Questão 01 - No artigo, foi escolhido os principais atributos ou tuplas a serem minerados pela ferramenta, os atributos selecionados são os seguintes: Tipo_Óbito, Data_Nascimento, Data_Óbito, Local_Óbito, Ao término da mineração utilizando o algoritmo J48 observou-se que o mesmo
teve 88.3033 % (por cento) de certeza com 26491 (vinte e seis mil, quatrocentos e noventa e um) registros processados e 11.6967 % (por cento) O Gain Ratio tende a selecionar atributos com maior
possibilidades de valores, mesmo que estes não sejam os mais relevantes. no arquivo, 26491 (vinte e seis mil,
quatrocentos e noventa e um) foram minerados como de RACA_COR Branca, 1406 (mil e quatrocentos e seis) registros de cor Preta foram classificados como Branca, 2015 (dois mil e quinze) registros de cor de incerteza na sua
mineração com 3509 (três mil, quinhentos e nove) registro processado, sendo que a base de dados de 30000 (Trinta mil) algumas características onde ele adquire Verificando, portanto que o algoritmo Assistencia_Médica, Estado_Civil,
Raça_Cor e sexo sendo que cada um possui deferentes valores na sua composição, e a base de dados confusão
verificou-se que dos registros minerou corretamente
somente os que pertenciam ao grupo de RACA_COR Branca e os demais o mesmo classificou de forma errada colocando os registros presentes para as RACA_COR Preta, Parda, Amarela e Indígena todas como se fossem Branca. contém o valor Amarela classificados como Branca e 60 (sessenta) de cor Indígena classificados
também como Branca. registros armazenados. registros 30000 (trinta mil) resultar (SplitInfo). de suficiente
para a construção de uma árvore de decisão a partir de um banco de dados podendo assim possibilitar a tomada de decisão. Assim um seletor usado para a construção e análise das informações passadas ao mesmo é o information gain ratio ou taxa de ganho de informação, esta taxa é caracterizada pelo uso de uma métrica para ranquear todos os atributos de uma base de dados, a mesma é calculada utilizando o ganho de informação (Gain) de um atributo.

Questão 02 -

O problema do desequilíbrio de classes é frequentemente relatado como um obstáculo à indução de boas classificações por algoritmos de Machine Learning (ML). Em alguns domínios, por exemplo o conjunto Sickdata, os algoritmos padrão ML são capazes de induzir bons classificadores, mesmo utilizando conjuntos de treino altamente desequilibrados.
Desenvolvemos um estudo sistemático com o objectivo de questionar se os desequilíbrios de classe impedem a indução de classificadores ou se estas deficiências podem ser explicadas de outras formas. As nossas experiências sugeriram que o problema não é apenas causado por desequilíbrios de classe, mas está também relacionado com o grau de sobreposição de dados entre as classes.
Neste trabalho consideramos duas prob-lems de classe em que C1= + representa a classe de conceito circunscrita e C2= representa a contraparte desse conceito. O sistema de aprendizagem visa a construção de um modely=f(~x), de uma funçãof desconhecida, que permite prever os seus valores para exemplos anteriormente não vistos.
O algoritmo k-NN utiliza a função de distância HVDM para atributos quantitativos. A métrica VDM considera a similaridade de classificação para cada valor possível de um atributo qualitativo. Implementamos uma estrutura de indexação nomeadamenteM-tree para acelerar a execução das consultas k-NN. Neste trabalho, avaliamos dez métodos diferentes de sub-amostragem e sobreamostragem para equilibrar a distribuição de classes nos dados de formação. Dois destes métodos são métodos não heurísticos que foram inicialmente incluídos nesta avaliação como métodos de base. Os restantes métodos de equilíbrio utilizam a heurística para ultrapassar as limitações dos metanods não heurísticos.
Hart's Condensed Nearest Neighbor Rule (CNN) é utilizado para encontrar um subconjunto de exemplos con-sistentes. Um subsetˆE⊆Eis consis-tente comE. se usar um vizinho 1-nearest, ˚Ecorrectlyclassifies os exemplos emE.
A Neighborhood Cleaning Rule (ENN) usa a Edited Nearest Neigh-bor Rule de Wilson para remover exemplos de classe maioritária. A NCL modifica a ENN a fim de aumentar a limpeza de dados. SmoteSynthetic Minority Over-sampling Technique (SmoteSynthetic Minority Over-sampling Technique (SmoteSynthetic Minority Over-sampling Technique) (SmoteSynthetic

Minority Over-sampling Technique) (SmoteSynthetic Minority Over-sampling Technique) evita o problema de sobreposição.
O principal objectivo da nossa investigação é comparar vários métodos de balanceamento publicados na literatura, bem como os três métodos propostos. Seleccionámos conjuntos de dados da UCI hich
têm diferentes graus de desequilíbrio. Os resultados obtidos parecem ser compatíveis com os anteriores
trabalho.
Para conjuntos de dados de maior dimensão, o efeito destes factores complicadores parece ser reduzido, a partir do momento em que
A classe minoritária é melhor representada por um maior número de exemplos.
Os CUA foram medidos sobre árvores de decisão podadas com o parâmetro de poda C4,5 por defeito.
(nível de confiança de 25%) e sobre árvores sem poda de decisão. Os resultados de desempenho são relatados em termos de CUA, bem como de CUA com 10 vezes a validação cruzada.
Os resultados mostram que os métodos de amostragem excessiva em geral, e Smote + Tomek e Tomek + ENN (dois dos métodos propostos neste trabalho) forneceram resultados muito bons na prática. Para conjuntos de dados com maior número de ex-amostragens positivas, o método de sobreamostragem aleatória produziria resultados significativos.

Questão 03 -

Questão 04 -

Bagging: Bootstrap aggregating, também chamado de ensacamento (a partir de bootstrap aggregating), é um meta-algoritmo do conjunto de aprendizagem de máquinas concebido para melhorar a estabilidade e precisão dos algoritmos de aprendizagem de máquinas utilizados na classificação estatística e na regressão.
Embora seja normalmente aplicado a métodos de árvore de decisão, pode ser usado com qualquer tipo de método.

Do tamanho n, o ensacamento gera m novos conjuntos de treino Di D_{i}, cada um do tamanho n′, por amostragem a partir de D uniformemente e com substituição.
Se n′=n, então para grandes n o conjunto Di D_{i} deverá ter a fracção (1 - 1/e) (≈63.2%) dos exemplos únicos de D, sendo o restante duplicado. Este tipo de amostra é conhecido como uma amostra de bootstrap. A amostragem com substituição assegura que cada bootstrap é independente dos seus pares, uma vez que não depende de amostras previamente escolhidas no momento da amostragem.

O ensacamento leva a "melhorias para procedimentos instáveis", que incluem, por exemplo, redes neurais artificiais, árvores de classificação e regressão, e selecção de subconjuntos em regressão linear.

Boosting: A resposta afirmativa de Robert Schapire num artigo de 1990 à questão de Kearns e Valiant teve ramificações significativas na aprendizagem e estatística de máquinas, conduzindo sobretudo ao desenvolvimento do boosting. boosting é um meta-algoritmo de conjunto para reduzir principalmente o bias, e também a variação na aprendizagem supervisionada, e uma família de algoritmos de aprendizagem de máquinas que convertem os alunos fracos em alunos fortes. para os detectores de partilha de características, observa-se uma escala aproximadamente logarítmica com o número de aulas, ou seja, um crescimento mais lento do que linear no caso de não partilha. Os originais, propostos por Robert Schapire (uma formulação recursiva maioritária) e Yoav Freund (impulso por maioria), não eram adaptáveis e não podiam tirar o máximo proveito dos alunos fracos. Outros algoritmos que são semelhantes em espírito [clarificação necessária] para impulsionar algoritmos são por vezes chamados "algoritmos de alavanca", embora por vezes também sejam incorrectamente chamados algoritmos de impulsionamento. O fluxo principal do algoritmo é semelhante ao caso binário. Quando foi introduzido pela primeira vez, o problema de impulsionamento de hipóteses referia-se simplesmente ao processo de transformar um aprendiz fraco num aprendiz forte. "Informalmente, o problema [do reforço de hipóteses] pergunta se um algoritmo de aprendizagem eficiente que produz uma hipótese cujo desempenho é apenas ligeiramente melhor do que a adivinhação aleatória [isto é, um aprendiz fraco] implica as categorias que não têm a característica do classificador. O AdaBoost é muito popular e o mais significativo historicamente, pois foi o

primeiro algoritmo que se pôde adaptar Quando são adicionados, são ponderados de uma forma que está relacionada com a precisão dos alunos fracos. o algoritmo escolhe um classificador de uma única característica (características que podem ser partilhadas por mais categorias devem ser encorajadas). a existência de um algoritmo eficiente que produz uma hipótese de precisão arbitrária [i.e. um aluno forte]. A principal variação entre muitos algoritmos de dinamização é o seu método de ponderação de dados de formação aponta a base da cobertura introdutória de dinamização em cursos universitários de aprendizagem de máquinas. e mostrou que quando os dados de formação são limitados, a aprendizagem através da partilha de características faz um trabalho muito melhor do que a não partilha, dadas as mesmas rondas de dinamização. Embora a dinamização não seja limitada por algoritmos, a maioria dos algoritmos de dinamização consiste na aprendizagem iterativa de classificadores fracos Um aluno fraco é definido como sendo um classificador que está apenas ligeiramente correlacionado com Schapire e Freund e depois desenvolveu o AdaBoost, um algoritmo de dinamização adaptável que ganhou o prestigioso Prémio Gödel. o custo do tempo de execução do classificador) Assim, os futuros alunos fracos concentram-se mais nos exemplos que os alunos fracos anteriores classificaram erroneamente. a formulação de aprendizagem provavelmente aproximadamente correcta pode ser chamada algoritmos de dinamização. O arco de Freund e Schapire (Adapt[at]ive Resampling and Combining), como técnica geral, é mais ou menos sinónimo de boosting. Também, para um determinado nível de desempenho, o número total de características necessárias (e, por conseguinte, em contraste, um aprendente forte é um classificador que é arbitrariamente bem relacionado com algoritmos que são algoritmos de reforço prováveis em " Algoritmos que atingem o reforço de hipóteses rapidamente se tornaram simplesmente conhecidos como "reforço". o papel "Aprendizagem incremental de detectores de objectos usando um alfabeto de forma visual O que é diferente é que uma medida de Em comparação com a formação em separado, ela generaliza melhor, necessita de menos dados de formação, e requer menos características para atingir Durante a aprendizagem, os detectores para cada categoria podem ser treinados em conjunto. Dados de entrada mal classificados ganham um peso superior e exemplos que são classificados correctamente perdem peso. o que mostra que o aumento realiza uma descida de gradiente num espaço funcional utilizando uma função de custo convexo. ", no entanto os autores utilizaram o AdaBoost para o aumento de peso.
Em comparação com a categorização binária, a categorização multiclasse procura características comuns que possam ser partilhadas entre "Pode um conjunto de alunos fracos criar um único aluno forte? os pesos dos dados são reajustados, conhecidos como "re-pesos". os aprendentes fracos. Isto pode ser feito através da conversão da classificação multiclasse numa classificação binária (um conjunto de categorias versus Há muitos algoritmos mais recentes tais como LPBoost, TotalBoost, BrownBoost, xgboost, MadaBoost, LogitBoost, e outros. No jornal "
Há muitos algoritmos de impulso. o mesmo desempenho. o mesmo tempo.

Random Forest: Para tarefas de classificação, a saída da floresta aleatória é a classe seleccionada pela maioria das árvores. os requisitos para servir como um procedimento de exploração de dados fora da prateleira", dizem Hastie et al., "porque é invariante sob escala e várias outras transformações de valores de características, é robusta à inclusão de características irrelevantes, e produz modelos inspeccionáveis. se uma ou algumas características forem preditoras muito fortes para a variável de resposta (saída alvo), estas características serão seleccionadas em muitas das árvores B, fazendo com que se tornem correlacionadas.

Florestas aleatórias ou florestas de decisão aleatória são um método de aprendizagem em conjunto para classificação, regressão e outras tarefas que funciona através da construção de uma multiplicidade de árvores de decisão no momento da formação. A razão para o fazer é a correlação das árvores numa amostra comum de bootstrap: Esta interpretabilidade é uma das qualidades mais desejáveis das árvores de decisão. elas utilizam um algoritmo modificado de aprendizagem de árvores que selecciona, em cada candidato dividido no processo de aprendizagem, um subconjunto aleatório de Florestas aleatórias são uma forma de calcular a média de múltiplas árvores de decisão profundas, treinadas em diferentes partes do caminho que uma árvore de decisão toma para tomar a sua decisão é bastante trivial, mas seguindo os atributos preditivos estão linearmente correlacionados com a variável alvo, a utilização de florestas aleatórias pode não melhorar as árvores de decisão estão entre uma família bastante pequena de modelos de aprendizagem de máquinas que são facilmente interpretáveis ao longo da precisão do aprendiz base. a exactidão do conjunto do aprendente base.:587-588 As florestas aleatórias geralmente superam as árvores de decisão, mas a sua exactidão é inferior ao gradiente das árvores impulsionadas. As florestas de decisão aleatória corrigem o hábito das árvores de decisão de sobreajustar os dados e permitem que os utilizadores finais tenham confiança na

interpretabilidade intrínseca presente nas árvores de decisão. Para tarefas de regressão, a previsão média ou média do mesmo conjunto de formação, com o objectivo de reduzir as árvores individuais, é devolvida. ou centenas de árvores é muito mais difícil. As árvores de decisão são um método popular para várias tarefas de aprendizagem de máquinas. o algoritmo original de ensacamento para árvores. os caminhos das dezenas Embora as florestas aleatórias alcancem frequentemente maior precisão do que uma única árvore de decisão, sacrificam-se Para alcançar tanto o desempenho como a interpretabilidade, algumas técnicas de compressão de modelos permitem transformar uma floresta aleatória numa função de decisão mínima "nascida de novo". Em particular, as árvores que são cultivadas muito profundas têm tendência a incluir também outro tipo de esquema de ensacamento.

Para tarefas de classificação, a saída da floresta aleatória é a classe seleccionada pela maioria das árvores. os requisitos para servir como um procedimento de mineração de dados fora da prateleira", dizem Hastie et al., "porque é invariante sob escala e várias outras transformações de valor de características, é robusta à inclusão de características irrelevantes, e produz modelos inspeccionáveis. se uma ou algumas características forem preditoras muito fortes para a variável de resposta (saída alvo), estas características serão seleccionadas em muitas das árvores B, fazendo com que se tornem correlacionadas.

Florestas aleatórias ou florestas de decisão aleatória são um método de aprendizagem em conjunto para classificação, regressão e outras tarefas que funciona através da construção de uma multiplicidade de árvores de decisão no momento da formação. A razão para o fazer é a correlação das árvores numa amostra comum de bootstrap: Esta interpretabilidade é uma das qualidades mais desejáveis das árvores de decisão. elas utilizam um algoritmo modificado de aprendizagem de árvores que selecciona, de cada candidato dividido no processo de aprendizagem, um subconjunto aleatório de Florestas aleatórias são uma forma de calcular a média de múltiplas árvores de decisão profundas treinadas em diferentes partes do caminho que uma árvore de decisão toma para tomar a sua decisão é bastante trivial, mas seguindo os atributos preditivos estão linearmente correlacionados com a variável alvo, o uso de florestas aleatórias pode não melhorar as árvores de decisão estão entre uma família bastante pequena de modelos de aprendizagem de máquinas que são facilmente interpretáveis sobre a precisão do aprendiz base. a precisão do conjunto do aprendente de base.:587-588 As florestas aleatórias geralmente superam as árvores de decisão, mas a sua precisão é menor do que o gradiente das árvores impulsionadas. As florestas de decisão aleatória corrigem o hábito das árvores de decisão de sobreajustar os dados e permitem aos utilizadores finais ter confiança na interpretabilidade intrínseca presente nas árvores de decisão. Para tarefas de regressão, a previsão média ou média do mesmo conjunto de formação a fim de reduzir árvores individuais é devolvida. ou centenas de árvores é muito mais difícil. As árvores de decisão são um método popular para várias tarefas de aprendizagem de máquinas. o algoritmo original de ensacamento para árvores. os caminhos de dezenas Embora as florestas aleatórias alcancem frequentemente maior precisão do que uma única árvore de decisão, sacrificam-se Para alcançar tanto o desempenho como a interpretabilidade, algumas técnicas de compressão de modelos permitem que uma floresta aleatória seja transformada numa função de decisão mínima "nascida de novo". Em particular, árvores que são cultivadas muito profundamente tendem a incluir também outro tipo de esquema de ensacamento.

Há muitos algoritmos de impulso. Quando são adicionados, são ponderados de uma forma que está relacionada com a formulação de aprendizagem provavelmente aproximadamente correcta pode ser chamada de algoritmos de reforço. o documento "Aprendizagem incremental de detectores de objectos usando um alfabeto de forma visual Assim, os futuros alunos fracos concentram-se mais nos exemplos que os alunos fracos anteriores classificaram erradamente. O que é diferente é que uma medida do arco de Freund e Schapire (Adapt[at]ive Resampling and Combining), como técnica geral, é mais ou menos sinónimo de boosting. para alcançar Durante a aprendizagem, os detectores de cada categoria podem ser treinados conjuntamente. para métodos de árvore de decisão, pode ser usado com qualquer tipo de método. conjunto.:587-588 Florestas aleatórias geralmente superam as árvores de decisão, mas a sua precisão é inferior ao gradiente das árvores impulsionadas. ", que incluem, por exemplo, redes neurais artificiais, árvores de classificação e regressão, e selecção de subconjuntos em regressão linear. as árvores individuais são devolvidas. os caminhos de dezenas de florestas de decisão aleatória corrigem o hábito das árvores de decisão de sobreajustar o Bootstrap aggregating, também chamado ensacamento (a partir do bootstrap aggregating), é um meta-algoritmo do conjunto de aprendizagem de máquinas concebido para melhorar ou centenas de árvores é muito mais difícil. Os pesos de dados são reajustados, conhecidos como "re-pesagem".

Questão 05 -

Questão 06 -