33\%$	$12,67\%$	$98,67\%$	$1,33\%$
30	$84,44\%$	$15,56\%$	$84,56\%$	$15,44\%$	$96,89\%$	$3,11\%$

Foram ainda geradas as matrizes confusão destes resultados sendo estes os valores resultantes com $K = 5$,

- KNN acertou $87,33\%$
- Instâncias classificadas corretamente: 393 - 0,873
- Instâncias classificadas incorretamente: 57 - 0,127

	Classe 1	Classe 2	Classe 3
Classe 1	304	10	0
Classe 2	27	66	7
Classe 3	5	8	23

Os resultados aqui apresentados são muito favoráveis em termos de acerto, supondo que todos os passos do algoritmo foram bem implementados.

D. Naive Bayes

O classificador Naive-Bayes baseia-se na aplicação do Teorema de Bayes para o cálculo das probabilidades necessárias para a classificação. Apesar da sua simplicidade o classificador Bayesiano é bastante utilizado pois por vezes supera os resultados de métodos de classificação mais sofisticados.

Fundamentos Teóricos

O teorema de Bayes providência um método de calcular a probabilidade à posteriori,$P(c|x)$, a partir de $P(c)$,$P(x)$, e $P(x|c)$. O classificador assume que o efeito do valor de um predictor(x) numa dada classe (c) é independente dos valores das outras previsões. A probabilidade à posteriori é calculada da seguinte forma,

$$P(c|x) = \frac{P(c|x)P(c)}{P(x)}$$

Como o denominador $P(x_1...x_n)$ é uma constante, pois não depende da variavel classe que estamos á procura, este pode ser anulado no teorema de Bayes, resultando a seguinte formula,

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

- $P(c|x)$ É a probabilidade à Posteriori da classe c dado atributo x.
- $P(c)$ É a probabilidade à priorí da classe.
- $P(x|c)$ É a likelihood ou seja, a probabilidade do atributo $(x_1, x_2, \cdots, x_n)$ dada a classe c.
- $P(x)$ É a probabilidade á prorí de $(x_1, x_2, \cdots, x_n)$.

Cálculo da Probabilidade Condicional $P(x|c)$

A probabilidade Likelihood é resultante da formula,

$$P(x|c) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

em que a σ e a μ utilizadas são correspodentes ao atributo x.
O valor da *Likelihood* para uma dada instância é o resultado do produtório das Likelihood's dos seus atributos.

No caso de o resultado de algum dos termos calculados ser zero, a sua Likelihood tomará o valor de 0.99, evitando assim o caso de termos um termo absorvente na multiplicação. Desta forma resolvemos o problema da frequência zero.

Escolha da Classe

O cálculo da classe de uma nova intancia consiste no calculo da probabilidade de todas as possiveis classes, escolhendo-se, a seguir a classe com maior probabilidade:

$$argmax P(c|x_1 \cdots x_n) = argmax \prod_{i=0}^{21} P(x_i|c)P(c)$$

Resultados da Análise Automática de Sinais Cardiotocográficos

No data set utilizado pelo algoritmo *Naive-Bayes* aplicamos a mesma formatação do data set utilizado anteriormente no *KNN*. A separação de ambos os conjuntos de treino e teste seguiu a mesma lógica. Como os dados resultantes deste algoritmo utilizando os vários tipos de normalização são bastante semelhantes, é apresentada a média de acerto de todos os testes efectuados.

% de Acerto	% de Erro	% de Acerto WEKA	% de Erro WEKA
81.28%	18.72%	64.44%	35.56%

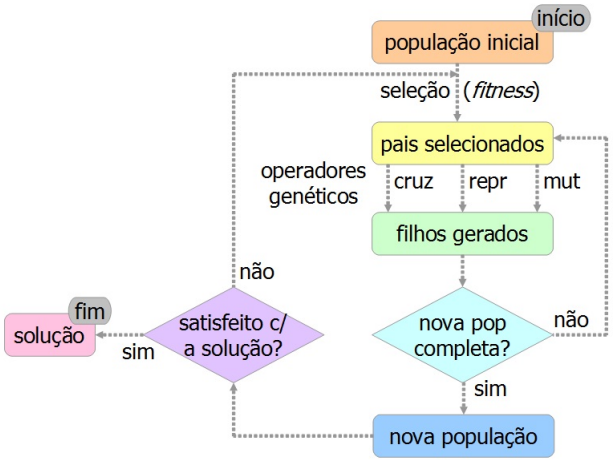
Foram ainda geradas as matrizes confusão destes resultados sendo estes os valores resultantes,

- Naive Byes acertou $80,88\%$
- Instâncias classificadas corretamente: 364
- Instâncias classificadas incorretamente: 86

	Classe 1	Classe 2	Classe 3
Classe 1	171	0	0
Classe 2	61	58	0
Classe 3	22	3	135

E. Algoritmo Genético

Os algoritmos genéticos utilizam conceitos provenientes do principio da seleção natural para abordar uma série ampla de problemas, em especial de optimização. Robustos, genéricos e facilmente adaptáveis, consistem de uma técnica amplamente estudada e utilizada em diversas áreas.



Funcionamento Algoritmo Genético

O funcionamento do Algoritmo Genético foi inspirado na maneira como o darwinismo explica o processo de evolução das espécies. Holland decompos o funcionamento dos algoritmos genéticos nas etapas de inicialização, avaliação, seleção, cruzamento, mutação, atualização e finalização.

1) Inicialização: Na maioria dos algoritmos a população inicial é gerada aleatoriamente. O tamanho da população é um dos parâmetros de um algoritmo genético. Assim como na natureza para haver evolução é necessário diversidade, é importante que a população inicial cubra a maior área possível do espaço de busca, sem interessar se são boas soluções ou não.

2) Avaliação: Podem ser utilizadas várias formas de avaliação. Cada individuo da população é avaliado, determinando assim o seu grau de adaptação. O valor de retorno da função mede a capacidade de resposta do individuo ao problema proposto. Este grau de adequabilidade também é denominado de *fitness*.

3) Seleção: A seleção é responsável pela perpetuação de boas caracteristicas na espécie. Nesta fase, os individuos são escolhidos para posterior cruzamento. Fazendo uso do grau de adaptação de cada um dos individuos é efectuado um sorteio onde os mais aptos possuem maior probabilidade de se reproduzirem.

4) Cruzamento: As caracteristicas das soluções escolhidas são recombinadas gerando novos indivíduos pela combinação das caracteristicas de dois individuos. Fragmentos das caracteristicas de um individuo são trocadas por um fragmento equivalemente oriundo de outro. O resultado desta operação é um individuo que combina caracteristicas potencialmente melhores que a dos seus pais.

5) Mutação: As caracteristicas dos individuos resultantes do processo de reprodução são alteradas acrescentando diversidade à população.

6) Atualização: Os individuos criados nesta geração são inseridos na população.

7) Finalização: Verifica se as condições para o fim do processo evolutivo foram atingidas. Podem ser utilizados vários critérios de paragem. Desde o numero de gerações já criadas até ao grau de convergência da população atual.

Algoritmo Genético na Análise Automática de Sinais Cardiotocográficos

Como visto anteriormente neste relatório, os algoritmos genéticos são frequentemente utilizados em casos de otimização. De forma a maximizar possiveis beneficios, devem ser removidas carateristicas irrelevantes ao processo de classificação. Fazendo uso do principio de seleção natural, o algoritmo genético mede a adequação de uma resposta ao problema. Com base nesta afirmação, ao fim de N gerações o individuo mais adaptado será o que possui o conjunto de caracteristicas que melhora o processo de classificação.
Para o desenvolvimento do trabalho pratico, desenvolveu-se um algoritmo genético que inicialmente cria uma população de possíveis respostas ao problema e as submete ao processo de evolução. Com base no algoritmo de Holland, definem-se os seguintes parámetros para a seleção de carateristicas do individuo:

- Inicialização - A população é iniciada com um conjunto de soluções candidatas codificadas em *array's* binários de tamanho fixo, cromossomas. O tamanho da população é definida como parámetro inserido pelo utilizador e influencia diretamente a eficácia do algoritmo. Populações constituidas por um numero demasiado pequeno de individuos convergem rápido e prematuramente. Populações com um numero elevado de individuos incrementa exponencialmente o tempo de execução do algoritmo, sem que se ganhem resultados significativos no processo de convergência.

- - Avaliação - É dificil calcular com exatidao completa o grau de adequação dos individuos. Pode ser uma tarefa complexa que é repetida ao longo do processo de evolução. Acarreta consigo um elevado custo computacional.

  A cada individuo corresponde um determinado codigo genético constituido por diferentes alelos com valores de 0's e 1's. Este código geńetico é passado ao algoritmo de classificação e no caso de o valor do alelo ser 1, a caracteristica do individuo é utilizada no processo de classificação. Caso contrário, a caracteristica não é utilizada. O valor de retorno dos algoritmos de classificação*KNN* e *Naive-Bayes* corresponde ao fitness do individuo.

- Seleção - Após a avaliação, cada individuo tem associado um valor numérico que traduz o seu nivel de adaptação ás restrições do problema. É imporante permitir que individuos pouco adaptados tenham probabilidade, embora reduzida, de ser selecionados.
  Para atingir esse fim, é escolhido o método de seleção por torneio, pelo qual um sub-conjunto da população com N individuos é sorteado e os melhores individuos dese grupo sao selecionados para decidir qual irá reproduzir. Utiliza-se o torneio de dois individuos(obtidos aleatoriamente) que conpetem entre si e o vencedor torna-se um dos pais.

Uma propriedade importante da seleção por torneio e o motivo pel qual levou á escolha deste algoritmo é que este não depende de um conhecimento global da população. Além disso não leva em consideração o *rank* que o individuo ocupa na população, permitindo uma seleção com menos tendencias.

- Cruzamento - Uma vez selecionados dois individuos, estes passam o seu material genético a uma próxima geração de filhos.Foram implementados dois tipos de recombinação. Um desses tipos é a recombinação de um ponto. Nesta operação seleciona-se aleatoriamente um ponto de corte nos cromossomas, dividindo-o em duas partições. Cada um dos filhos é composto pela junção de uma das partições dos seus pais. Outro tipo de recombinação utilizada é a de dois pontos. Neste caso, o filho 1 recebe a partição central do pai 2 e as partições á esquerda e á direita do pai 1. No caso do filho 2, este recebe a partição central do pai 1 e as partições á esquerda e á direita do pai 2.

- Mutação - O operador de mutação modifica aleatoriamente todos os cromossomas de um individuo. A probabilidade de ocorrência de uma mutação é denominada taxa de mutação. Foi atribuido um valor de 0.001 a essa taxa de mutação uma vez que esta operação pode gerar um individuo pior que o original.

- Atualização - Nesta fase os individuos resultantes do processo de cruzamento e mutação são inseridos na população. A população mantém um tamanho fixo e os individuos são criados em mesmo numero que os antecessores, substituindo-os.
  É adicionado a este conjunto a melhor solução da geração anterior aumentado assim o desempenho do algoritmo genético. Á sobrevivencia dos melhores N individuos numa população é dado o nome de elitismo. O elitismo total no entanto pode diminuir significativamente a diversidade dos individuos, podendo estagnar em locais ótimos e/ou aumentar o tempo de convergencia do algoritmo.

- Finalização - A finalização é composta por um teste que dá término ao processo de evolução. Caso o algoritmo genético tenha chegado ás 100 gerações é atingido o critério de paragem.

III. Resultados Obtidos do Algoritmo Genético

Nesta secção são apresentados conjuntos de resultados comparativos entre dois algoritmos incluidos no WEKA e que nao foram discutidos neste relatório. Os algoritmos escolhidos para comparação foram o K* e o Bayes-Net. O critério comparativo escolhido foi a precisão avaliativa de cada um dos algoritmos. O K* é um classificador baseado em exemplos, isto é, baseia-se na classe das instância de formação semelhante, conforme é determinado por uma função de similidaridade. Este difere de outros algoritmospor utilizar funções de distância na entropia e assume que os exemplos similares terão classes similares. A rede Bayesiana é um modelo probabilístico que representa um conjunto de variáveis aleatórias e as suas dependências condicionais através de um grafo dirigido acíclico (DAG). Os algoritmos devem considerar somente as variáveis mais relevantes para a tomada de decisão, isto é, deve delimitar ao máximo o número de variáveis sem perder informações. Para alcançar esse objectivo a solução obtida pelo algoritmo genético após N gerações foi aplicada em ambos os casos. Os algoritmos têm em consideração o código genético do indíviduo mais adaptado da nossa população. Isto é, utilizam unicamente as características em que o alelo tem valor 1. Foram comparados os resultados obtidos com a totalidade das caracteristicas e com o exemplo do código genético seguinte, 01110111101001111000O.

Na tabela são apresentados os resultados obtidos pelo algoritmo *K** com os parâmetros padrão estabelecidos pelo *WEKA*.

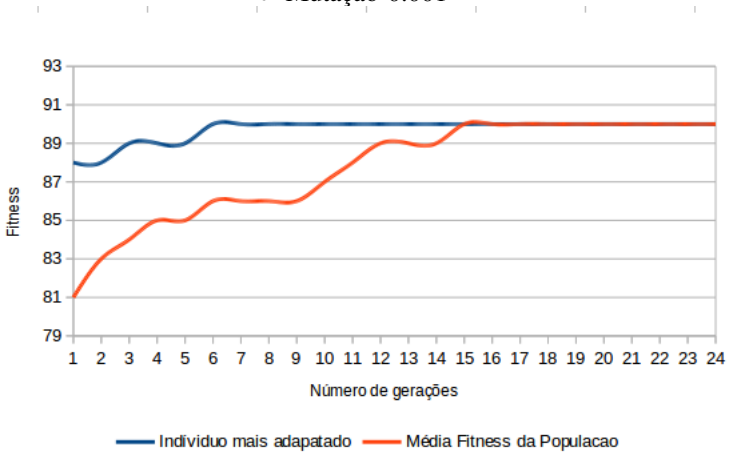
Qtd. Caracteristicas	% de Acerto	% de Erro
21	87.77%	12.23%
12	91.76%	8.24%

Na tabela são apresentados os resultados obtidos pelo algoritmo *BayesNet* com os parâmetros padrão estabelecidos pelo *WEKA*.

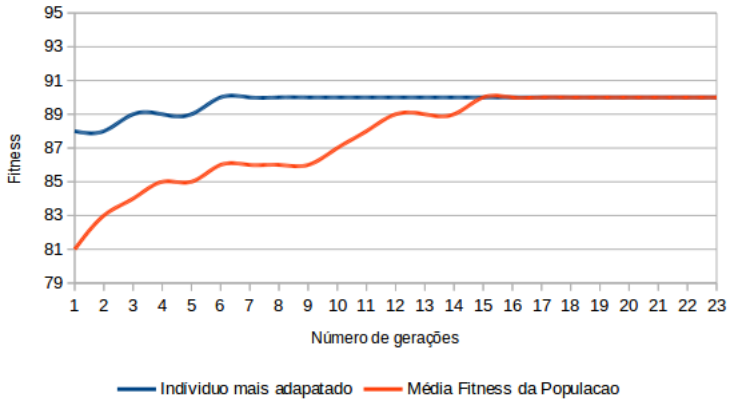
Qtd. Caracteristicas	% de Acerto	% de Erro
21	93.87%	6.13%
12	96.67%	3.33%

Gráficos do fitness da população para o KNN

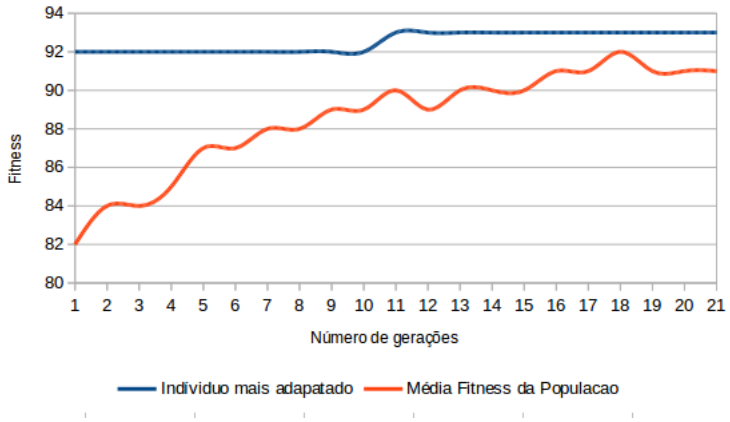
- Mutação 0.001



Mutação 0.01

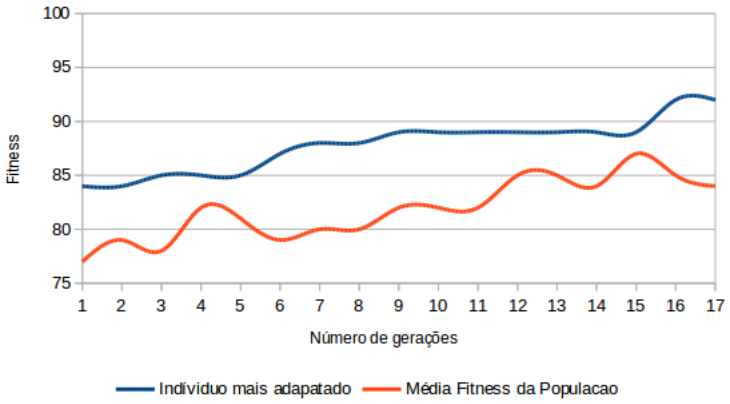


Mutação 0.1

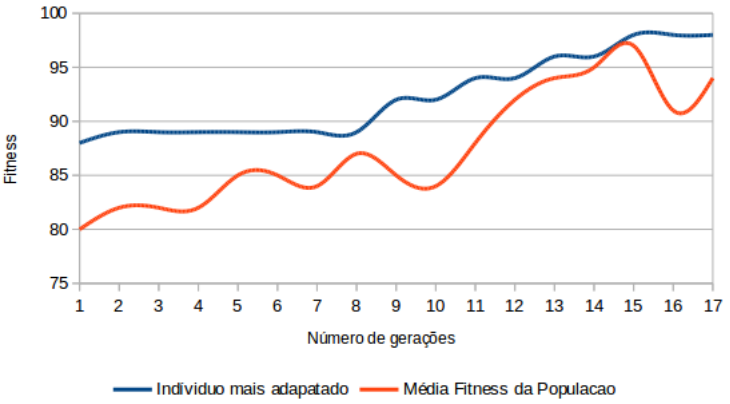


Gráficos do Fitness da População para o Naive-Bayes

- Mutação 0.001



Mutação 0.01



Comparação entre Classificadores

	KNN	Naive-Bayes
Sem AG	87.77%	93.87%
Com AG	91, 76%	96, 67%

IV. Reflexão Critíca e Problemas encontrados

Em relação ao trabalho desenvolvido durante o semestre, compreendemos que implementar algoritmos de classificação pode aumentar bastante a tomada de decisão em situações reais. No caso do tema do trabalho desenvolvido, uma boa percentagem de acerto significa um bom reconhecimento do sinal cardiotocográficos e consequentemente um bom diagnóstico. Relativamente ao projeto, foram implementados dois algoritmos de classificação. Embora os resultados obtidos fossem satisfatórios, verificou-se a presença de caracteristicas irrelevantes no processo de classificação. Com o objetivo de conhecer o conjunto de caracteristicas que optimizam o resultado de classificação foi implementado um algoritmo genético capaz de nos apresentar a melhor solução. Existiram alguns precalços na realização do trabalho, nomeadamente na implementação do *Naive-Bayes* no cálculo da probabilidade *likelihood* . Durante o desenvolvimento do algoritmo genético existiram opiniões divergentes no que consta á atualização da geração. Inicialmente foi definido que ambas as gerações de pais e filhos competiam pelos recursos existentes. Posteriormente foi decidido implementar a forma mais tradicional do algoritmo genético simples.

V. Conclusões e Trabalho Futuro

Concluimos que este foi um trabalho produtivo e rentável em termos de aquisição de conhecimentos sobre o funcionamento dos mecanismos de aprendizagem automática. Neste trabalho, os algoritmos de classificação nomeadamente o Naive-Bayes e o K-NN são usados para a classificação automática de sinais cardiotocográficos. Pela análise dos resultados obtidos observa-se que o KNN tem consideravelmente melhores resultados. Conclui-se que o KNN embora seja um algoritmo simples não paramétrico de reconhecimento de padrões está ao nivel de aplicações mais complexas. Considerando o processo de tomada de decisão na prática clínica médica , apenas os sintomas mais importantes são considerados necessários para a tomada de decisão sobre um determinado diagnóstico. Ou seja, o médico não elege um diagnóstico baseado em muitos sintomas, mas sim, a partir dos sintomas mais significativos para a tomada de decisão. Com o intuito de tornar esse processo automático foi implementado o algoritmo genético. Desse processo resultam uma série de respostas que representam com grande fidelidade as características mais relevantes para a análise automática de sinais cardiotocográficos. Como apresentado nos resultados experimentais o uso do algoritmo genético permitiu que se alcançassem resultados com uma melhoria bastante significativa. Como trabalho futuro pretende-se desenvolver um classificador baseado em regras difusas, tentando assim compreender qual dos dois algoritmos genéticos possui um fator de optimização mais alargado.