Universidade Lusófona de Humanidades e Tecnologias
Licenciatura em Informática de Gestão

TRABALHO FINAL DE CURSO

Relatório

Versão final

07 Abril 2016

M. Clara José N.º 21200686 Turma 3N1

ÍNDICE

Contents

Introdução 7
O que se passa na área da Saúde? 7
Porquê analisar dados 8
Descrição do Processo 9
Ferramentas 10
Caso de estudo – Insuficiência Renal Crónica 11
Definição do Problema 11
Recolha de dados 12
Tratamento dos dados 12
Estudo das variáveis 13
Redução do número de variáveis do modelo 17
Situação das varáveis após o pré-processamento dos dados 20
Construção dos Modelos 21
Partição de dados 21
Modelos preditivos 23
Árvores de Decisão 23
Regressão estatística 25
Comparação dos Modelos Criados 30
Conclusões 30
Referências 33
Referências do caso de estudo: 33
Anexo I – Quadrante Mágico Gartner 35
Anexo II – Estudo das variáveis em detalhe 36
Anexo III – Diagrama do caso do estudo 72

Agradecimentos

Este trabalho foi realizado com o contributo de várias pessoas, às quais quero expressar os meus agradecimentos:

- Professor Francesco Costigliola, orientador deste projeto, pelo todo apoio e disponibilidade para acompanhar a elaboração deste caso de estudo;
- Professor Sérgio Guerreiro, pelo apoio prestado na fase inicial e por toda a documentação fornecida;
- Dr.ª Manuela Lucas, da Direção Geral de Saúde, pela explicação sobre a utilização de modelos preditivos em medicina;
- Dr.ª Lourdes Tavares, do Hospital dos Lusíadas, pelo apoio na interpretação dos dados e dos resultados obtidos.

Introdução

Vivemos numa época sem precedentes. A rápida evolução tecnológica, que se intensificou nos últimos vinte anos, potenciou o desenvolvimento económico e social e abriu novas fronteiras e áreas de conhecimento. A utilização de sistemas de informação contribuiu para uma progressiva digitalização da sociedade que, impulsionada pela facilidade de acesso à tecnologia e pela existência de uma rede global de comunicações, é considerada hoje uma força transformadora cujo verdadeiro alcance ainda não é possível determinar.

A par da evolução tecnológica, a evolução das ciências da vida, em geral, e da medicina, em particular, contribuiu para o aumento da esperança de vida nos países mais desenvolvidos ou em vias de desenvolvimento. Existem hoje meios de diagnóstico que permitem ver os mínimos detalhes do que se passa no interior do corpo humano. Os tratamentos são cada vez mais precisos e menos invasivos. A descodificação do genoma humano promete personalizar o tratamento de certas patologias, como nunca se pensou que fosse possível.

No entanto, num estudo de 2013 divulgado pela OMS, conclui-se que a diabetes, as doenças coronárias e os tumores malignos das mais variadas origens são as principais causas de mortes prematuras no mundo desenvolvido. Isto significa que apesar de vivermos vidas mais longas, não vivemos vidas mais saudáveis: factores como o sedentarismo, a ingestão de gorduras saturadas, o tabagismo, os ambientes poluídos, o excesso de açúcar e sal que ingerimos diariamente contribuem para o aumento dessas doenças e representam elevados custos de saúde pública.

O objectivo deste trabalho é mostrar como se pode alinhar a evolução tecnológica aos desafios com que a medicina se defronta: pretende-se demonstrar que os modelos preditivos (largamente usados em diversas áreas da medicina) dispõem, nos dias de hoje, de ferramentas de análise poderosas para tratar grandes quantidades de dados. Essa análise poderá expor factores de risco desconhecidos ou relações não evidentes entre a doença, as características do paciente e as terapêuticas aplicadas. Novos conhecimentos e *insights* podem ser obtidos e aplicados no tratamento e na prevenção de certas doenças o que, a médio prazo, terá um impacto significativo e no aumento da esperança de vida dos pacientes.

O que se passa na área da Saúde?

Em meados dos anos 90 do século passado iniciou-se a informatização da área da Saúde, numa perspectiva administrativa: os dados pessoais dos pacientes passaram a ser registados informaticamente, os pedidos de exames e análises clínicas também e as consultas passaram a ser efectuadas com base numa agenda electrónica. No entanto, nenhuma destas alterações teve impacto direto nem no diagnóstico nem no tratamento dos doentes. Nessa altura, os objectivos da informatização na Saúde eram semelhantes aos de qualquer outra área: automatizar tarefas repetitivas, reduzir tempos de espera, aumentar a eficiência operacional administrativa e capacitar os sistemas para faturar consultas e exames de forma mais rápida e eficaz.

Progressivamente a informatização na saúde foi-se alargando: os dados clínicos dos pacientes começaram a ser registados electronicamente, bem como os resultados dos exames médicos. Em meados da primeira década dos anos 2000, os hospitais públicos portugueses avançaram de forma concertada para a implementação do registo electrónico dos dados médicos dos pacientes. As fichas

dos doentes, os resultados dos exames e as notas registadas pelos médicos nas consultas passaram a estar em formato electrónico e acessíveis aos profissionais de saúde que deles necessitassem.

No entanto, dezasseis anos depois, o historial médico continua a estar disperso e não existe uma visão única sobre o paciente. O tratamento dos doentes continua a focar-se no imediato: o objectivo é tratar as condições e sintomas existentes. O registo centralizado dos dados médicos está longe de estar concretizado. Esta visão fragmentada condiciona o tratamento e a prevenção das doenças. No entanto, o conjunto de dados clínicos atualmente disponíveis em formato digital cresceu exponencialmente e não pára de aumentar, quer em quantidade quer em variedade. O potencial é enorme e o desafio também: pretende-se que esses dados sejam usados para diagnosticar de forma mais eficaz as patologias, identificar factores que contribuem para o seu aparecimento e adequar a aplicação das terapêuticas tendo por base as características genéticas e metabólicas dos pacientes.

Porquê analisar dados

Enormes quantidades de dados das mais variadas origens e formatos são recolhidas e armazenadas todos os dias. Uma estimativa recente refere que a quantidade de dados em circulação duplica a cada 14 meses e a tendência é que este aumento exponencial continue por mais uns anos. Mas porque é que esses dados são importantes?

A resposta a esta questão é fácil: quanto mais dados temos mais e melhor informação podemos obter. Mas para isso há que analisar os dados de forma a identificar padrões e traçar tendências. É necessário reduzir um grande número de variáveis a modelos que possamos compreender e que permitam tomar decisões mais informadas e adequadas.

A área dos sistemas de informação dedicada à análise e tratamento de grandes quantidades de dados com o objectivo de extrair conhecimento é denominada *data mining*. Nos últimos anos esta área das ciências da informação tem ganho preponderância: a promessa de identificar padrões na enorme quantidade de dados disponível pode representar uma vantagem num Mundo cada vez mais competitivo. Trata-se de uma área de conhecimento multidisciplinar que envolve:

- Tecnologias de bases de dados, que providenciam formas de armazenamento e obtenção de grandes quantidades de dados;
- Estatística, que providencia modelos para análise e comportamento dos dados recolhidos;
- Inteligência artificial e aprendizagem máquina, que são áreas particularmente importantes no reconhecimento de padrões e classificação dos dados;
- Técnicas de visualização para apresentar de forma gráfica os padrões detectados na análise.

A análise de dados tem, então, dois objectivos principais:

1. Prever o comportamento de uma variável, em função dos valores de outras variáveis;
2. Classificar os dados, segundo determinados critérios.

Descrição do Processo

A análise de dados segue um processo bem definido que se equematiza na figura seguinte:

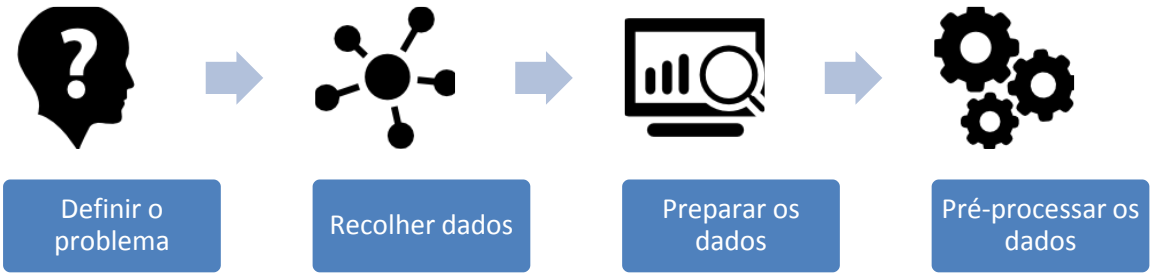


Figura 1 - Etapas do Processo de Data Mining

Etapa	Objectivo
Definir o problema	É a primeira atividade do processo de *data mining*. Consiste na identificação do problema que se pretende analisar. Pressupõe que sejam definidas as variáveis do problema (ou seja, os *inputs*) e qual o objectivo (o *output* ou *target* a atingir).
Recolher os dados	A recolha de dados deve obedecer a alguns critérios: • A amostra deve ser significativa (maior que 5000 registos); • Os dados recolhidos devem ser o mais aproximados possível da realidade; • Durante a recolha de dados deve ser garantido que não existem alterações estruturais no período de recolha.
Preparar os dados	Na maior parte dos *data sets* obtidos existem dados em falta ou dados que estão claramente errados. Este tipo de situações causa problemas aos modelos a desenvolver, pois podem enviesar seriamente os resultados. Assim, é necessário analisar as varáveis no sentido de verificar se existem ajustes a fazer ou se há variáveis que não podem ser usadas. É também nesta fase que são identificados os *outliers*, ou seja, as observações (corretas ou não) que representam casos extremos e que também podem enviesar os modelos criados.
Pré-processar os dados	Esta é uma das fases mais importantes de todo o processo. Nesta fase, é feita uma análise que identifica as variáveis que podem estar relacionadas entre si e de que forma. Com esta identificação é possível reduzir o número de variáveis em estudo, na medida em que se duas variáveis estão correlacionadas então, basta usar uma delas na construção do modelo. Dessa forma reduz-se a redundância e contribui-se para a precisão e qualidade do modelo. No fim desta fase, é expectável que os dados se encontrem normalizados e que as redundâncias estejam eliminadas.

Ferramentas

Existe hoje uma grande oferta de ferramentas com as quais se podem explorar e analisar os dados, com o intuito de obter conhecimento. Nos últimos anos, temos vindo a assistir à disponibilização de plataformas *open source* dedicadas à análise de dados com recurso a métodos quantitativos e a algoritmos específicos de *data mining*. Num estudo recente da Gartner foram analisadas 16 ferramentas de diferentes fabricantes, que permitem:

- Incorporar dados de diferentes fontes e formatos;
- Explorar, preparar e visualizar dados;
- Criar modelos preditivos e descritivos e integrá-los no contexto do negócio;
- Aplicar os modelos criados aos dados reais.

O resultado dessa análise está patente na figura seguinte:

Figura 2 – Análise de ferramentas de *data mining- Garter Inc.*

Neste estudo pode-se ver que os líderes de mercado indiscutíveis continuam a ser a SAS e a IBM. No quadrante dos líderes e a ganhar terreno estão alguns fabricantes com soluções *open source*, como é o caso do KNIME e do Rapid Miner (que também tem uma versão paga) e que têm vindo a ganhar preponderância e credibilidade.

Para o caso em estudo foi usada a ferramenta de **SAS® Enterprise Miner™**, que permite analisar problemas de modelação descritiva e preditiva, de forma visual e rápida e que é a ferramenta mais disseminada no mercado empresarial.

Para este trabalho usei a versão *online*, com a licença especial para estudantes, disponível na plataforma **SAS® On Demand for Academics**.

Caso de estudo – Insuficiência Renal Crónica

Definição do Problema

O caso de estudo que decidi apresentar está relacionado com a detecção precoce da insuficiência renal crónica tendo por base exames de rotina. A pergunta que coloquei foi: **É possível determinar a partir da observação dos pacientes e dos resultados obtidos de análises ao sangue e à urina quais os factores que apontam para uma possível insuficiência renal crónica?**

A insuficiência renal crónica é uma doença progressiva que afecta a função renal de forma irreversível. Esta doença surge, muitas vezes, associada a outras patologias, como por exemplo:

- Hipertensão arterial
- Diabetes Mellitus
- Doença poliquística renal (aparecimento quistos nos rins)
- Infecções do trato urinário superior com carácter repetitivo

Quando os pacientes apresentam alguma destas patologias é normal que sejam vigiados relativamente à insuficiência renal, mas o contrário já não é verdade. Sendo a insuficiência renal uma doença silenciosa, na maior parte das vezes instala-se sem que o paciente apresente quaisquer sinais ou sintomas. Isto significa que, normalmente, a doença só é diagnosticada quando a função renal já está comprometida de forma irreversível.

Trata-se de uma patologia com custos humanos e financeiros muito elevados. O único tratamento definitivo é o transplante renal, com todos os riscos que acarreta: os doentes transplantados têm de seguir terapêuticas rigorosas para evitar rejeições. Em tratamentos mais convencionais os pacientes são submetidos a diálise (normalmente várias vezes por semana) e têm de seguir uma dieta com poucas proteínas, uma vez que quando a insuficiência renal está instalada, os rins deixam de conseguir excretar os produtos tóxicos resultantes do metabolismo associado às proteínas. Muitas vezes têm também de tomar medicação para prevenir e tratar outras patologias associadas. Os pacientes mais graves têm bastantes limitações e os longos anos à espera de um transplante (que pode nunca vir a acontecer) têm efeitos psicológicos danosos que se refletem na saúde emocional dos pacientes. Todos estes aspectos reduzem bastante a qualidade de vida destes pacientes.

A detecção da insuficiência renal numa fase precoce é determinante para o atraso aa progressão da doença. Quando detetada numa fase inicial, a doença pode ser controlada durante bastante tempo apenas com uma dieta apropriada e um estilo de vida saudável. Muitas vezes, a doença pode até ser travada e apesar do doente ter de ser seguido por especialistas e ter de fazer exames periodicamente, poucos mais incómodos terá, durante um período de tempo considerável.

Recolha de dados

Para o caso em estudo, obtive dados públicos, disponíveis no repositório **UCI Machine Learning** relacionados com insuficiência renal crónica (Para detalhes sobre os dados, consultar a secção de Referências).

O conjunto de dados utilizados teve por base a recolha de dados quantitativos e qualitativos relativamente a um grupo de 400 pacientes em que uma percentagem apresenta insuficiência renal crónica. A recolha foi efetuada no Apollo Reach Hospital, em Karaikudi, uma pequena cidade no sudoeste da Índia. O *data set* foi previamente preparado e formatado por um estudante de pós-graduação de uma Universidade indiana para poder ser usado neste tipo de análises e foi disponibilizado no repositório público referido.

O número de observações é claramente abaixo do desejável para a construção de um modelo (como vimos anteriormente, a construção de modelos deve ter uma base com pelo menos 5000 observações), ainda assim, para o objectivo deste trabalho, a amostra é suficiente pois permite demonstrar a aplicabilidade das técnicas e o recurso a ferramentas especializadas na análise de dados.

Tratamento dos dados

Na importação dos dados foi acrescentado um identificador para os registos e cada uma das variáveis foi classificada de acordo com o seu contributo para a construção do modelo:

- ID →Identificador dos registos
- *Input* → Variáveis independentes
- *Target* → Variável dependente, neste caso corresponde à existência ou não de insuficiência renal crónica

Variável	Descrição	Tipo	Contributo para o Modelo
ID	Identificador de cada registo	Números inteiros sequenciais	Identificador
age	Idade do paciente	Numérica, contínua, em anos	*Input*
al	Albumina na urina	Categórica, com seis valores possíveis: 0, 1, 2, 3, 4, 5	*Input*
ane	Anemia	Binária (Sim, Não)	*Input*
appet	Apetite	Binária (Normal, Pouco)	*Input*
ba	Bactérias	Binária (Presentes, Ausentes)	*Input*
bgr	Glicemia	Numérica, contínua, em mg/dl	*Input*
bp	Pressão arterial	Numérica, contínua, em mm/Hg	*Input*
bu	Ureia no sangue	Numérica, contínua, em mg/dl	*Input*
cad	Doença coronária	Binária (Sim, Não)	*Input*
dm	Diabetes Mellitus	Binária (Sim, Não)	*Input*
has_cdk	Indicação se o paciente apresenta insuficiência renal	Binária (Sim, Não)	***Target***
hemo	Hemoglobina	Numérica, contínua, em gms	*Input*
htn	Hipertensão	Binária (sim, não)	*Input*
pc	Piúria (leucócitos na urina)	Binária (normais, anormais)	*Input*

Variável	Descrição	Tipo	Contributo para o Modelo
pcc	Aglomerados de leucócitos na urina	Binária (Presentes, Ausentes)	*Input*
pcv	Volume globular, (tamanho de glóbulos vermelhos)	Numérica, contínua, em percentagem	*Input*
pe	Edema dos membros inferiores	Binária (Sim, não)	*Input*
pot	Potássio	Numérica, contínua, em mEq/L	*Input*
rbc	Glóbulos vermelhos	Binária (normais, anormais)	*Input*
rbcc	Contagem de glóbulos vermelhos	Numérica, contínua, em milhões de células por cm^3	*Input*
sc	Creatinina	Numérica, contínua, em mg/dl	*Input*
sg	Densidade urinária	Categórica, com cinco valores possíveis: 1.005, 1.010, 1.015, 1.025	*Input*
sod	Sódio	Numérica, contínua, em mEq/L	*Input*
su	Açúcar na urina – Glicosúria	Categórica, com seis valores possíveis: 0, 1, 2, 3, 4, 5	*Input*
wbcc	Leucócitos na urina Contagem de glóbulos brancos	Numérica, contínua, em nº de células por cm^3	*Input*

Estudo das variáveis

Foi efetuada uma análise preliminar dos dados, no sentido de identificar as situações a tratar antes de se avançar com a construção dos modelos[1].

Tal como referido anteriormente, o tratamento das variáveis é uma das fases mais críticas do processo de construção de um modelo, pois condiciona fortemente o resultado. Neste caso em particular, dado que a amostra é muito pequena, este tratamento assumiu particular relevância, pois quaisquer inconsistências nos dados têm grande impacto nos modelos criados.

O objetivo desta fase foi:

- Identificar *outliers*, ou seja, casos que estão claramente fora dos padrões, quer se tratem de situações reais, quer se trate de eventuais erros;
- Identificar valores omissos e decidir como tratá-los;
- Identificar variáveis relacionadas entre si, com o intuito de validar se era possível reduzir o espaço de input, no sentido de criar um modelo mais fácil de interpretar e aplicar, mas sem comprometer os factores que podem contribuir para a detecção da patologia.

[1] O estudo detalhado das variáveis é apresentado no Anexo II.

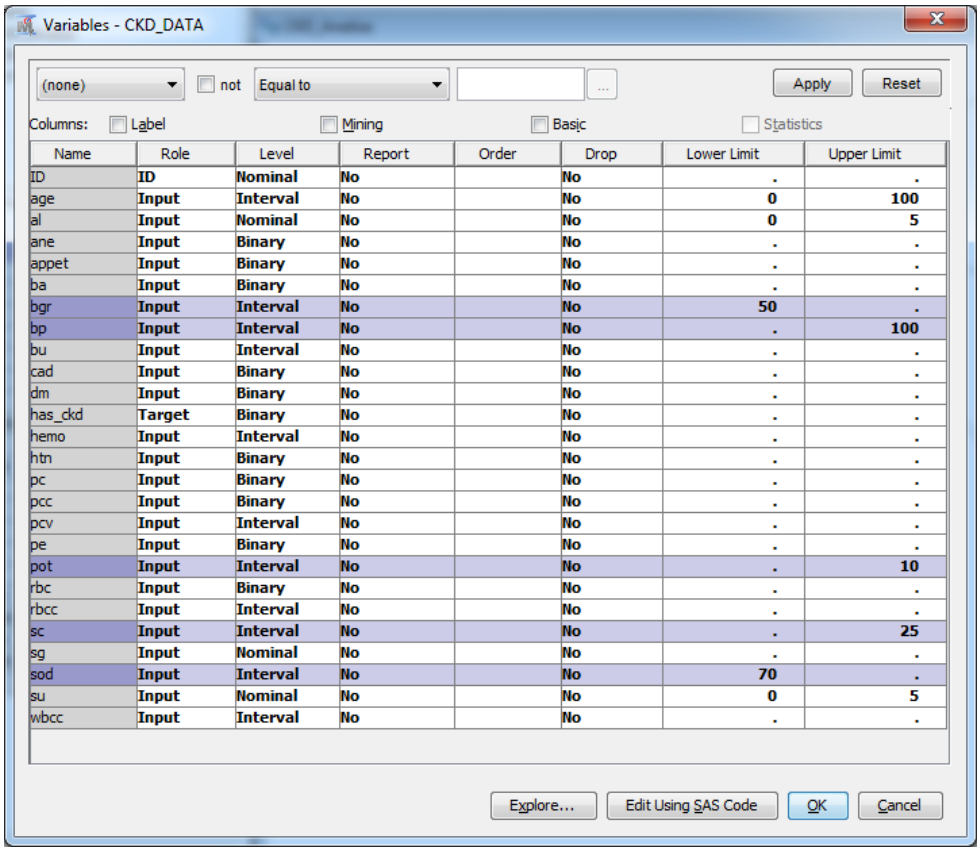
Outliers

Foram identificados alguns *outliers*, provavelmente relativos a situações de erro de registo. Os casos mais evidentes são os seguintes:

- Glicémia (**bgr**) – 1 paciente com um valor anormalmente baixo, na ordem dos 22 mg/dl;
- Tensão arterial (**bp**) – 6 pacientes com valores anormalmente elevados para tensão diastólica. Para estes casos estão registados valores acima de 100 mmHg, pelo que presumi que se trate de um erro de registo;
- Potássio (**pot**) – 2 pacientes com valores anormalmente elevados, provavelmente trata-se de um erro de registo.
- Creatinina (**sc**) – 4 pacientes com valor excessivos, acima dos 25 mg/dl
- Sódio (**sod**) – 1 paciente com valor anormalmente baixo, na ordem dos 4.5 mEq/l

No sentido de proceder ao tratamento destes casos, optei por impor limites mínimos e máximos às varáveis associadas aos *outliers*. Nestas situações, o **EMiner™** trata os *outliers* como casos omissos, ou seja sem valor para as variáveis respectivas. Isto significa que na construção do modelo vão ser aplicadas a estas observações as regras definidas para os casos omissos.

Assim, foram colocados os limites necessários, de acordo com a figura seguinte:



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
ID	ID	Nominal	No		No	.	.
age	Input	Interval	No		No	0	100
al	Input	Nominal	No		No	0	5
ane	Input	Binary	No		No	.	.
appet	Input	Binary	No		No	.	.
ba	Input	Binary	No		No	.	.
bgr	Input	Interval	No		No	50	.
bp	Input	Interval	No		No	.	100
bu	Input	Interval	No		No	.	.
cad	Input	Binary	No		No	.	.
dm	Input	Binary	No		No	.	.
has_ckd	Target	Binary	No		No	.	.
hemo	Input	Interval	No		No	.	.
htn	Input	Binary	No		No	.	.
pc	Input	Binary	No		No	.	.
pcc	Input	Binary	No		No	.	.
pcv	Input	Interval	No		No	.	.
pe	Input	Binary	No		No	.	.
pot	Input	Interval	No		No	.	10
rbc	Input	Binary	No		No	.	.
rbcc	Input	Interval	No		No	.	.
sc	Input	Interval	No		No	.	25
sg	Input	Nominal	No		No	.	.
sod	Input	Interval	No		No	70	.
su	Input	Nominal	No		No	0	5
wbcc	Input	Interval	No		No	.	.

Figura 3 – Outliers - Limites mínimos e máximos considerados

Tratamento dos valores omissos

Para a aplicação de técnicas de regressão não podem existir valores omissos. Assim, para esres casos podemos:

- Excluir do modelo os registos que apresentam valores omissos ou, em alternativa:
- Atribuir valores com base em determinados critérios e utilizar os registos na construção do modelo.

Com o intuito de criar um modelo o mais preciso possível e dado que a amostra é bastante reduzida, optei por não excluir nenhuns registos e tratar os valores omissos (para mais detalhes consultar o Anexo II – Estudo de variáveis em detalhe). No quadro seguinte é apresentado um resumo da análise e decisões tomadas para cada uma das variáveis de *input*:

Variável	Descrição	Valores Omissos	Imputação de valores
ID	Identificador de cada registo	Sem valores omissos.	Não aplicável.
age	Idade do paciente	9 pacientes dos quais não sabemos a idade	Foi aplicada a média de idades: ~52 anos
al	Albumina na urina	49 pacientes sem valor para esta variável	Substituído por 0, ou seja, negativo para a presença de albumina na urina
ane	Anemia	1 paciente para o qual não sabemos se apresenta anemia ou não	Substituído por 0, ou seja, negativo para a anemia
appet	Apetite	1 paciente para o qual não sabemos se apresenta alterações de apetite	Substituído por 0, ou seja, sem alterações de apetite
ba	Bactérias (na urina)	4 pacientes para os quais não sabemos se existe infecção bacteriana presente	Substituído por 0, ou seja, negativo para a presença de infecção
bgr	Glicémia (açúcar no sangue)	44 pacientes para os quais não sabemos o valor da glicemia	Deveria ser aplicada a média dos valores, ou seja 148 mg/dl, mas esta variável foi rejeitada.
bp	Pressão arterial	12 pacientes para os quais não temos informação da pressão arterial diastólica	Foi aplicada a média dos valores, ou seja, 77 mmHg
bu	Ureia no sangue	19 pacientes para os quais não temos informação sobre os valores de ureia no sangue	Foi aplicada a média dos valores, ou seja, 52,74 mm/dl.
cad	Doença coronária	2 pacientes para os quais não sabemos se apresentam doença coronária	Substituído por 0, ou seja, negativo para a presença de doença coronária.
dm	Diabetes Mellitus	3 pacientes para os quais não temos informação se apresentam diabetes ou não	Substituído por 0, ou seja, negativo para diabetes.

Variável	Descrição	Valores Omissos	Imputação de valores
has_cdk	Indicação se o paciente apresenta insuficiência renal	Sem valores omissos, variável *target*.	Não aplicável.
hemo	Hemoglobina	52 pacientes para os quais não sabemos os valores de hemoglobina	Deveria ter sido aplicada a média de 12,52 gms, mas a variável foi rejeitada.
htn	Hipertensão	2 pacientes para os quais não sabemos se apresentam hipertensão ou não	Substituído por 0, ou seja, negativo para hipertensão.
pc	Piúria (leucócitos na urina)	66 pacientes sem valor para esta varável	A análise dos dados e a comparação com a variável *target* não permitiu tirar conclusões sobre o valor a aplicar
pcc	Aglomerados de leucócitos na urina	4 pacientes sem valor para esta variável	Deveria ter sido substituído por 0, ou seja, negativo para a presença de leucócitos na urina, mas a variável foi rejeitada.
pcv	Volume globular	71 pacientes sem valor para esta variável	Foi aplicado o valor médio, ~39%
pe	Edema dos membros inferiores	1 paciente para o qual não temos informação se apresenta edema dos membros inferiores	Substituído por 0, ou seja, negativo para o edema dos membros inferiores.
pot	Potássio	88 pacientes para os quais não temos informação sobre os valores de potássio no sangue	Foi aplicado o valor médio, ou seja ~4,63 mEq/l.
rbc	Glóbulos vermelhos na urina	152 pacientes para os quais não sabemos se apresentam glóbulos vermelhos na urina	Deveria ter sido imputado o valor 1, ou seja, positivo para a presença de glóbulos vermelhos na urina, mas a variável foi rejeitada.
rbcc	Contagem de glóbulos vermelhos	131 pacientes para os quais não sabemos os valores de glóbulos vermelhos presentes no sangue	Foi aplicado o valor médio, ou seja, ~4,71 mEq/l
sc	Creatinina	17 pacientes para os quais não sabemos o valor da creatinina no sangue	Foi aplicado o valor médio, ou seja, ~3.07 mg/l.
sg	Densidade da urina	47 pacientes para os quais não sabemos o valor da densidade urinária	Foi aplicado o valor médio, ou seja: ~1.07.
sod	Sódio	87 pacientes para os quais não sabemos o valor de sódio no sangue	Foi aplicado o valor médio, ou seja, ~137.52 mEq/l.

Variável	Descrição	Valores Omissos	Imputação de valores
su	Açúcar na urina – Glicosúria	49 pacientes para os quais não sabemos o valor de açúcar na urina	Foi aplicado o valor médio: 0.45.
wbcc	Contagem de glóbulos brancos	106 pacientes para os quais não sabemos o valor dos glóbulos brancos.	Variável rejeitada devido ao facto de não se conseguir concluir que imputação de valores deve ser feita.

A imputação de valores foi feita considerando a média para variáveis contínuas (numéricas) e o valor 0 para as variáveis binárias.

Os resultados da execução da imputação podem ser vistos na figura seguinte:

Imputation Summary

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
age	MEAN	IMP_age	51.33948	INPUT	INTERVAL	age	7
al	CONSTANT	IMP_al	0	INPUT	ORDINAL	al	33
ane	CONSTANT	IMP_ane	0	INPUT	BINARY	ane	1
appet	CONSTANT	IMP_appet	0	INPUT	BINARY	appet	1
ba	CONSTANT	IMP_ba	0	INPUT	BINARY	ba	4
bp	MEAN	IMP_bp	76.33333	INPUT	INTERVAL	bp	8
bu	MEAN	IMP_bu	56.47586	INPUT	INTERVAL	bu	17
cad	CONSTANT	IMP_cad	0	INPUT	BINARY	cad	1
dm	CONSTANT	IMP_dm	.	INPUT	BINARY	dm	2
htn	CONSTANT	IMP_htn	0	INPUT	BINARY	htn	1
pc	CONSTANT	IMP_pc	0	INPUT	BINARY	pc	44
pcv	MEAN	IMP_pcv	38.63793	INPUT	INTERVAL	pcv	46
pe	CONSTANT	IMP_pe	0	INPUT	BINARY	pe	1
pot	MEAN	IMP_pot	4.521364	INPUT	INTERVAL	pot	58
rbcc	MEAN	IMP_rbcc	4.677157	INPUT	INTERVAL	rbcc	81
sc	MEAN	IMP_sc	3.197148	INPUT	INTERVAL	sc	15
sg	MEAN	IMP_sg	1.017247	INPUT	INTERVAL	sg	31
sod	MEAN	IMP_sod	137.1018	INPUT	INTERVAL	sod	57
su	MEAN	IMP_su	0.514286	INPUT	INTERVAL	su	33

Figura 4 – Resultado da imputação dos valores omissos

Redução do número de variáveis do modelo

Com o intuito de simplificar o modelo, analisei os dados no sentido de detectar variáveis correlacionadas. Esta análise foi baseada em conhecimento prévio sobre as relações que poderiam existir entre determinadas variáveis.

Comparação das variáveis que podem influenciar a anemia

No sentido de verificar as relações entre as variáveis comparei:

- Contagem de glóbulos vermelhos no sangue (**rbc**)
- Hemoglobina no sangue (**hemo**)
- Anemia (**ane**)
- Contagem de glóbulos vermelhos na urina (**rbcc**)

Pela análise, é possível verificar que as três primeiras variáveis estão relacionadas entre si: para valores baixos de hemoglobina e de glóbulos vermelhos, confirma-se a presença de anemia.

Já com a variável que indica a presença de glóbulos vermelhos na urina, não é possível estabelecer uma relação com a anemia.

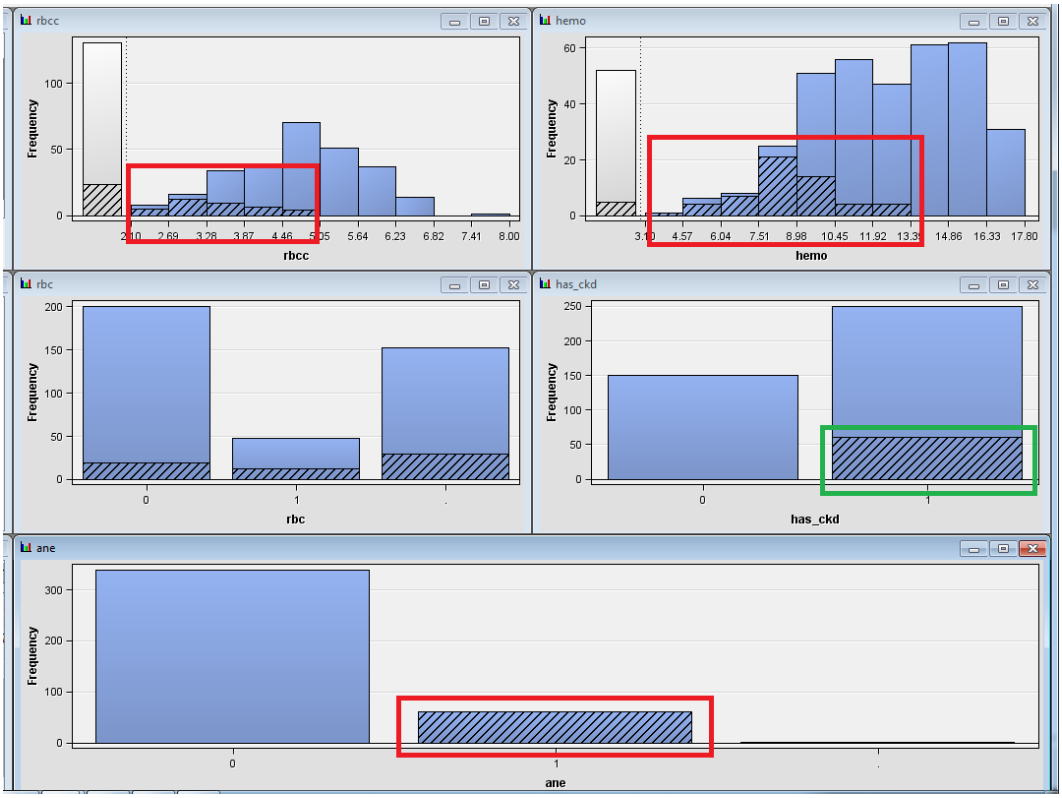


Figura 5 - Comparação das variáveis relacionadas com a anemia

Assim, decidi manter no modelo a variável relativa à presença de anemia nos pacientes e optei por rejeitar as variáveis **rbc** (contagem de glóbulos vermelhos) e **hemo** (hemoglobina).

Comparação das variáveis indicativas de diabetes

No sentido de reduzir o número de variáveis no modelo, comparei as variáveis relacionadas com a diabetes:

- Diabetes (**dm**)
- Glicemia (açúcar no sangue – **bgr**)
- Glicosúria (açúcar na urina - **su**)

Ao comparar estas três varáveis é possível determinar que a diabetes está relacionada com elevados níveis de glicemia, como se pode constatar na figura seguinte:

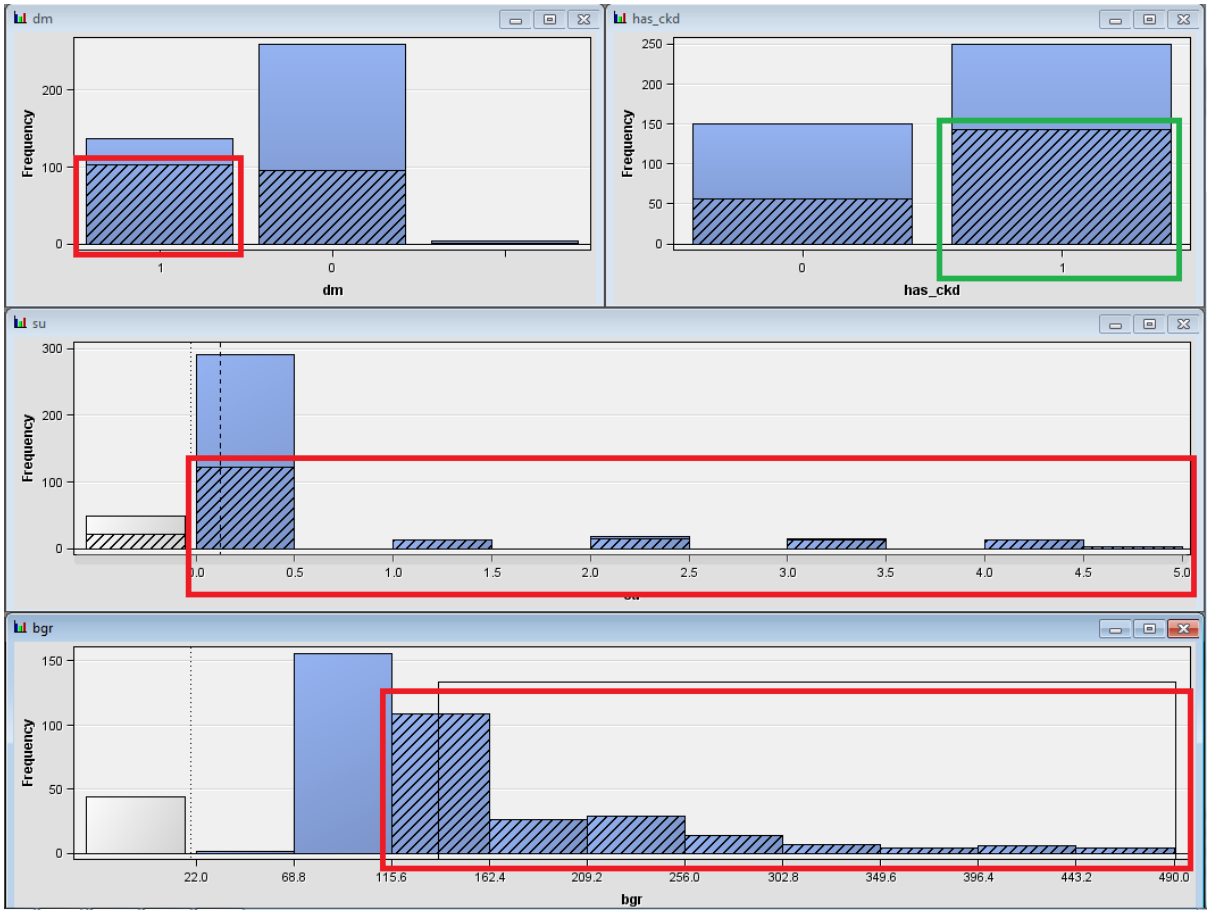


Figura 6 - Comparação das variáveis relacionadas com a diabetes

Já a presença de açúcar na urina não releva uma relação tão evidente com a diabetes, já que na maior parte dos casos, os valores apresentados estão dentro do normal.

Assim optei por rejeitar a variável relativa à glicemia (**bgr**) e manter as outras duas.

Comparação das variáveis indicativas de infecção urinária

As varáveis indicativas da presença de infecção urinária são:

- Bactérias na urina (**ba**)
- Piúria (**pc**)
- Aglomerados de leucócitos (**pcc**)

Ao comparar estas variáveis não há uma relação evidente entre as três, como se pode constatar nas figuras seguintes:

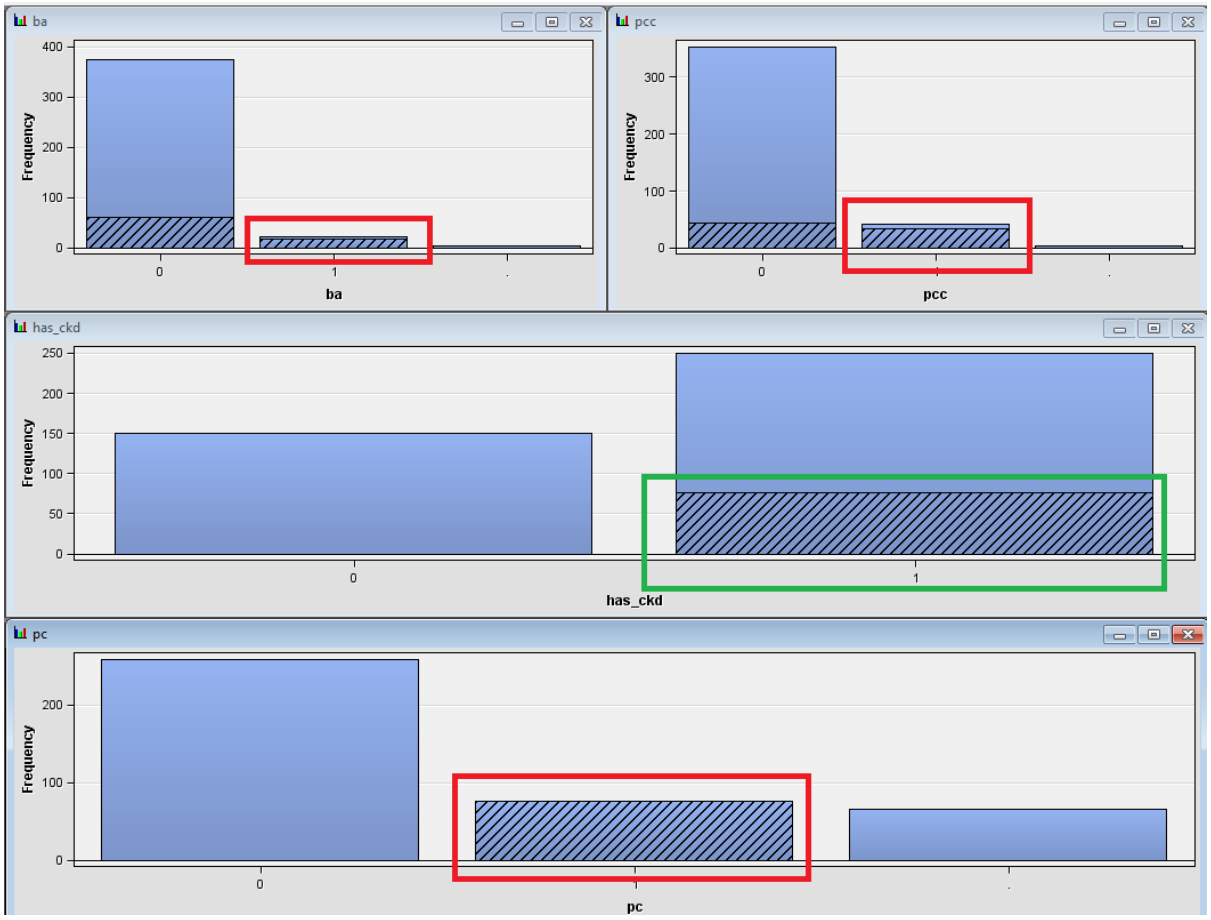


Figura 7 - Comparação da piúria com a presença de aglomerados de leucócitos e infecção bacteriana

Sabendo que os aglomerados de leucócitos são indicativos de infecção urinária de uma maior gravidade, optei por rejeitar esta variável (**pcc**) em detrimento variável que indica a presença de leucócitos degenerados (**pc**) na urina, uma vez que é expectável que quem tenha aglomerados de leucócitos já tenha sido diagnosticado para a insuficiência renal crónica.

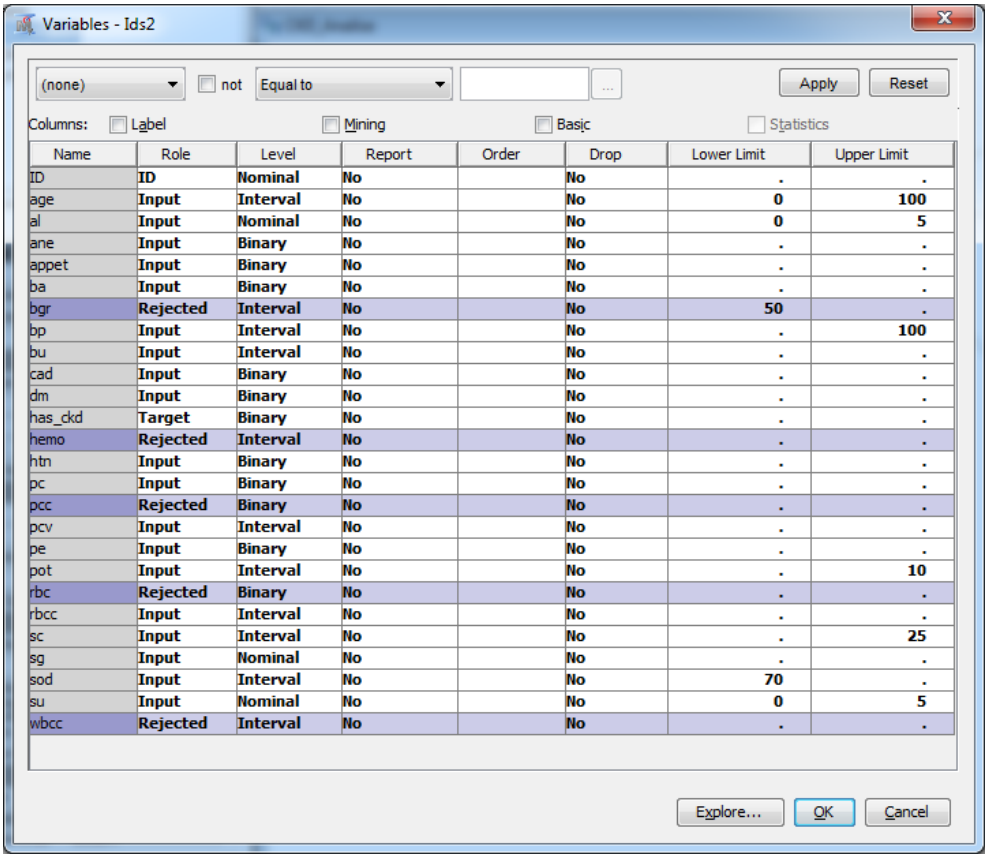
Situação das varáveis após o pré-processamento dos dados

Como vimos, a análise preliminar de variáveis permitiu identificar as situações com valores omissos e como tratá-las. Permitiu ainda identificar variáveis relacionadas e diminuir o espaço de *input*, com vista a simplificar o modelo a construir.

Assim, em termos de variáveis para a construção dos modelos, temos a seguinte situação:

- **ID do registo** = 1 variável com a identificação única de cada registo;
- **Variáveis de *input*** = 8 variáveis binárias + 3 variáveis categóricas + 8 variáveis contínuas
- **Variáveis rejeitadas** = 5 variáveis rejeitadas → Uma das variáveis foi rejeitada devido ao elevado número de valores omissos e devido ao facto de não ter sido possível identificar um valor a imputar. As outras quatro variáveis foram rejeitadas devido ao facto de se ter identificado que existia uma correlação entre elas e outras variáveis do modelo.
- **Variável *target*** = 1 variável, binária, indicativa da situação do paciente (com ou sem insuficiência renal crónica)

Na figura seguinte é apresentada a situação das variáveis antes da construção do modelo:



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
ID	ID	Nominal	No		No	.	.
age	Input	Interval	No		No	0	100
al	Input	Nominal	No		No	0	5
ane	Input	Binary	No		No	.	.
appet	Input	Binary	No		No	.	.
ba	Input	Binary	No		No	.	.
bgr	Rejected	Interval	No		No	50	.
bp	Input	Interval	No		No	.	100
bu	Input	Interval	No		No	.	.
cad	Input	Binary	No		No	.	.
dm	Input	Binary	No		No	.	.
has_ckd	Target	Binary	No		No	.	.
hemo	Rejected	Interval	No		No	.	.
htn	Input	Binary	No		No	.	.
pc	Input	Binary	No		No	.	.
pcc	Rejected	Binary	No		No	.	.
pcv	Input	Interval	No		No	.	.
pe	Input	Binary	No		No	.	.
pot	Input	Interval	No		No	.	10
rbc	Rejected	Binary	No		No	.	.
rbcc	Input	Interval	No		No	.	.
sc	Input	Interval	No		No	.	25
sg	Input	Nominal	No		No	.	.
sod	Input	Interval	No		No	70	.
su	Input	Nominal	No		No	0	5
wbcc	Rejected	Interval	No		No	.	.

Figura 8 – Situação das variáveis após o tratamento

Construção dos Modelos

A modelação preditiva permite inferir resultados desejados a partir de um conjunto de dados conhecidos. As técnicas utilizadas para criar modelos preditivos relacionam os dados entre si e vão criando aproximações ao resultado desejado. Este tipo de modelos permite prever resultados possíveis sobre uma determinada situação e identificar relações entre os dados que não são óbvias.

Neste caso de estudo pretende-se inferir, mediante um determinado conjunto de variáveis qual a probabilidade de um paciente apresentar insuficiência renal crónica.

Partição de dados

Para a criação de um modelo preditivo é necessário que exista um *subset* de dados que sirva de base à criação do modelo e um *subset* de dados para validar o rigor do modelo criado.

Ao primeiro *subset*, que comporta a maior percentagem de dados, chamamos **conjunto de treino**. Ao segundo *subset*, chamamos **conjunto de validação**, pois servirá para validar se o modelo criado está a responder adequadamente aos dados.

Poderá ainda existir um terceiro *subset* de dados para teste do modelo, que é precisamente o conjunto de testes.

Para o caso em estudo, optei por dividir os dados nos conjuntos de treino e validação, com 70% e 30% dos dados da amostra, como se pode ver na figura seguinte:

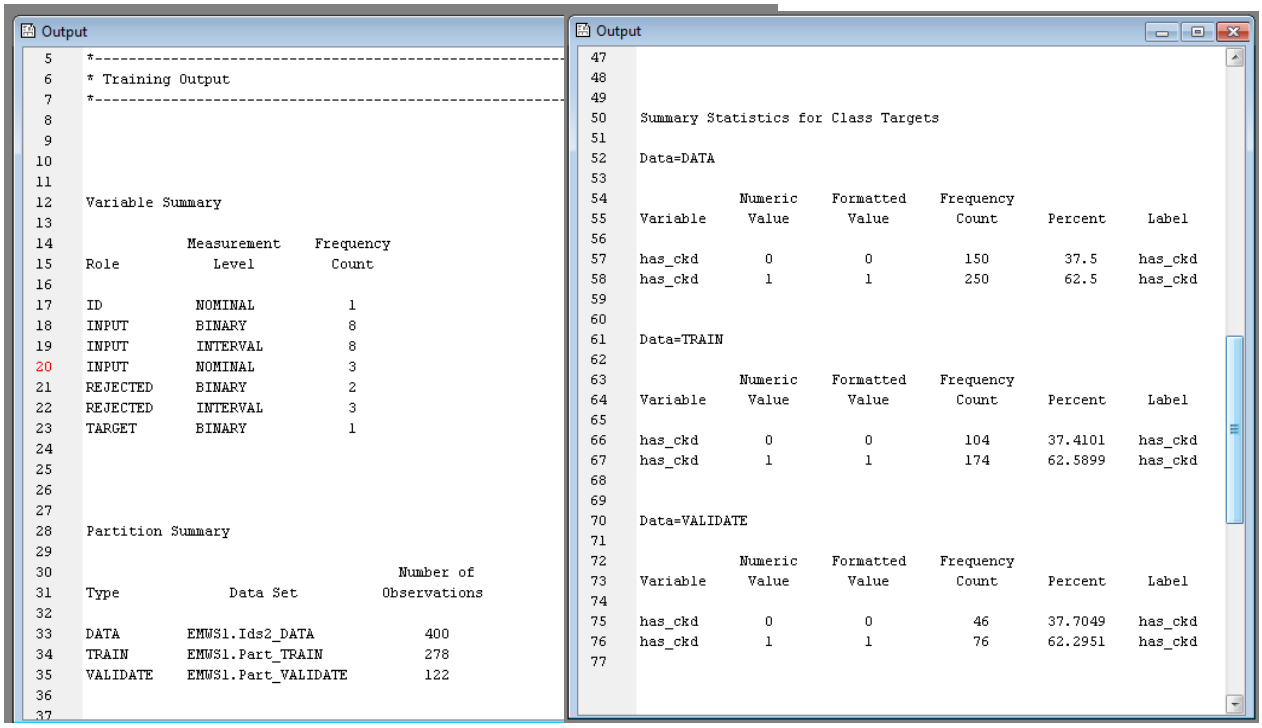


Figura 9 – Resultados da partição de dados

Em resumo:

- Conjunto de treino = 278 ocorrências, correspondentes a 69,50% do total
- Conjunto de validação = 122 ocorrências, correspondentes aos restantes 30,50% da amostra

Target	Data set	Conjunto de treino	Conjunto de validação
Sem insuficiência renal crónica	150 ocorrências, correspondentes a 37,50% do total do *data set*	104 ocorrências, correspondentes a 37,41% do total do *subset* de treino	46 ocorrências, correspondentes a 37,70% do total do *subset* de validação
Com insuficiência renal crónica	250 ocorrências, correspondentes a 62,50% do total do *data set*	174 ocorrências, correspondentes a 62,59% do total do *subset* de treino	76 ocorrências, correspondentes s 62,30% do total do *subset* de validação

Modelos preditivos

A criação de modelos preditivos é feita com base em técnicas que recorrem a algoritmos bem estabelecidos, cujo objetivo é evidenciar quais os parâmetros disponíveis influenciam o *target*. Neste caso específico, foram usadas duas técnicas diferentes:

- Árvores de decisão
- Regressão Logística

Árvores de Decisão

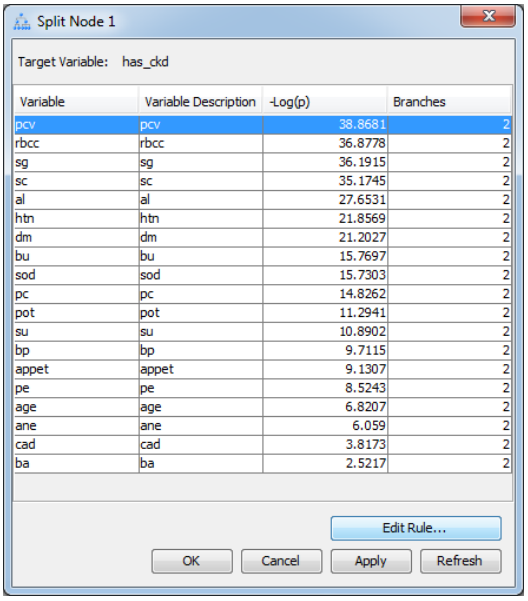
As árvores de decisão permitem representar as regras de um modelo de maneira relativamente simples. A construção de uma árvore de decisão é feita de tal forma que a cada divisão efectuada, o nível de conhecimento sobre os dados vai aumentando. A forma como são construídas implica que:

- As variáveis mais relevantes se encontram logo no topo da árvore;
- O conhecimento sobre os dados vai sendo cada vez maior à medida que a árvore vai sendo dividida (ou seja, o nível de conhecimento sobre os dados é maior nas folhas da árvore do que na raiz).

No **EMiner**™ podemos criar árvores de decisão de forma manual ou de forma automática. Mesmo no caso manual, o **EMiner**™ sugere a sequência da partição dos ramos, que é, obviamente, idêntica à usada no algoritmo automático.

Árvores de decisão manuais

No sentido de obter alguma sensibilidade relativamente à importância das variáveis na construção do modelo, optei, numa primeira fase, por criar uma árvore de decisão de forma manual. Ao criar os ramos manualmente, é logo apresentada a importância relativa das variáveis:



Variable	Variable Description	-Log(p)	Branches
pcv	pcv	38.8681	2
rbcc	rbcc	36.8778	2
sg	sg	36.1915	2
sc	sc	35.1745	2
al	al	27.6531	2
htn	htn	21.8569	2
dm	dm	21.2027	2
bu	bu	15.7697	2
sod	sod	15.7303	2
pc	pc	14.8262	2
pot	pot	11.2941	2
su	su	10.8902	2
bp	bp	9.7115	2
appet	appet	9.1307	2
pe	pe	8.5243	2
age	age	6.8207	2
ane	ane	6.059	2
cad	cad	3.8173	2
ba	ba	2.5217	2

Figura 10 – Árvore de decisão – importância das variáveis

Podemos concluir que, para o *target* pretendido, as varáveis mais relevantes (ou seja, aquelas que têm maior influência na identificação da insuficiência renal crónica), são, por esta ordem:

1. Volume globular (pvc)
2. Contagem de glóbulos vermelhos (rbcc)
3. Densidade da urina (sg)
4. Creatinina no sangue (sc)
5. Albumina na urina (al)
6. Hipertensão (htn)
7. Diabetes (dm)
8. Ureia no sangue (bu)
9. Sódio (sod)
10. Leucócitos na urina (pc)
11. Potássio (pot)
12. Açúcar na urina (su)
13. Pressão arterial (bp)
14. Apetite (appet)
15. Edema dos membros inferiores (pe)
16. Idade (age)
17. Anemia (ane)
18. Doença coronária (cad)
19. Bactérias na urina (ba)

Árvores de decisão automáticas

Na criação de uma árvore de decisão automática o **EMiner™** disponibiliza uma série de propriedades que caracterizam o tipo de árvore de decisão a obter:

- Qual o método a usar para proceder à divisão dos ramos e determinar a importância das variáveis;
- Possibilidade de usar a mesma variável (com diferentes valores) em níveis de decisão distintos, ou, em alternativa, não permitir que a mesma variável se repita em níveis de decisão diferentes;
- Número máximo de níveis criados de forma automática.

Por omissão, no **EMiner™**, as árvores de decisão automáticas caracterizam-se por:

- Apresentar 6 níveis de decisão
- Em cada iteração, efetuar a divisão em dois ramos;
- Repetição das variáveis nos vários níveis de decisão, o que significa que uma variável que já foi usada numa divisão anterior da árvore pode voltar a ser utilizada novamente num nível mais abaixo, para restringir ainda mais o modelo;

O resultado da execução da árvore de decisão automática com estas características é apresentado na figura seguinte:

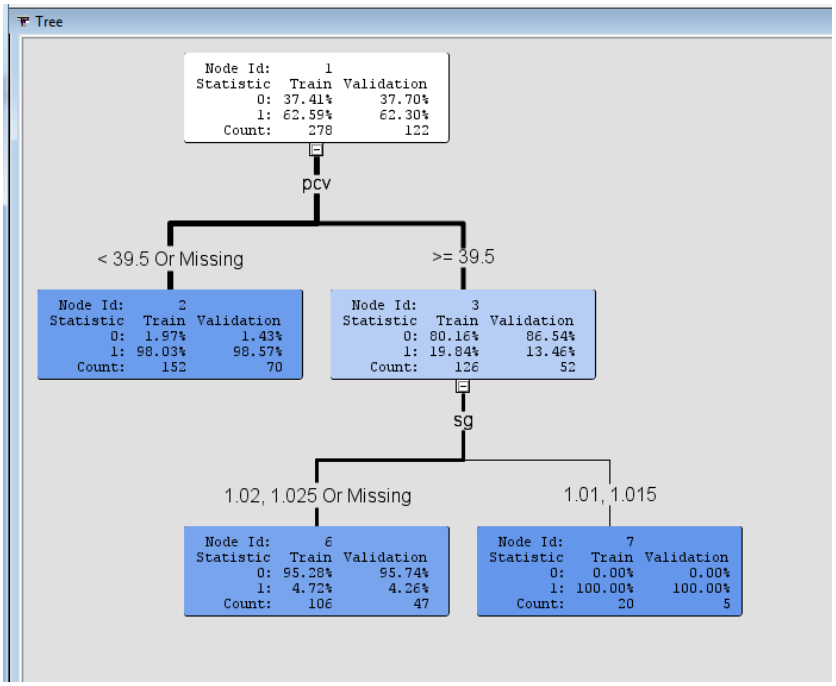


Figura 11 – Árvore de decisão automática – resultados

Como já era esperado, a divisão é efetuada pelem primeiro lugar pelo volume globular (**pvc**) e da mesma forma que foi sugerido na árvore de decisão interativa. Com os resultados da primeira iteração, o **EMiner**™ não efetua mais nenhuma divisão no ramo em que os pacientes apresentam volume globular inferior a 39,50%.

No ramo em que o volume globular é superior ou igual a 39,50% o **EMiner**™ efetuou uma segunda iteração, com base nos valores da densidade da urina. Para valores de 1.020, 1.025 todos os pacientes apresentam insuficiência renal. Apesar dos valores de referência se situarem entre os 1.005 e 1.030, com estes resultados podemos concluir que são os valores mais baixos que são indicativos de insuficiência renal crónica.

Análise dos resultados

Com este algoritmo pode-se inferir que o volume globular baixo e a densidade da urina baixa são indicativos de insuficiência renal, com especial destaque para a primeira variável. No entanto, conclui que a execução de uma árvore de decisão automática para este conjunto de dados não nos dá grande informação sobre as variáveis que podem ser indicativas de insuficiência renal crónica. Esta situação está provavelmente relacionada, uma vez mais, com o reduzido de observações desta amostra.

Regressão estatística

A regressão consiste no estudo do impacto das alterações de uma variável independente sobre uma variável dependente, ou seja, estuda a relação causa-efeito entre duas variáveis. A regressão pode ser:

- **Linear**, e nesse caso a relação causa-efeito é traduzida por uma função linear entre a variável que se está a estudar e as variáveis que irão modelar o estudo.
- **Logística** e neste caso, a relação causa-efeito é também traduzida por uma função linear. No entanto é baseada em logaritmos e permite efetuar previsões sobre um determinado

acontecimento, resultantes da análise de um conjunto de dados conhecidos sobre os eventos a predizer. Uma diferença importante em relação à regressão linear é que a variável objectivo do estudo tem de ser binária (ou categórica).

No caso em estudo, dado que se pretende efetuar uma previsão sobre a propensão para a insuficiência renal, o tipo de regressão adequado será a regressão logística. A regressão logística pode ser aplicada, recorrendo a várias técnicas de seleção de variáveis:

- **Foreward** – Abordagem do particular para o geral. Implica ir adicionando ao modelo as variáveis e a fiabilidade do modelo é testada a cada iteração.
- **Backward** – Abordagem do geral para o particular. O modelo é testado com as variáveis todas que vão sendo retiradas à medida que o modelo vai sendo construído. O modelo é testado a cada iteração.
- **Stepwise** – Corresponde a uma mistura dos dois métodos anteriores, em que em cada iteração as variáveis vão sendo incluídas ou retiradas, dependendo da fiabilidade que conferem ao modelo.

No caso da regressão, iremos aplicar os três métodos. Como já foi anteriormente referido, no caso da regressão não é possível existirem valores omissos nos dados, pelo que forma imputados valores aos valores omissos, de acordo com o explicado no ponto relativo ao tratamento de dados.

Regressão Foreward

Neste tipo de abordagem, as variáveis vão sendo progressivamente adicionadas ao modelo e em cada iteração o modelo é testado.

Ao aplicar este algoritmo, foram realizadas cinco iterações e as variáveis foram adicionadas ao modelo pela seguinte ordem:

Iteração #	Variável adicionada
1	Densidade da urina (sg)
2	Hipertensão (htn)
3	Albumina no sangue (al)
4	Volume globular (pcv)
5	Creatinina no sangue (sc)

Neste caso, o resultado final da aplicação do algoritmo foi:

Odds Ratio Estimates

Effect		Point Estimate
IMP_al	0 vs 4	<0.001
IMP_al	1 vs 4	999.000
IMP_al	2 vs 4	999.000
IMP_al	3 vs 4	999.000
IMP_htn	0 vs 1	<0.001
IMP_pcv		<0.001
IMP_sc		999.000
IMP_sg	0 vs 1.025	999.000
IMP_sg	1.005 vs 1.025	<0.001
IMP_sg	1.01 vs 1.025	999.000
IMP_sg	1.015 vs 1.025	999.000
IMP_sg	1.02 vs 1.025	999.000

Figura 12 – Regressão *foreward* – resultados

Com a aplicação deste algoritmo, verifica-se que existem duas variáveis que são indicativas da presença de insuficiência renal crónica:

- **Albumina na urina** (**al**) → Neste caso, para valores de 1, 2, 3 ou 4 a probabilidade dos pacientes apresentarem insuficiência renal crónica é elevadíssima, face aos que não apresentam albumina na urina.
- **Creatinina** (**sc**) → A presença de creatinina no sangue é também fortemente indicativa da insuficiência renal crónica.
- **Densidade da urina** (**sg**) → Valores acima dos 1.005 são fortemente indicativos de insuficiência renal.

Regressão Backward

Neste tipo de abordagem, opta-se por criar um modelo com todas as variáveis que vão sendo progressivamente eliminadas ao longo do processo. A eliminação das variáveis é processada pela ordem que torna o modelo mais fiável, ou seja, podemos dizer que vão sendo retiradas as variáveis que menos influenciam o modelo. Em cada iteração a fiabilidade do modelo é avaliada.

No caso em estudo, foram incluídas todas as variáveis, como se mostra na figura seguinte:

Backward Elimination Procedure

Step 0: The following effects were entered.

Intercept IMP_age IMP_al IMP_ane IMP_appet IMP_ba IMP_bp IMP_bu IMP_cad IMP_htn IMP_pc IMP_pcv IMP_pe IMP_pot IMP_rbcc IMP_sc IMP_sg IMP_sod IMP_su

Figura 13 – Variáveis usadas no início da regressão (*backward*)

O algoritmo de regressão *backward* realizou 14 iterações e as variáveis foram retiradas pela ordem que se apresenta no quadro seguinte:

Iteração #	Variável retirada
1	Açúcar na urina (su)
2	Albumina na urina (al)
3	Ureia no sangue (bu)
4	Doença coronária (cad)
5	Bactérias na urina (ba)
6	Sódio (sod)
7	Anemia (ane)
8	Idade do paciente (age)
9	Pressão arterial diastólica (bp)
10	Densidade da urina (sg)
11	Edema dos membros inferiores (pe)
12	Alterações do apetite (appet)
13	Leucócitos na urina (pc)
14	Hipertensão (htn)

O resultado da aplicação do algoritmo foi:

Odds Ratio Estimates

Effect		Point Estimate
IMP_appet	0 vs 1	<0.001
IMP_htn	0 vs 1	<0.001
IMP_pc	0 vs 1	999.000
IMP_pcv		0.135
IMP_pe	0 vs 1	<0.001
IMP_pot		0.109
IMP_rbcc		0.058
IMP_sc		51.171
IMP_sg	0 vs 1.025	999.000
IMP_sg	1.005 vs 1.025	999.000
IMP_sg	1.01 vs 1.025	999.000
IMP_sg	1.015 vs 1.025	999.000
IMP_sg	1.02 vs 1.025	84.559

Figura 14 – Regressão *backward* – resultados

Com a aplicação deste algoritmo, verifica-se que existem seis variáveis que são indicativas da presença de insuficiência renal crónica:

- **Leucócitos na urina** (**pc**) → A presença de leucócitos é indicativa de insuficiência renal;

- **Creatinina** (**su**)→ Valores elevados de creatinina no sangue são fortemente indicativa de insuficiência renal crónica;
- **Densidade da urina** (**sg**) → Neste caso os resultados são semelhantes aos obtidos na regressão *foreward,* com a diferença que mesmo os valores baixos são também indicativos de insuficiência renal.

Regressão Stepwise

Neste tipo de abordagem, como já referi, as variáveis vão sendo inseridas ou retiradas do modelo, em cada iteração. Neste caso, foram realizadas três iterações e a ordem pela qual as variáveis foram processadas foi seguinte:

Iteração #	Variável processada
1	Densidade da urina (sg) – adicionada
2	Hipertensão (htn) - adicionada
3	Hipertensão (htn) - retirada

Os resultados da aplicação da abordagem *stepwise* foram são apresentados na figura seguinte:

Odds Ratio Estimates

Effect		Point Estimate
IMP_sg	0 vs 1.025	45.321
IMP_sg	1.005 vs 1.025	999.000
IMP_sg	1.01 vs 1.025	999.000
IMP_sg	1.015 vs 1.025	999.000
IMP_sg	1.02 vs 1.025	2.660

Figura 15 – Regressão *stepwise* – resultados

Neste modelo, verifica-se que a densidade da urina é fortemente indicativa de insuficiência renal crónica.

Comparação dos Modelos Criados

Para comparar os diferentes modelos criados, usei a funcionalidade que o **EMiner™** disponibiliza e estabelece qual o melhor modelo, face ao comportamento deste em relação aos dados.

A fiabilidade do modelos obtidos pode ser validada pela análise dos seguintes aspectos:

- **Sensibilidade**: Traduz a capacidade de um modelo de prever os casos positivos. É dada pela razão entre os valores positivos identificados para um modelo e o total de positivos (Verdadeiros positivos e falsos negativos)
- **Especificidade:** Traduz a resposta de um modelo na identificação de casos negativos (verdadeiros negativos). É dada pela razão entre os verdadeiros negativos identificados no modelo e o total de negativos (Verdadeiros negativos e Falsos Positivos);

Os resultados da comparação foram os seguintes:

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Valid: Misclassification Rate
Y	Reg2	Reg2	Regression Backward	has_ckd	has_ckd	0.008197
	Reg	Reg	Regression Forward	has_ckd	has_ckd	0.016393
	Tree2	Tree2	Auto Decision Tree	has_ckd	has_ckd	0.02459
	Reg3	Reg3	Regression Stepwise	has_ckd	has_ckd	0.04918

Figura 16 – Resultado da comparação dos modelos

Desta comparação podemos ver que o modelo que apresenta melhor performance é a regressão *backward* pelo valor do parâmetro *Missclassification rate*, que mede a proporção de casos que estão a ser incorretamente associados a um determinado grupo.

Conclusões

Com este trabalho pretendi demonstrar que a aplicação de algoritmos de *data mining* sobre dados médicos pode fornecer informação relevante acerca de uma determinada patologia. Isso poderá permitir antecipar o diagnóstico de uma doença e ter influência direta no aumento da esperança e qualidade de vida dos pacientes.

Nesse sentido, usei a informação disponível relativa à insuficiência renal crónica, com vista a criar um modelo que permitisse identificar quais as variáveis mais determinantes para a identificação da patologia. Na construção do modelo foram aplicadas as seguintes técnicas:

- Construção de árvores de decisão;
- Regressão logística

De todos os modelos criados, o **EMiner™** considerou a regressão numa abordagem *backward* como sendo o modelo mais adequado para o objectivo do estudo. Assim, com base nos resultados obtidos pela aplicação deste modelo (ver figura 14) é possível concluir que as variáveis que apontam para uma maior probabilidade do paciente apresentar insuficiência renal são:

- **Presença de leucócitos na urina** (pc) → Os leucócitos na urina são indicativos de infecção bacteriana no trato urinário. Este factor encarado de forma isolada não é suficiente para indicar de forma segura que um paciente tenha insuficiência renal, no entanto, sabe-se que infecções renais repetidas poderão ser indicativas de insuficiência renal.

- **Valores de creatinina acima do normal** (**sc**) → A creatinina é um subproduto da atividade muscular e está diretamente relacionada com a atividade física. Os rins excretam a creatinina em excesso, excepto no caso da função renal estar comprometida. Concentrações elevadas de creatinina no sangue são fortemente indicativas de insuficiência renal crónica. Assim, valores elevados de creatinina podem ser um alerta para o médico aprofundar o diagnóstico do paciente. Há exames médicos específicos para avaliar a forma como a creatinina está a ser processada por um individuo e que podem dar mais informação.

- **Densidade da urina** (**sg**) → Para valores baixos, pode indicar função renal comprometida, na medida os rins deixam de excretar adequadamente os subprodutos resultantes do metabolismo normal. No entanto, para valores altos, poderá ser indicativo de outras patologias associadas à insuficiência renal crónica. Assim, a avaliação desta variável deverá ser feita em conjunto com outros indicadores. No entanto é um sinal de alerta importante em que os médicos se podem sustentar para efetuar um diagnóstico.

Este é um exemplo de como a informação médica disponível pode ajudar na identificação de patologias numa fase inicial e contribuir para o controlo de uma doença numa fase mais precoce.

As fontes de dados na área da saúde são inúmeras e variadas: estudos clínicos e registos médicos representam um manancial de informação sobre os efeitos das terapêuticas aplicadas. A evolução da medicina e da bioquímica colocou à nossa disposição o conhecimento sobre o DNA, a síntese de proteínas e os processos metabólicos que vieram lançar novas luzes sobre o funcionamento do corpo humano. A evolução da tecnologia associada à imagem médica permite ver em detalhe o funcionamento dos órgãos internos em geral e do cérebro, em particular.

A resposta à questão: "Este tratamento vai funcionar?" pode ser hoje dada com muito maior precisão. Muitos tratamentos falham ou diminuem a sua eficácia devido aos hábitos dos pacientes e à falta de adesão ao tratamento preconizado. Existem medicamentos para o tratamento do cancro da mama que são 100% eficazes na erradicação dos tumores em pacientes com um determinado marcador genético, mas totalmente ineficazes nas pacientes que não têm esse marcador. Assim, o sequenciamento do DNA e RNA dos das pacientes poderá dar indicações precisas de que terapêuticas devem ser aplicadas.

Assim, conclui-se que na área da saúde, os dados coligidos digitalmente até ao momento detêm um enorme potencial de informação que pode ser usado para melhor a qualidade dos cuidados de saúde a aplicar aos pacientes. Normalmente, quando um paciente apresenta determinados sintomas ou lhe diagnosticada uma doença, é aplicada uma série de tratamentos que o médico considera que funcionam melhor naquele caso. Esse tratamento depende de factores extrínsecos como a experiência prévia do médico e/ou estudos efectuados. Na verdade, não existe conhecimento sobre que tratamentos são mais eficazes em determinados pacientes. Isso faz com que muitas vezes sejam aplicadas terapêuticas que se revelam ineficazes.

No futuro próximo, o conhecimento da relação entre determinadas terapêuticas e resultados positivos irá produzir uma revolução na prática da medicina: os pacientes podem ser classificados de acordo com as suas características, hábitos, perfil genético e metabólico e os tratamentos serão mais adequados às suas necessidades e características, prolongando a sua esperança de vida e diminuindo de forma significativa os custos associados à área da saúde.

Referências

Coelho, I. (2015) **"Slides de apoio às Aulas Teóricas de Data Mining"**, Universidade Lusófona;

World Health Organization, (2013) **"The top 10 causes of death"**, disponível em http://www.who.int/mediacentre/factsheets/fs310/en/, último acesso em 22 Março de 2016;

Kart, L., Herschel, G., Linden A., Hare, J. (9 Feb 2016) **"Magic Quadrant for Advanced Analytics Platforms"**, Gartner Inc.;

O'Reilly, T., Steele, J., Loukides, M. e Hill, C. (2012) **"Data Science and the Health Care Revolution",** disponível em http://www.forbes.com/sites#/sites/oreillymedia/2012/08/20/data-science-and-the-health-care-revolution/#2aa09acb4369, ultimo acesso em 22 de Março de 2016;

BioCampello, **"Interpretação das suas Análises Clinicas",** disponível em http://www.biocampello.com/Interpretacao-das-suas-Analises-Clinicas, último acesso em 22 Março de 2016;

Against All Odds Productions, (2016) **"The Human Face of Big Data"** [Ficheiro em video], disponível em: https://www.youtube.com/watch?v=r6v15Z60eUI, último acesso em 22 de Março de 2016.

Kart, L., Herschel, G., Linden A. & Hare, J. (9 Feb 2016) "Magic Quadrant for Advanced Analytics Platforms", Gartner Inc.

Referências do caso de estudo:

L.Jerlin Rubini (Research Scholar) Dr.P.Soundarapandian.M.D. D.M (Senior Consultant Nephrologist) and (Alagappa University) Dr.P.Eswaran. **"Chronic kidney disease data set"** – UCI machine learning repository, 2015.

Fonte original dos dados: Soundarapandian P. M. D. (2015), Tamil Nadu, Karaikudi, Apollo Hopitals

Dataset criado por: Rubin L. Jerlin (2015), Eswaran, P. (2015). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Tamilnadu, Karaikudi, Alagappa University, Department of Computer Science and Engineering

ANEXOS

Anexo I – Quadrante Mágico Gartner

As análises efetuadas pela Gartner Inc. resultam num posicionamento dos vários fabricantes numa matriz, denominada Quadrante Mágico, cujo significado se encontra explicado na figura seguinte:

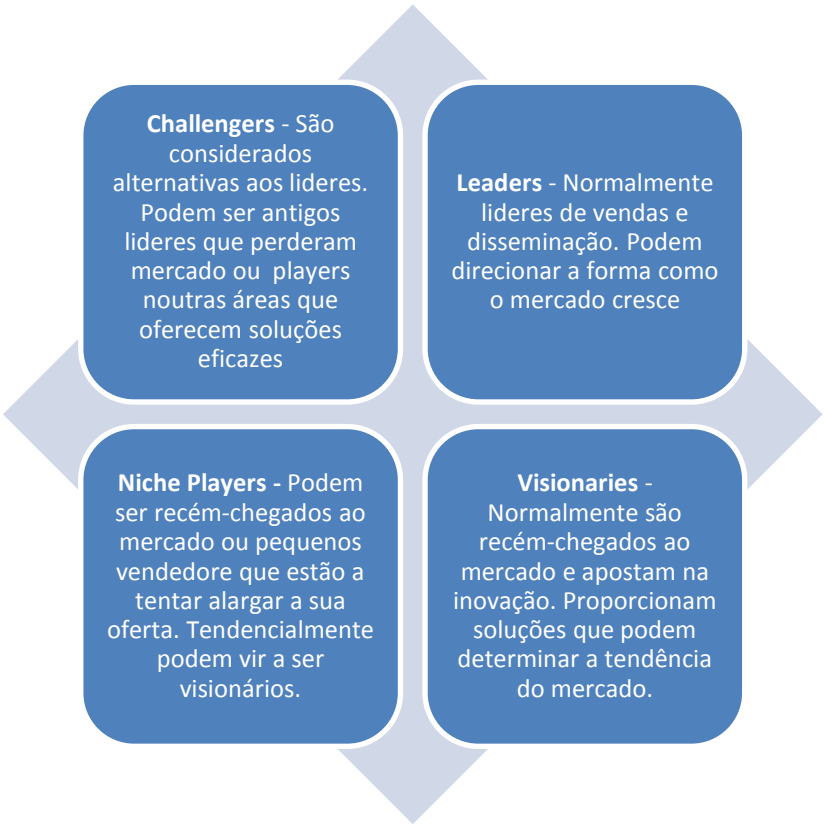
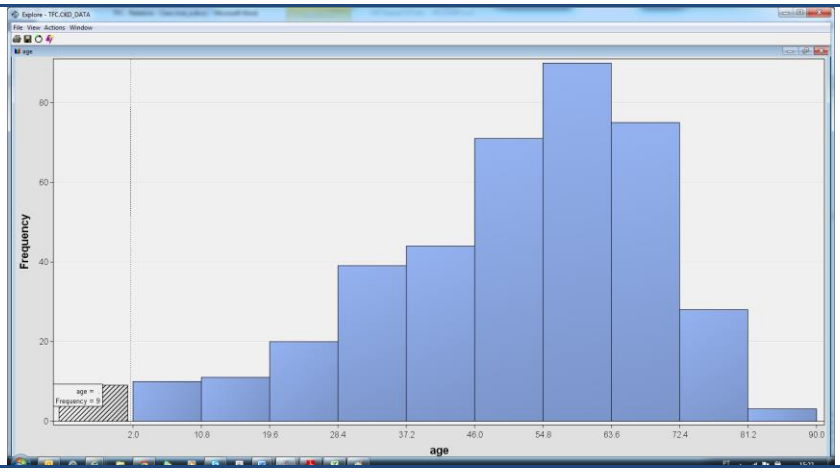


Figura 17 – Breve explicação do Quadrante Mágico[2] da Gartner

[2] Adaptado de Kart, L., Herschel, G., Linden A. & Hare, J. (9 Feb 2016) "Magic Quadrant for Advanced Analytics Platforms", Gartner Inc.

Anexo II – Estudo das variáveis em detalhe

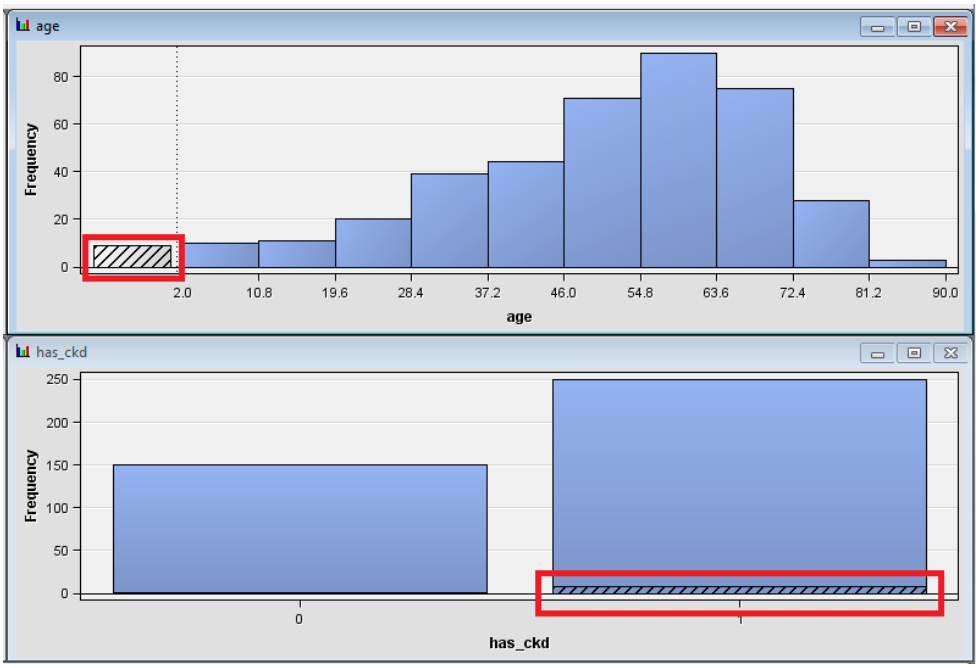
Idade do paciente (age)

Age	
Descrição	Idade do paciente. Trata-se de uma variável de *input*, com valores definidos entre 0 e 100.
Histograma	
Situação	Identificaram-se 9 registos com valor omisso, ou seja, pacientes dos quais não sabemos a idade. Pela distribuição podemos ver que as idades dos pacientes observados variam entre 2 e 90 anos, sendo que a maior frequência se situa entre os 40 e os 70 anos.
Decisão	Analisando cada intervalo de idades, não foi possível estabelecer uma co-relação directa entre a idade do paciente e a existência de insuficiência renal, com excepção do caso dos pacientes muito jovens que fazem parte deste *data set*. Apesar de ser um número bastante reduzido de pacientes entre os 2 e os 11 anos, podemos extrapolar que a probabilidade de pacientes muito jovens apresentarem sintomas de doença renal pode representar um forte indício de uma insuficiência renal crónica. Ainda assim, o número de observações não nos permite tirar essa conclusão. 

Age

O mesmo se pode dizer para os pacientes com idades muito avançadas, dado que no intervalo entre os 81 e os 90 anos, os 3 pacientes que fazem parte deste *data set* também apresentam insuficiência renal crónica, mas neste caso pode ser devido precisamente à idade.

Relativamente aos pacientes para os quais não sabemos a idade constatou-se que todos eles apresentam a patologia, como se pode verificar na figura seguinte:

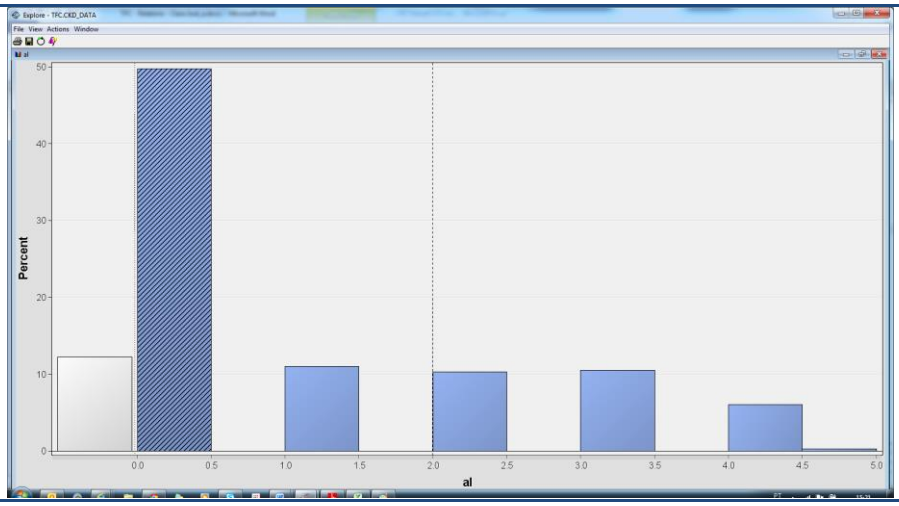
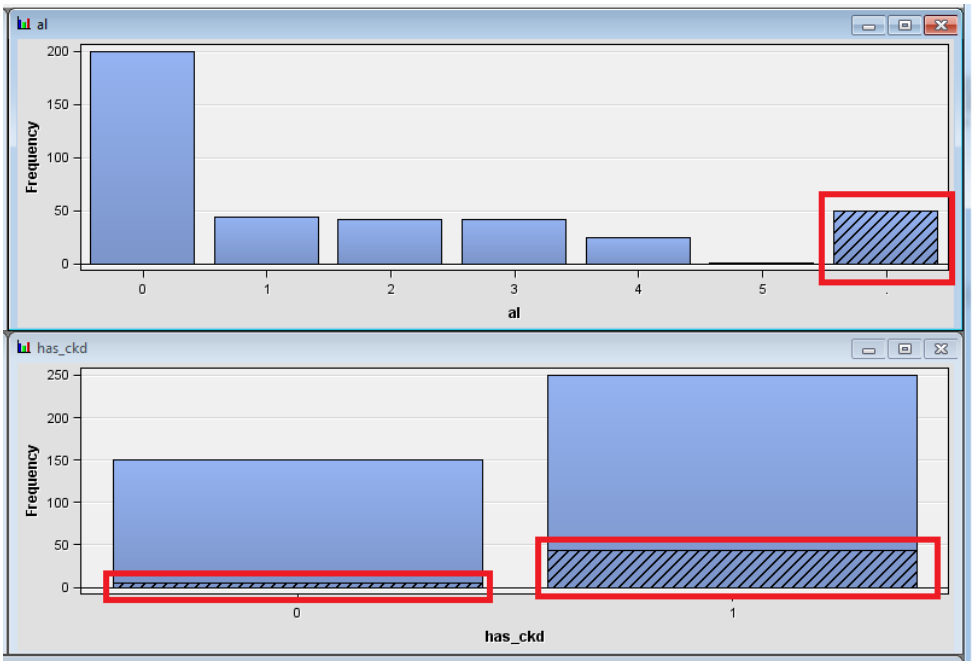


Assim, não é imediato o valor que devemos atribuir a variável deve ter, pois surgem três hipóteses: ou o paciente é muito jovem ou muito idoso ou nem uma coisa nem outra.

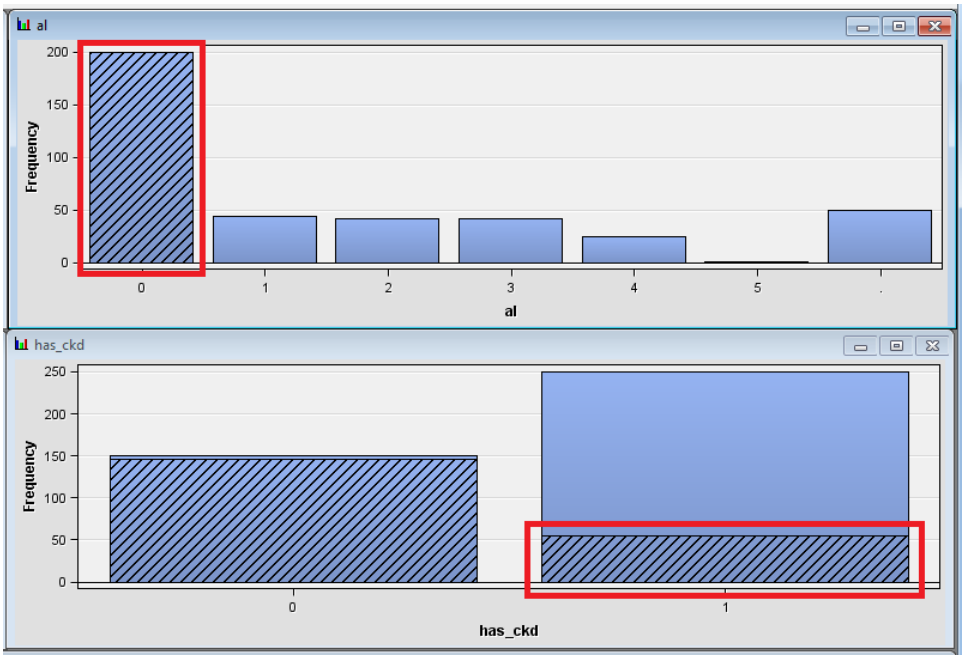
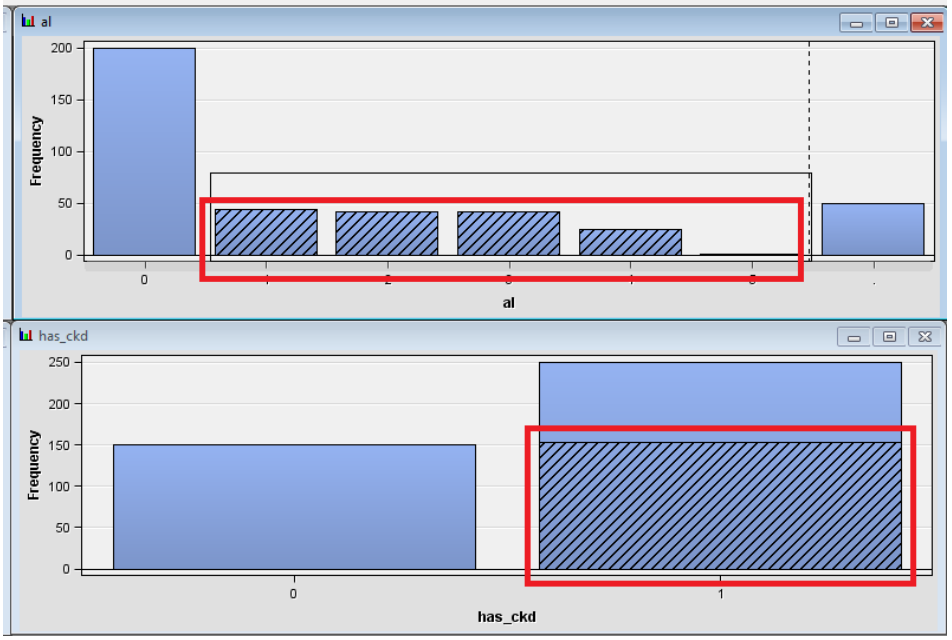
Ao analisar esta questão comparei esta variável com outras, tais como a presença de anemia e a prevalência de diabetes, mas também não foi possivel estabelecer nenhuma relação com a idade.

Assim, para tratamento desta varável, decidi usar a média das idades dos pacientes, que ronda os 52 anos.

Albumina (al)

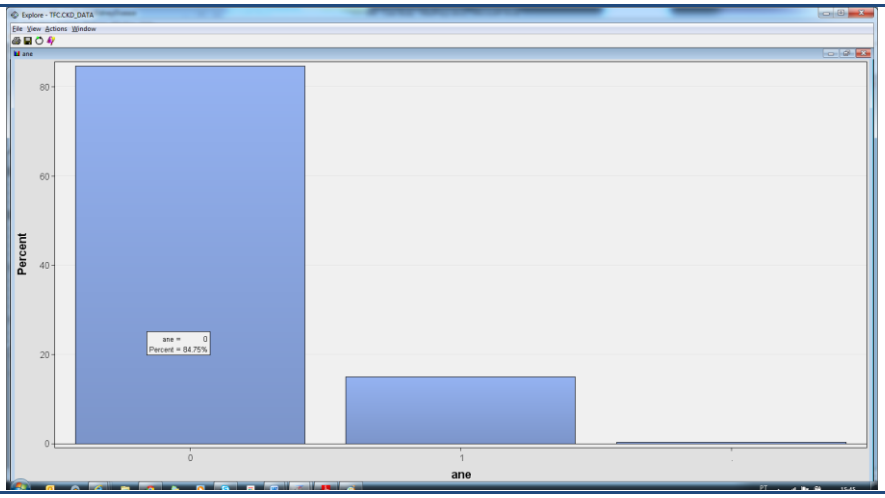
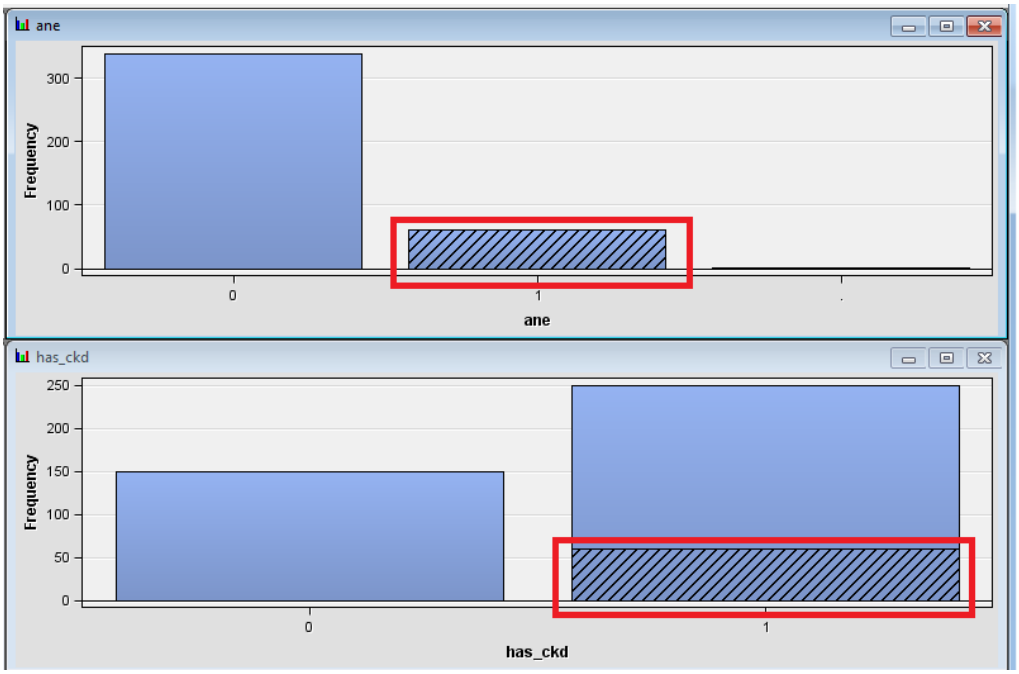
al	
Descrição	Trata-se de uma variável de *input*, com valores inteiros específicos entre 0 e 5. A presença de albumina na urina pode indicar uma lesão renal e sabe-se hoje que pode ser indicativa de uma futura insuficiência renal crónica.[3]
Histograma	
Situação	Identificaram-se 49 registos sem valores para esta variável, o que representa 12,25% da amostra:  Foi possível determinar que a maior parte dos pacientes apresenta valores de albumina 0 (199 do total dos 400), ou seja cerca de 50% do total da amostra. No entanto, uma parte destes pacientes apresenta a patologia, apesar de não acusar a

[3] Ver o artigo "Albumina na urina pode prever doenças renais", baseado em estudo publicado no PLoS Genetics, http://www.alert-online.com/pt/news/health-portal/albumina-na-urina-pode-prever-doencas-renais

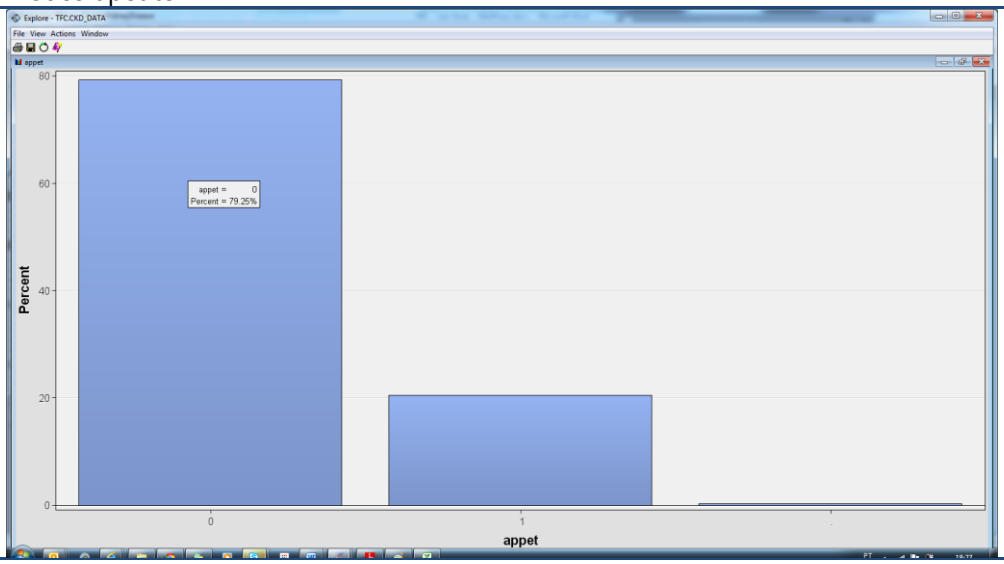
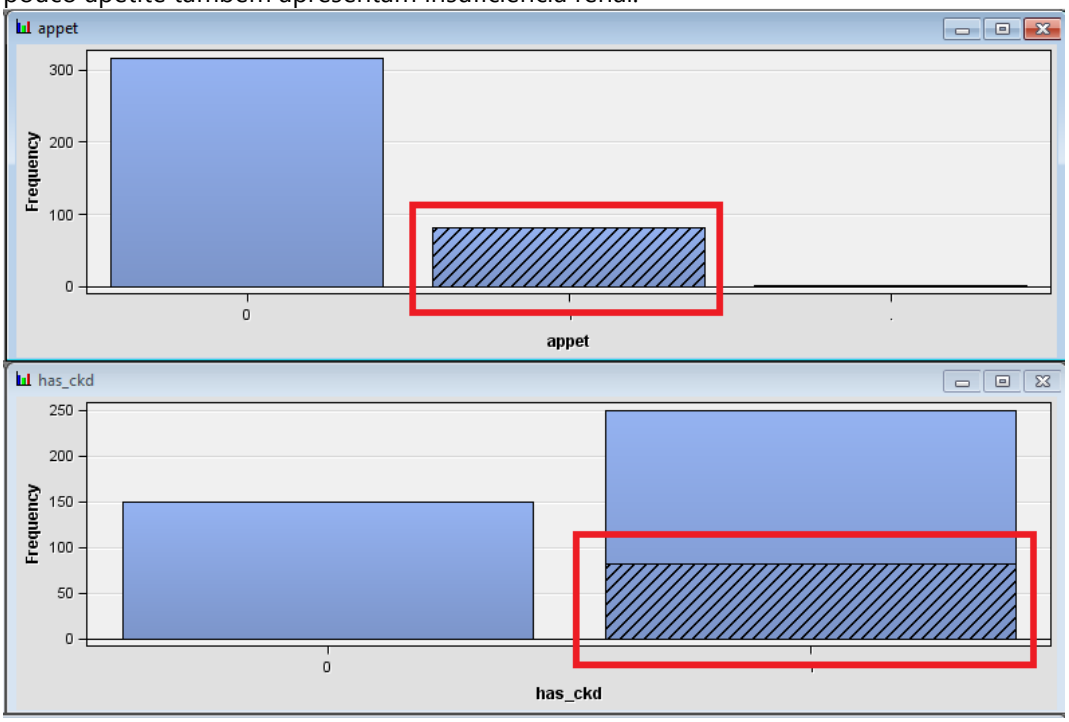
al	
	presença de albumina na urina:  Por outro lado, todos os pacientes que acusam a presença de albumina, todos eles, sem exceção são doentes renais crónicos, como se pode ver na figura seguinte: 
Decisão	Com este conjunto de dados e no sentido de tratar os valores omissos, tinha duas hipóteste igualmente viáveis: 1. Ou considerava o valor 0 (ou seja, sem a presença de albunina na urina) → Uma vez que a maior parte dos pacientes com valor omisso tem insuficiência renal, considerar 0 poder comportar uma diminuição do peso desta variável na determinação da doença renal crónica.

al	
	2. Ou aplicava a média (que é aproximadamente 1) → Neste caso, os pacientes que não têm insuficiência renal, podem exercer um ligeiro desvio nos resultados do modelo, uma vez que ,pelos dados que dispomos sempre que aparece albunina na urina isso é indicativo da patologia. Assim e sabendo de antemão quais os riscos, optei por considerar que os valores omissos devem ser preenchidos com 0, pois é a situaçao que melhor traduz a realidade que os dados apresentam.

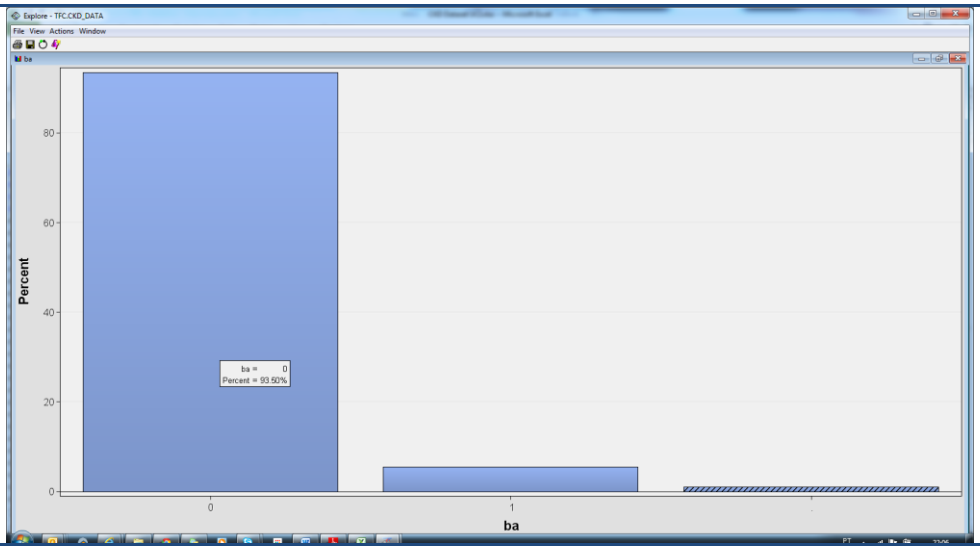
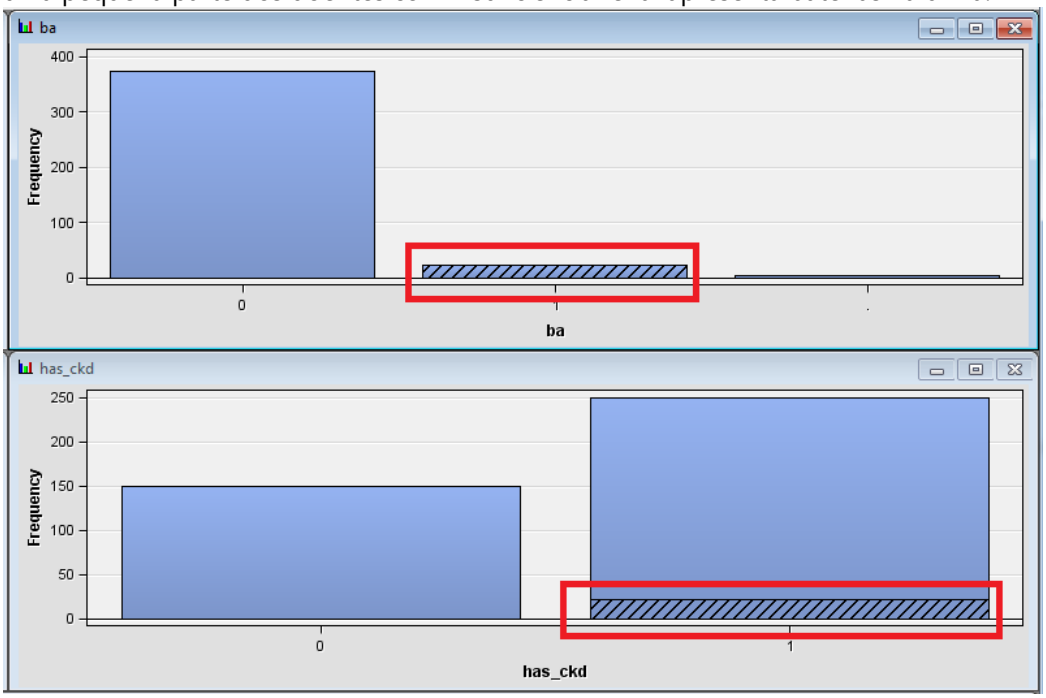
Anemia(ane)

ane	
Descrição	Variável de *input*, binária indicativa se o paciente apresenta anemia ou não. 0 – Sem anemia 1 – Com anemia
Histograma	
Situação	Uma larga percentagem dos pacientes (cerca de 85% não apresenta anemia). Neste caso existe apenas um paciente que não apresenta valor para esta variável.
Decisão	Apesar da percentagem de pacientes com anemia ser relativamente baixa, por comparação com a variável target é possível inferir que possa ser uma variável com significado na patologia, uma vez que existe uma relação direta entre os pacientes com anemia e a presença da patologia:  Assim, uma vez que o paciente com o valor omisso não apresenta insuficiência renal crónica optou-se por considerar que não tinha anemia, ou seja, optou-se por colocar o valor 0.

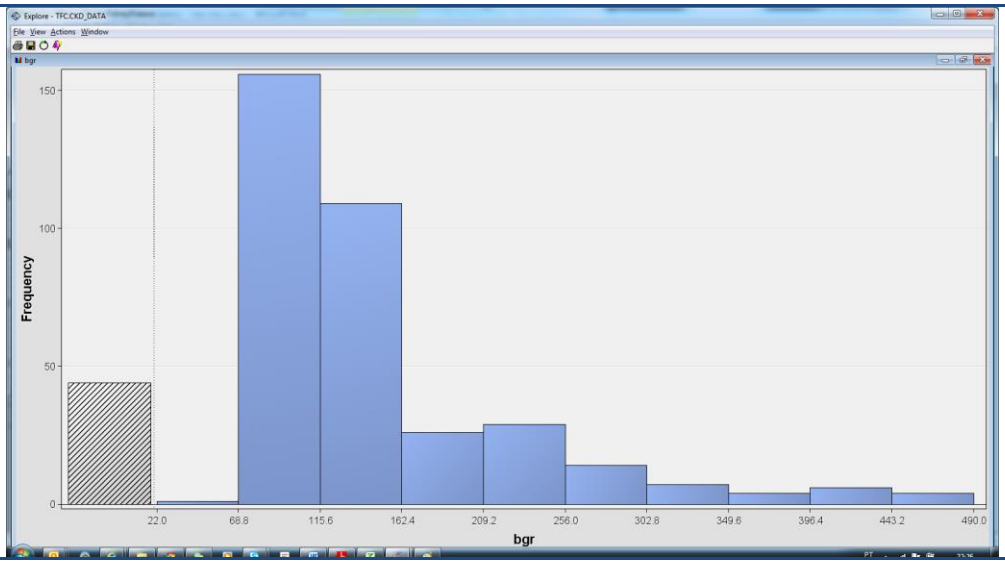
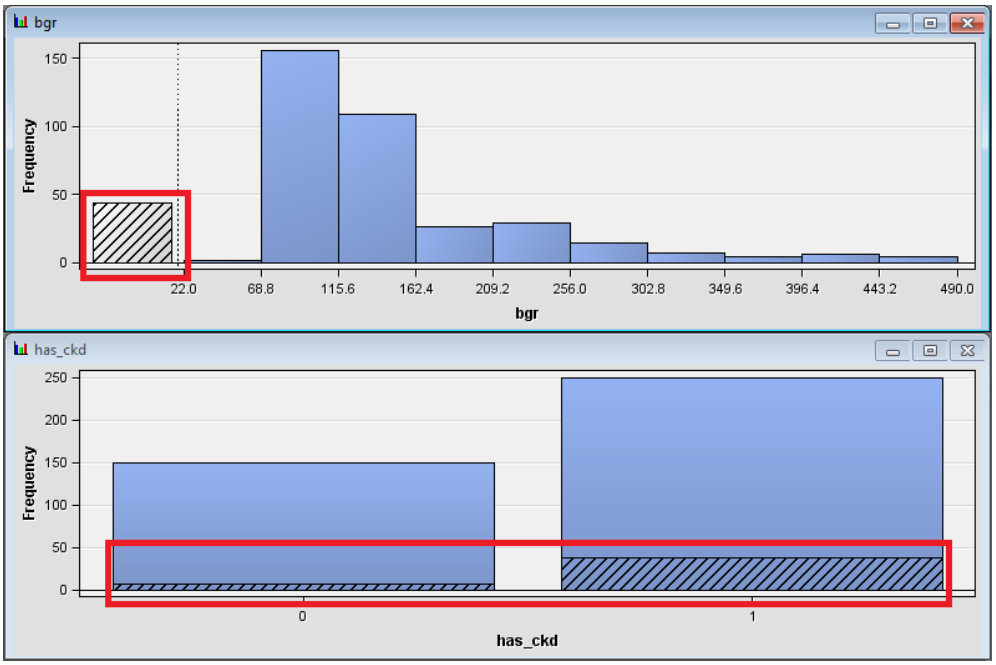
Apetite (appet)

appet	
Descrição	Variável de *input*, binária indicativa se o paciente tem alterações do apetite 0 – Sem alterações 1 – Pouco apetite
Histograma	
Situação	Cerca de 80% dos pacientes não apresentam alterações do apetite. Existe um paciente em que não foi indicado se existiam alterações no apetite.
Decisão	Por comparação com a variável target, foi possível constatar que todos os pacientes com pouco apetite também apresentam insuficiência renal:  Dado que o paciente com o valor omisso não apresenta insuficiência renal, decidi atribuir o valor 0 a esta varável (sem alterações de apetite).

Bactérias (ba)

ba	
Descrição	Variável binária indicativa da presença de bactérias na urina: 0 - Ausentes 1 - Presentes
Histograma	
Situação	93,50 % dos pacientes não acusam a presença de bactérias. Nesta variável existem ainda 4 pacientes para os quais não se sabe se têm ou não bactérias na urina.
Decisão	Pela comparação desta varável com o target, não existe uma relação evidente entre a ausência de bactérias e a insuficiência renal. Como se pode ver na comparação, apenas uma pequena parte dos doentes com insuficiência renal apresenta batérias na urina.  Relativamente aos valores omissos, uma vez que todos pertencem a pacientes que não apresentam a doença, optei por considerar que não apresentavam bactérias na urina, atribuindo valor 0 a esta variável.

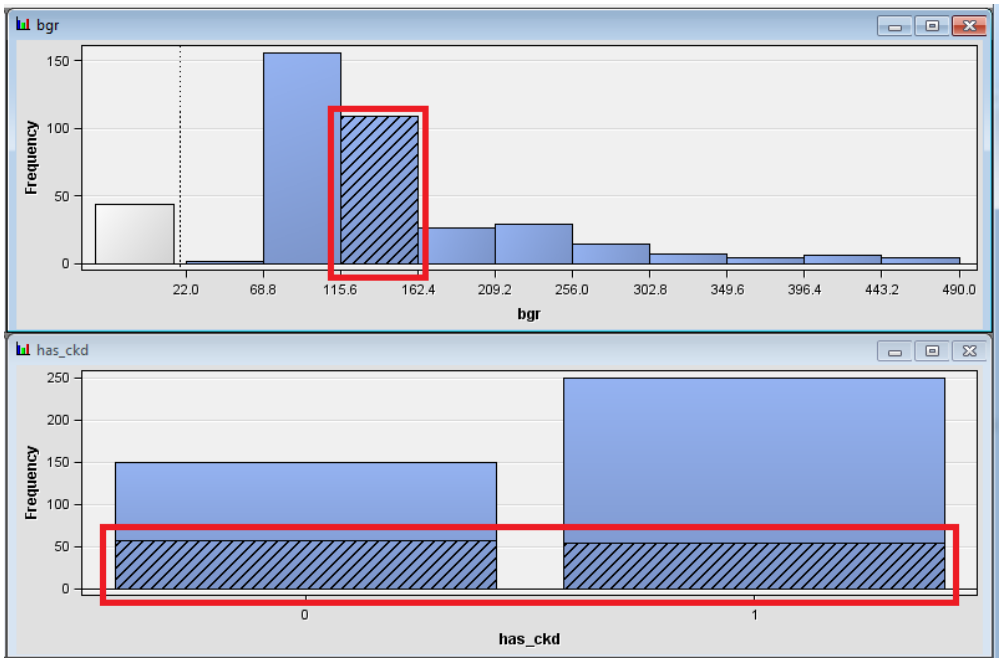
Glicémia(bgr)

bgr	
Descrição	Variável de *input* numérica, contínua, indicativa do nível de glicemia, ou seja, da concentração de glicose no sangue.
Histograma	
Situação	Foram identificados 44 casos em que não existe valor para esta variável. A distribuição destes casos com valor omisso entre os pacientes é a seguinte:  Ou seja, existe uma maior prevalência nos pacientes que apresentam insuficiência renal crónica. A maior parte dos pacientes apresenta valores nos intervalos: • [68,8 .. 115,6[• [115,6 .. 162,4]. sendo que valores acima dos 110 mg/dl já é considerado um valor anormalmente alto

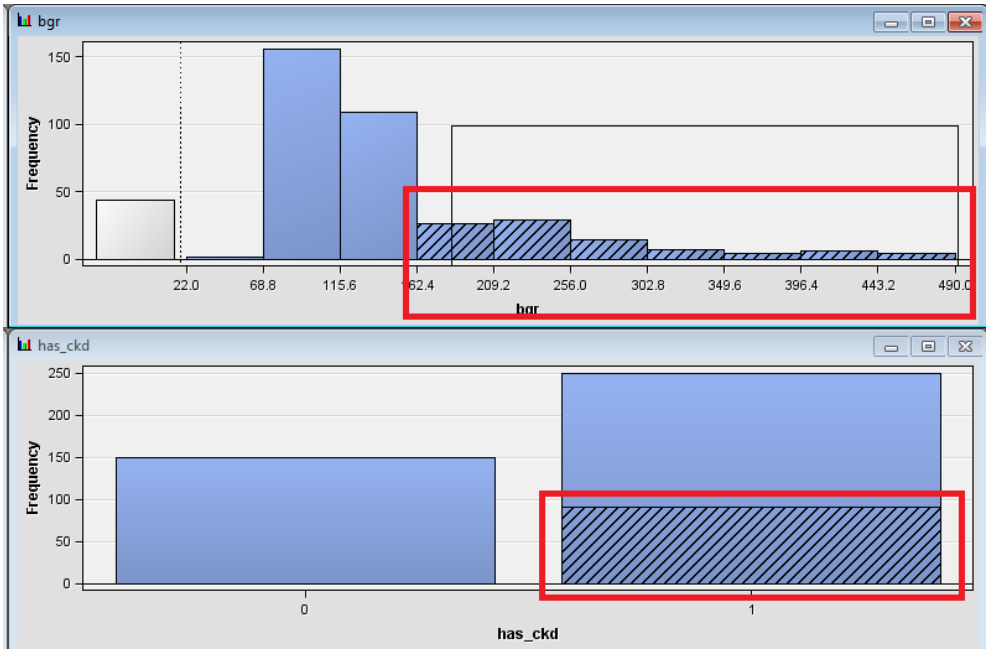
bgr

para o índice de glicemia.

No entanto, não há uma relação clara entre o índice de glicemia já considerado acima do normal e a prevalência da insuficiência renal crónica, pois os pacientes do segundo intervalo encontram-se distribuídos de forma homogénea entre os que apresentam a patologia e aqueles que não a apresentam, como se pode constatar na figura seguinte:

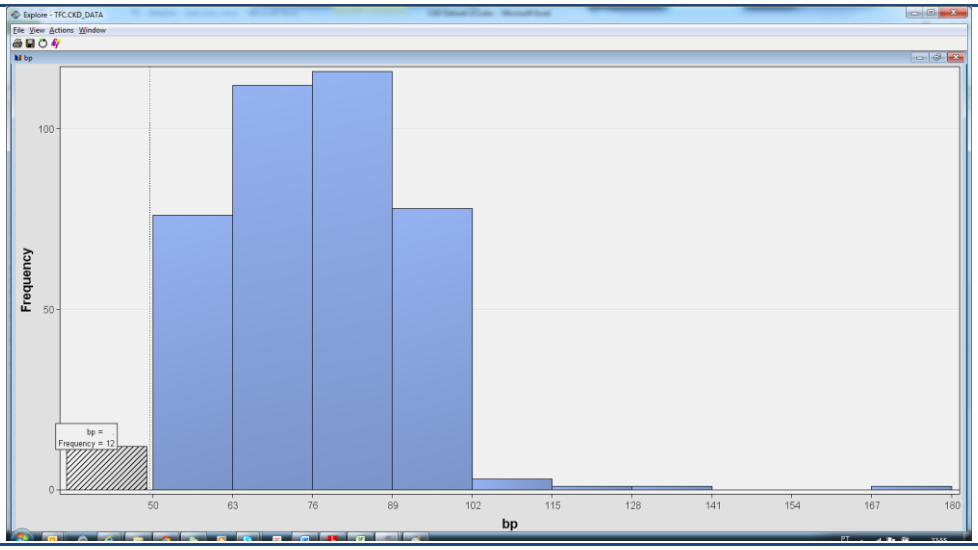


Ainda assim, por comparação com a ocorrência da doença, foi possível constatar que os pacientes com valores de açúcar no sangue superiores 162,4 mg/dl apresentam insuficiência renal crónica:

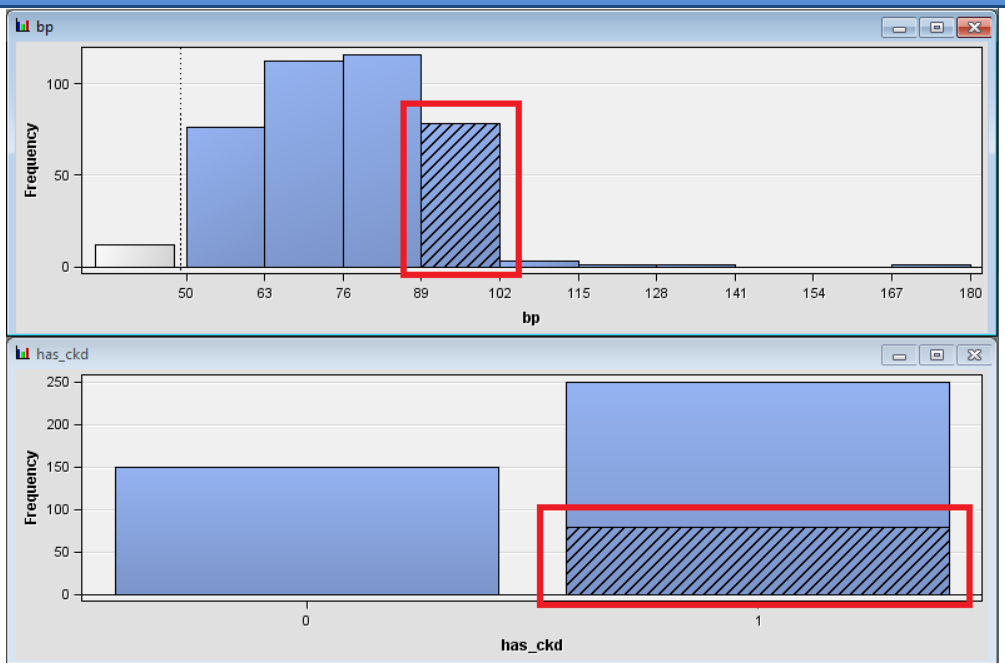


bgr	
	Existe um caso isolado de um paciente que apresenta um valor anormalmente baixo de 22 mg/dl e que apresenta a doença. Do ponto de vista deste data set é um *outlier*.
Decisão	Como neste caso, o valor médio desta variável é aproximadamente 148 mg/dl, optei por atribuir a média aos valores omissos. Dessa forma, os pacientes ficam distribuídos no intervalo [115,6 ; 162,4[em que há pacientes dos dois tipos (com e sem insufifcência renal crónica), tal como se verifica nos pacientes sem valor para esta variável.

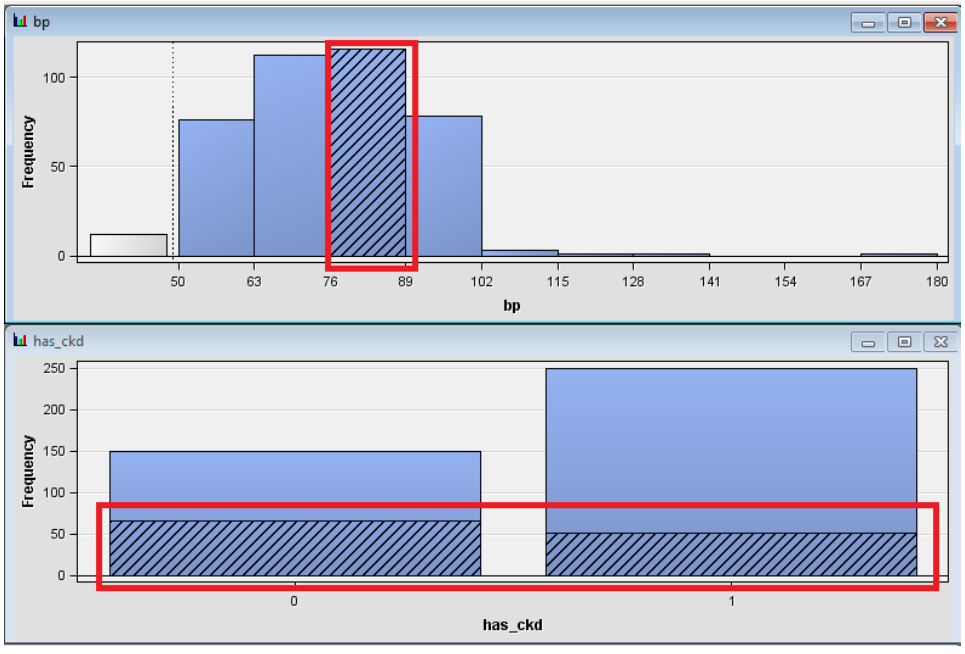
Tensão arterial (bp)

bp	
Descrição	Variável numérica, com indicação da tensão arterial diastólica (ou seja, a tensão medida quando o coração está em repouso, normalmente identificada com o valor mais baixo da medida da tensão arterial).
Histograma	
Situação	Existem 12 pacientes para os quais não sabemos quais os valores de tensão arterial. Pela análise da distribuição, podemos ver que os valores da tensão arterial da maior parte dos pacientes se situa entre 50 e 89 mm/Hg, ou seja valores de tensão diastólica considerados normais. Existem ainda 6 pacientes cuja tensão está acima dos valores considerados normais. Em alguns casos, deverá ser erro de registo, dados os valores anormalmente altos:
Decisão	Pela comparação com a ocorrência da patologia foi possível verificar que para valores acima dos 89 mm/Hg, todos os pacientes apresentam insuficiência renal:

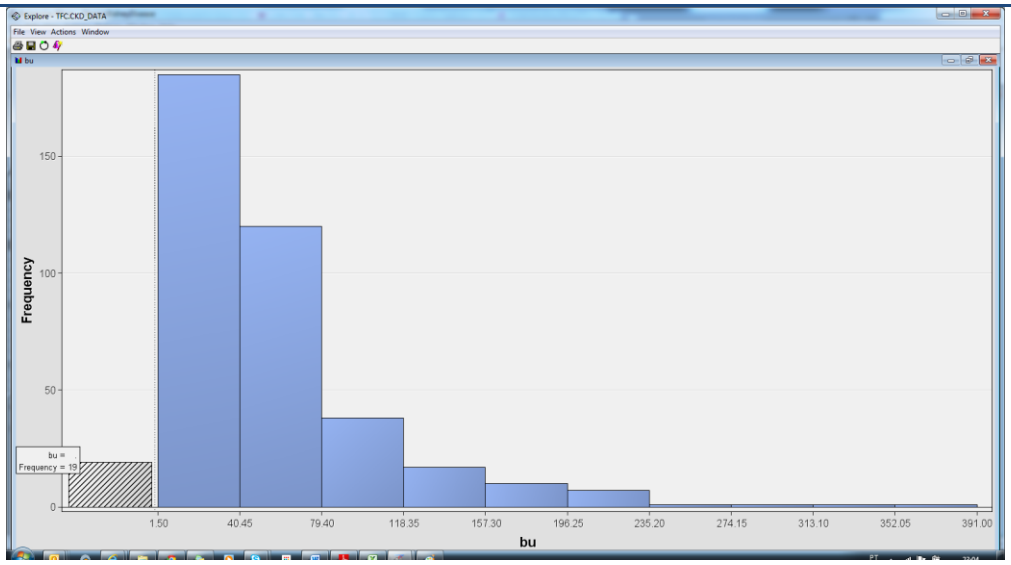
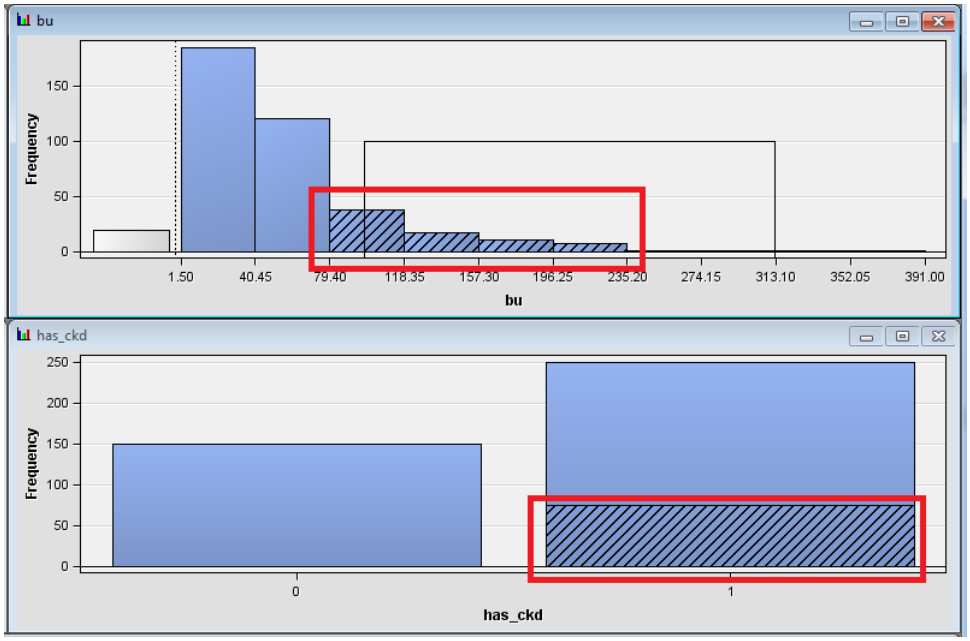
bp

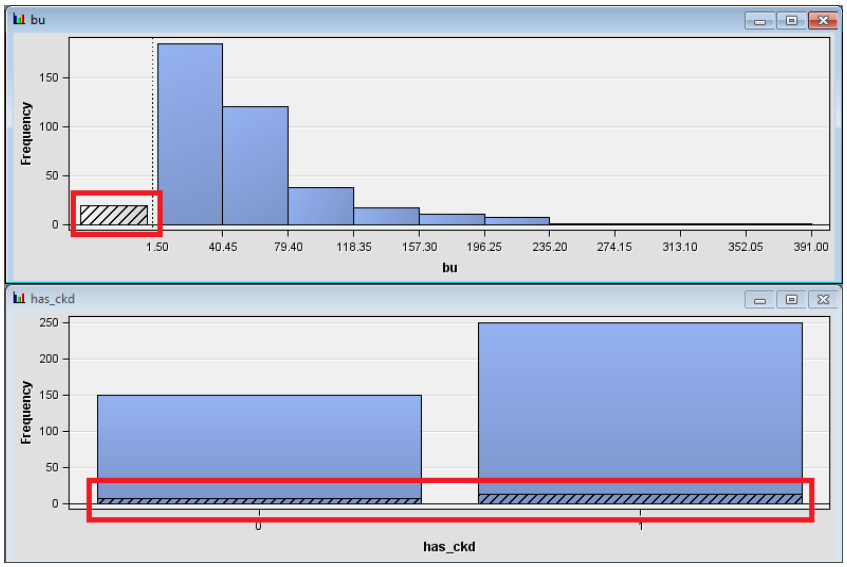


Uma vez que relativamente aos valores omissos, existem alguns que apresentam a doença e outros que não, decidi assumir para estes valores omissões o valor médio encontrado que se situa nos 77 mm/Hg e em que os pacientes são dos dois tipos:

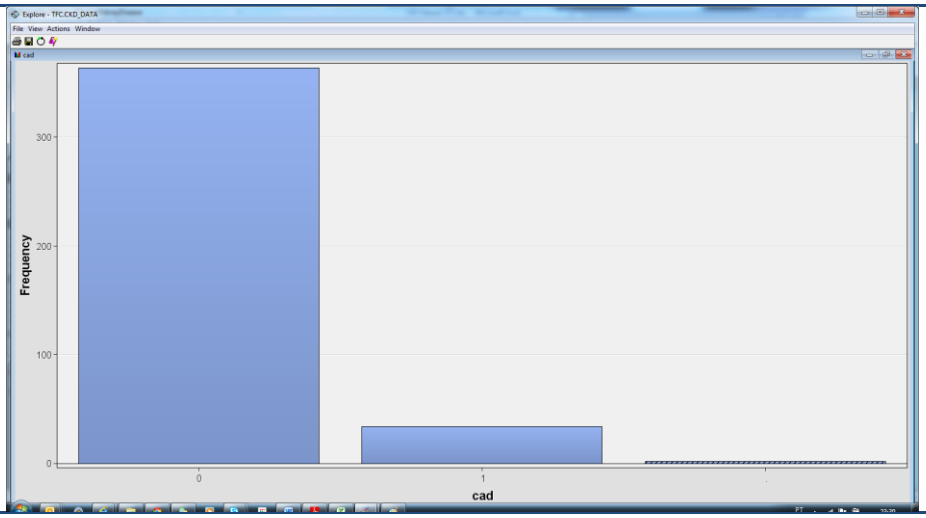


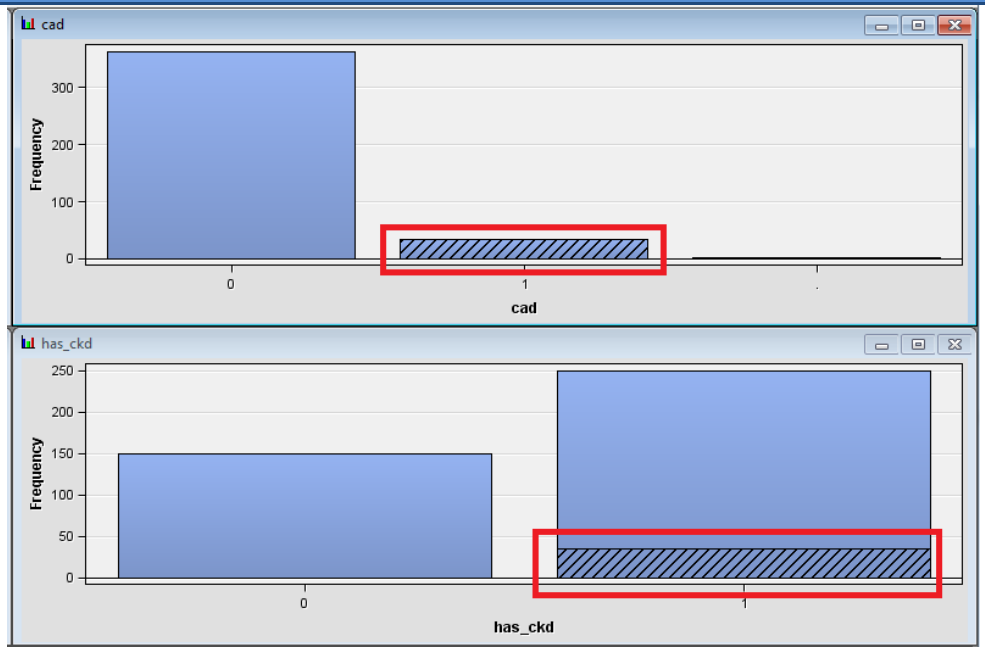
Ureia no sangue (bu)

bu	
Descrição	Variável numérica, contínua, medida em mg/dl que indica a quantidade de ureia presente no sangue dos pacientes. Sendo a ureia um produto tóxico resultante do metabolismo das proteínas, é normalmente excretada pelos rins. Se a função renal estiver comprometida, a ureia acumula-se no organismo e pode causar danos graves.
Histograma	
Situação	Os valores considerados normais situam-se entre 3,6 e 8,3 mg/dl. Neste caso, existem 19 pacientes para os quais não temos informação sobre os valores de ureia no sangue. Por comparação com o *target*, foi possível constatar que os pacientes com valores acima dos 79,4 mm/dl apresentam a doença. 
Decisão	Uma vez que os pacientes com valores omissões estão distribuídos entre os que apresentam a doença e os que não a apresentam, neste caso também optei por preencher os valores omissos com a média que se situa nos 52,74 mm/dl.

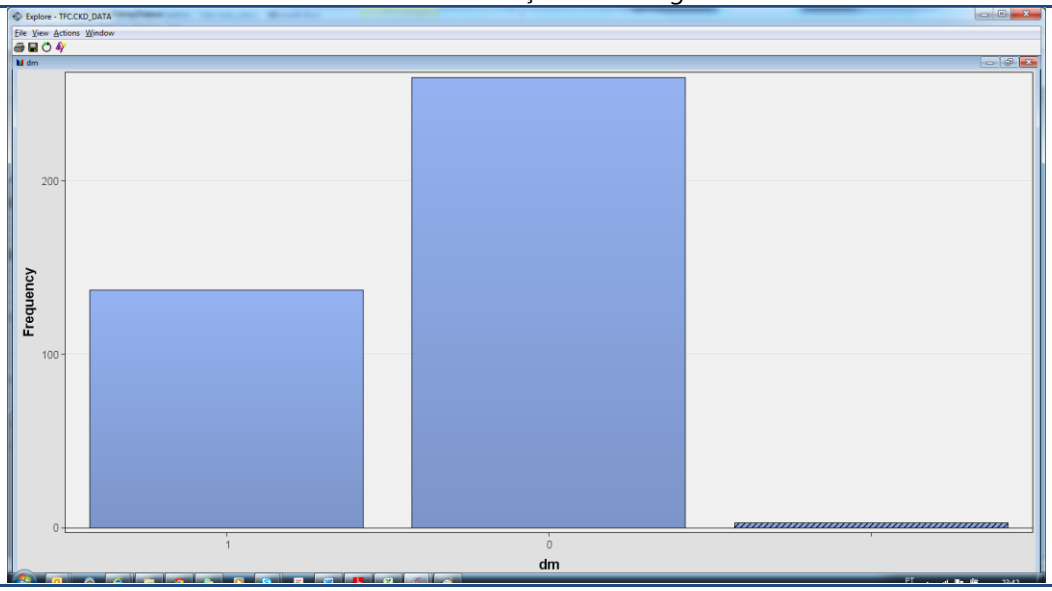
bu	
	

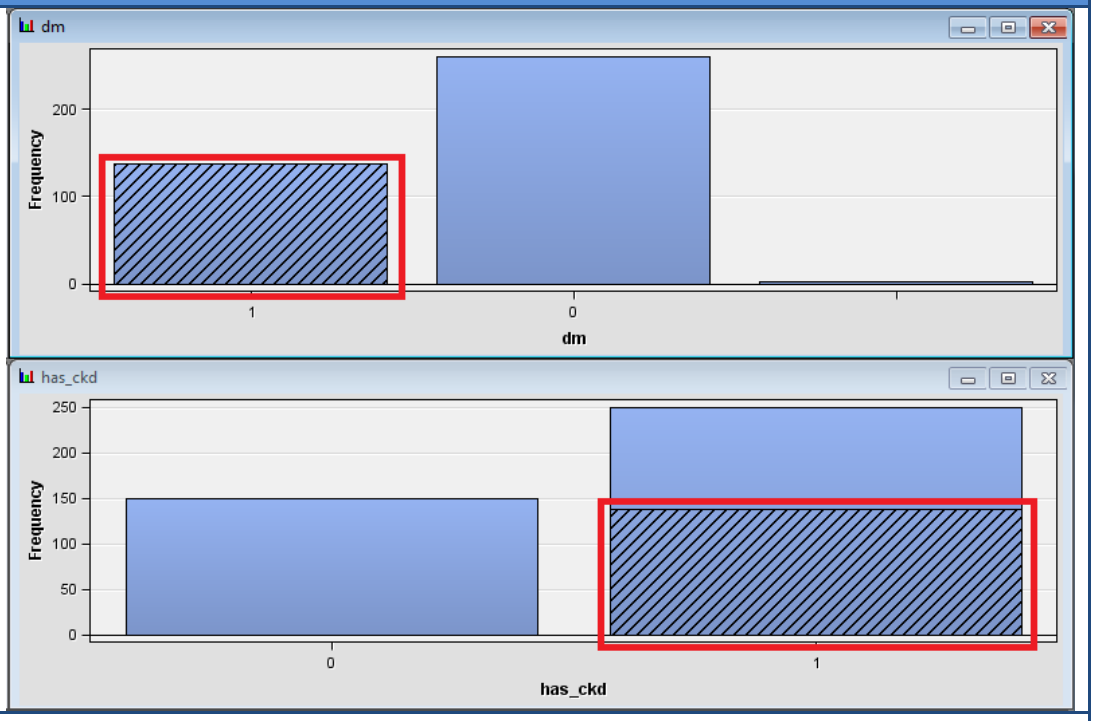
Doença coronária (cad)

cad	
Descrição	Variável binária, indicativa da presença de doença coronária, que resulta da degradação dos vasos sanguíneos do músculo cardíaco.
Histograma	
Situação	Neste caso, a maior parte dos pacientes não apresenta doença coronária. Existem 2 pacientes sem valor para esta variável.
Decisão	No entanto foi possível constatar que todos os pacientes que apresentam doença coronária, também apresentam a patologia, conforme se pode constatar na figura seguinte:

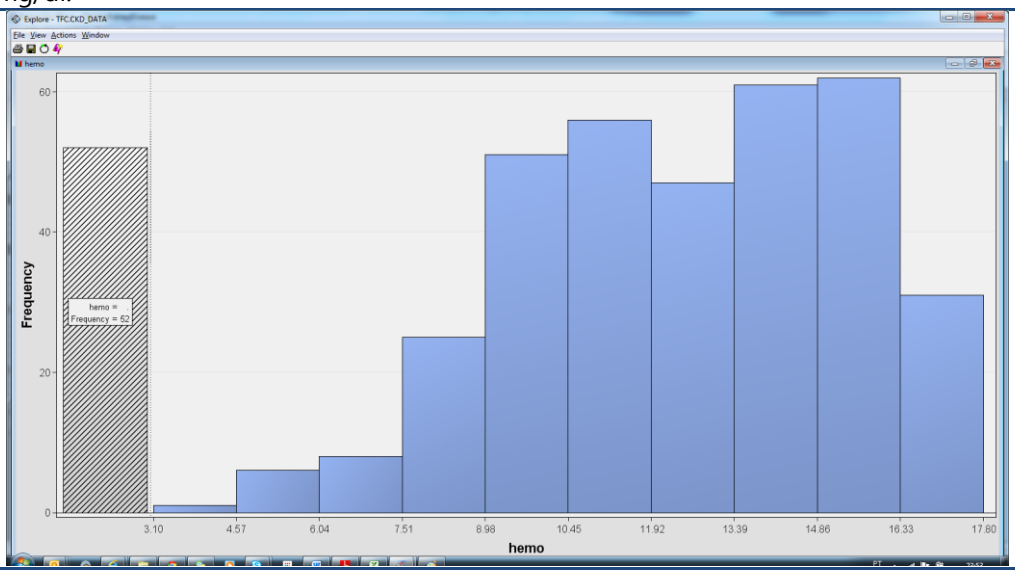
cad	
	 Assim, uma vez que os dois pacientes sem valor para esta variável não apresentam insuficiência renal, optei por preencher os valores omissos com 0 (ou seja, sem doença coronária.)

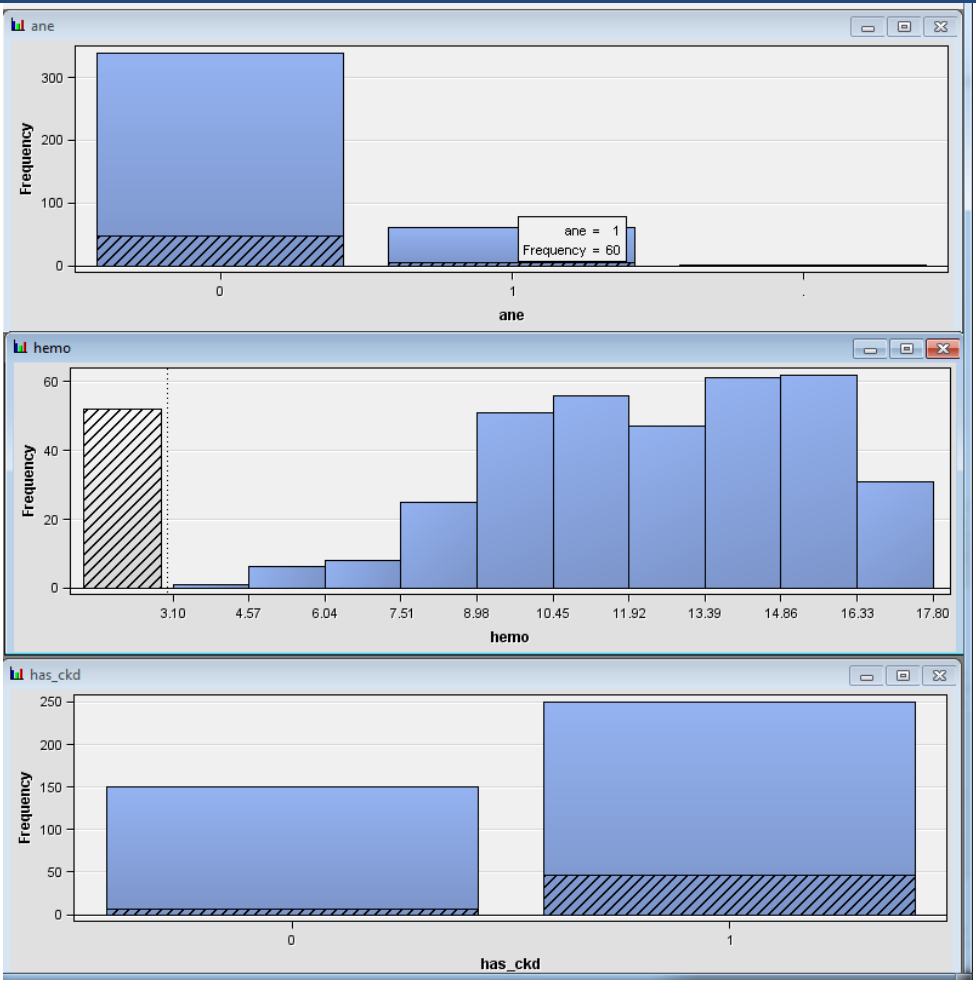
Diabetes Mellitus (dm)

dm	
Descrição	Variável binária indicativa de diabetes. A diabetes é uma doença metabólica com origens diversas e consiste no aumento anormal de açúcar no sangue.
Histograma	
Situação	Para esta variável existem 3 pacientes para os quais não sabemos se apresentam diabetes ou não. Por comparação os pacientes com insuficiência renal, foi possível constatar que todos os pacientes com diabetes também apresentam insuficiência renal, como se pode constatar na figura seguinte:

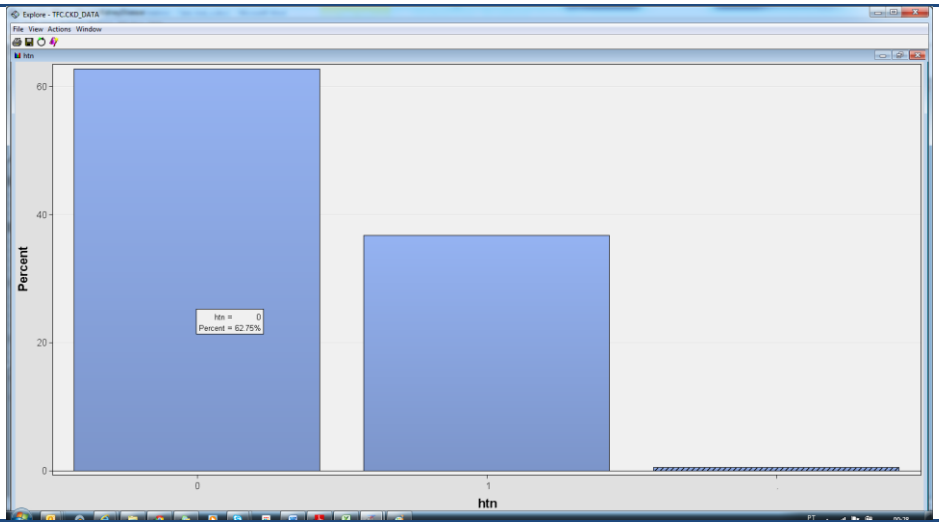
dm	
	
Decisão	Uma vez que os valores omissos correspondem a pacientes que não apresentam insuficiência renal, optei por considerar que estes pacientes também não apresentavam diabetes. Assim, neste caso, a variável foi preenchida com 0.

Hemoglobina (hemo)

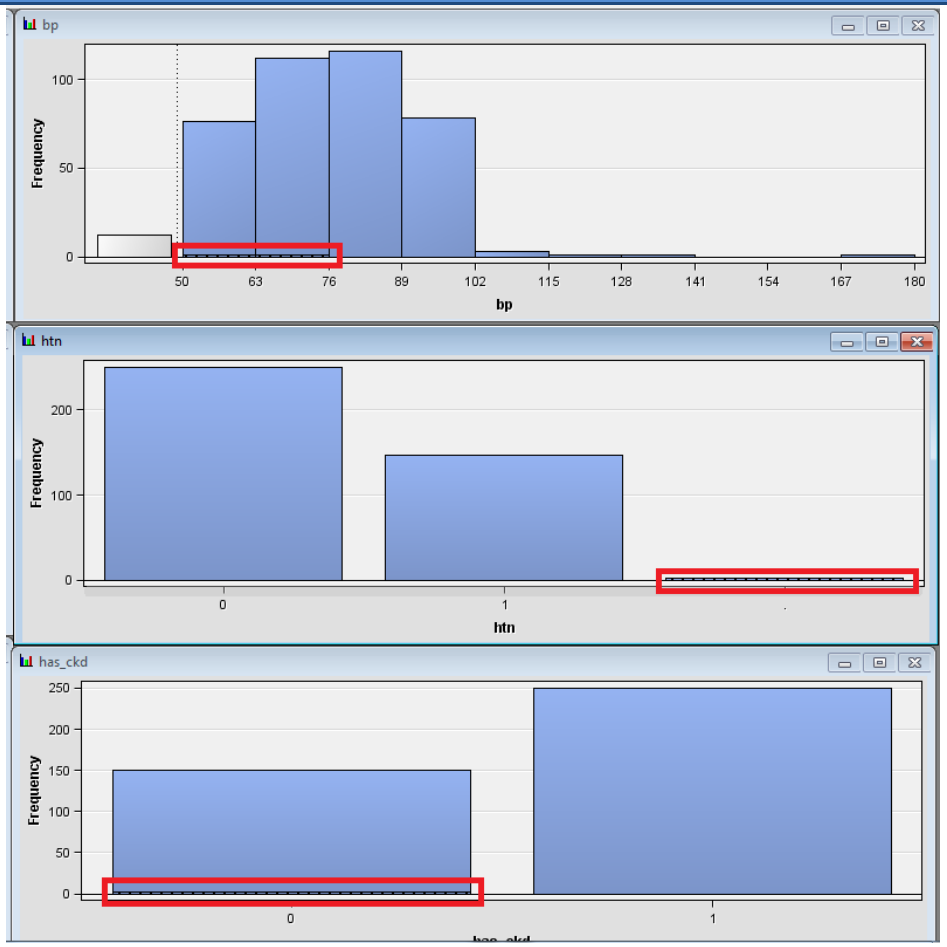
hemo	
Descrição	Variável numérica, com os valores de hemoglobina apresentados pelos pacientes, em mg/dl.
Histograma	
Situação	Para esta variável, existem 52 pacientes que não apresentam qualquer valor para esta variável. É um valor bastante elevado, face à amostra. No entanto, como os valores de hemoglobina no sangue estão relacionados com a anemia, optei por comparar estas duas variáveis e analisar a relação com a variável *Target*:

hemo	
	
Decisão	De todos os casos omissos, apenas um deles apresenta anemia. Assim, optei por preencher os valores omissos com a média dos valores de hemoglobina, ou seja, 12,52 gms.

Hipertensão (htn)

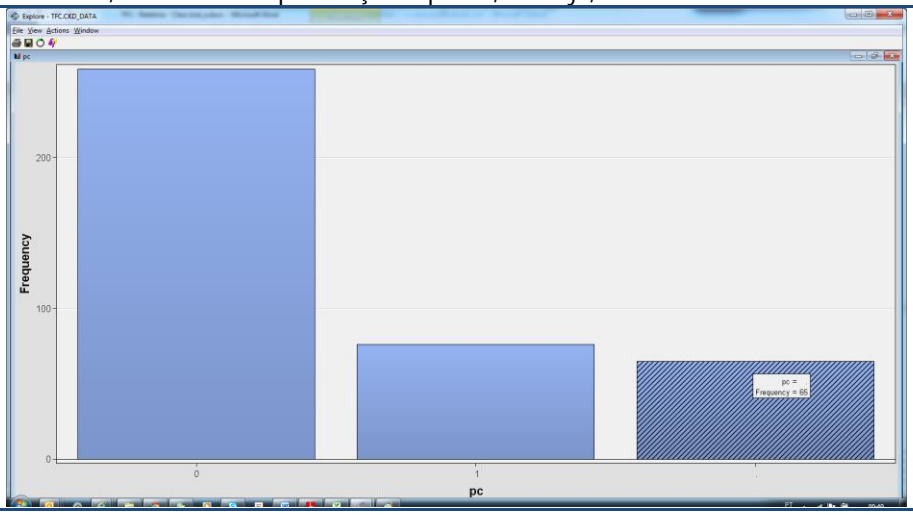
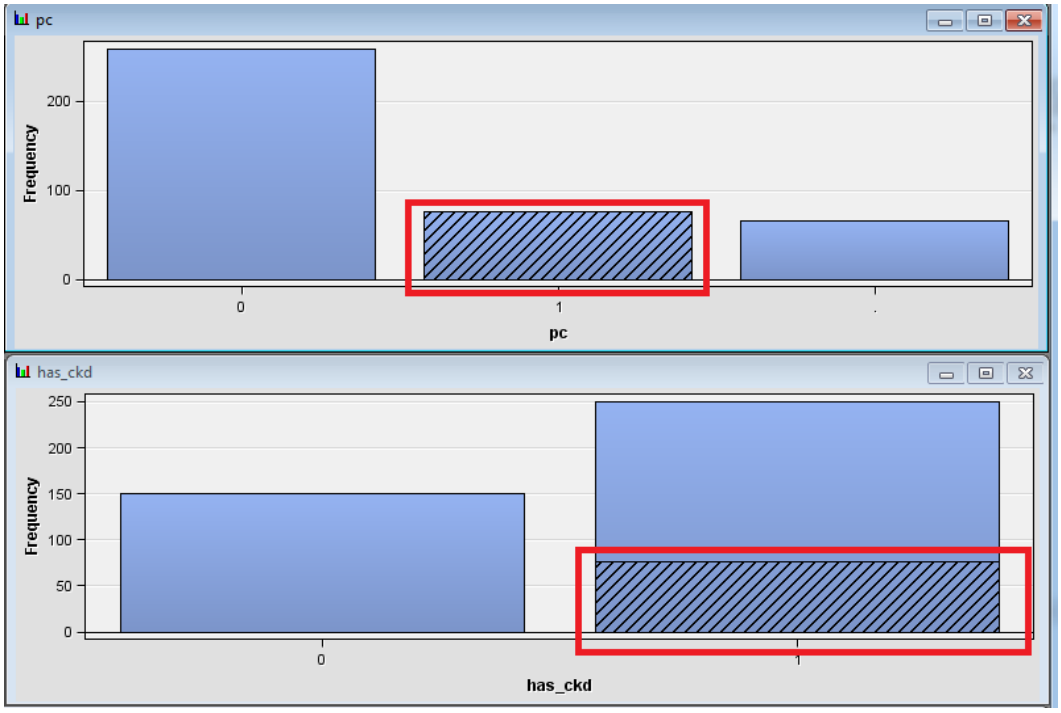
htn	
Descrição	Variável binária, indicativa se o paciente apresenta hipertensão.
Situação	Pela análise do gráfico, é possível constatar que cerca de 63% dos pacientes não apresenta hipertensão. Também nesta variável existem dois pacientes que não têm qualquer valor para a hipertensão.
Histograma	
Decisão	Uma vez que a hipertensão está relacionada com a pressão arterial, optei por comparar os pacientes com valores omissos, com a variável relativa à pressão arterial. Como se pode constatar, os pacientes sem valor para a hipertensão (gráfico do meio) estão relacionados com os pacientes com baixos valores de pressão arterial.

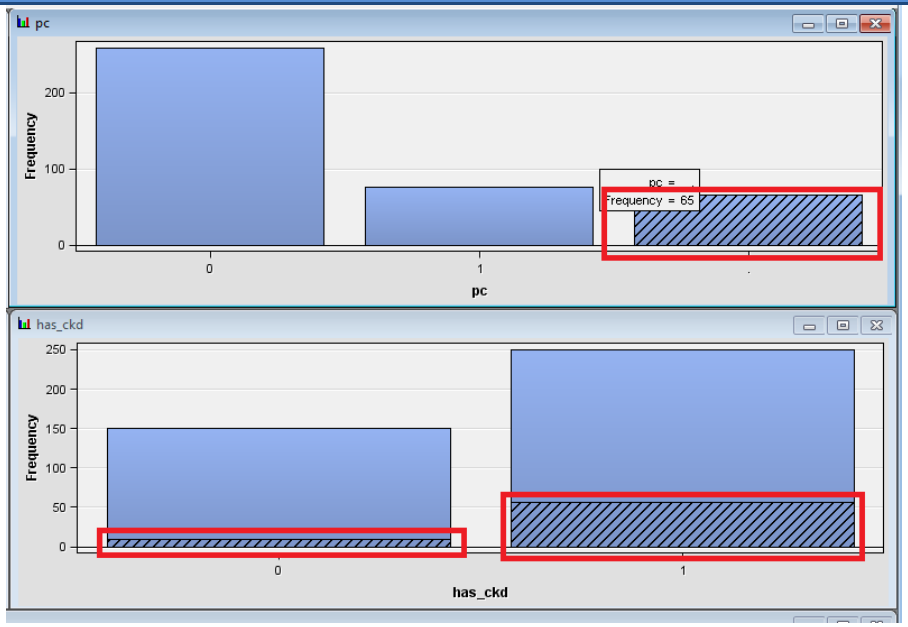
htn



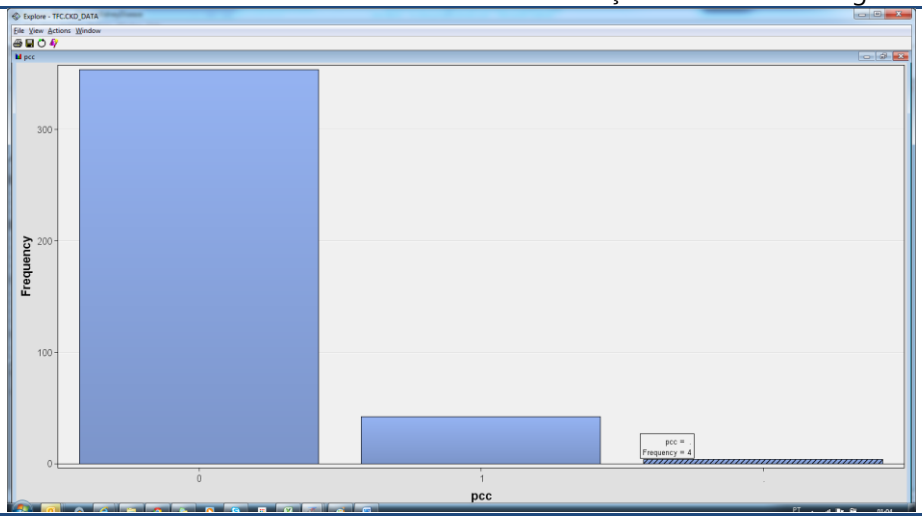
Assim, optei por considerar que estes pacientes não tinham hipertensão, assumindo 0 para os valores desta variável.

Piúria (pc)

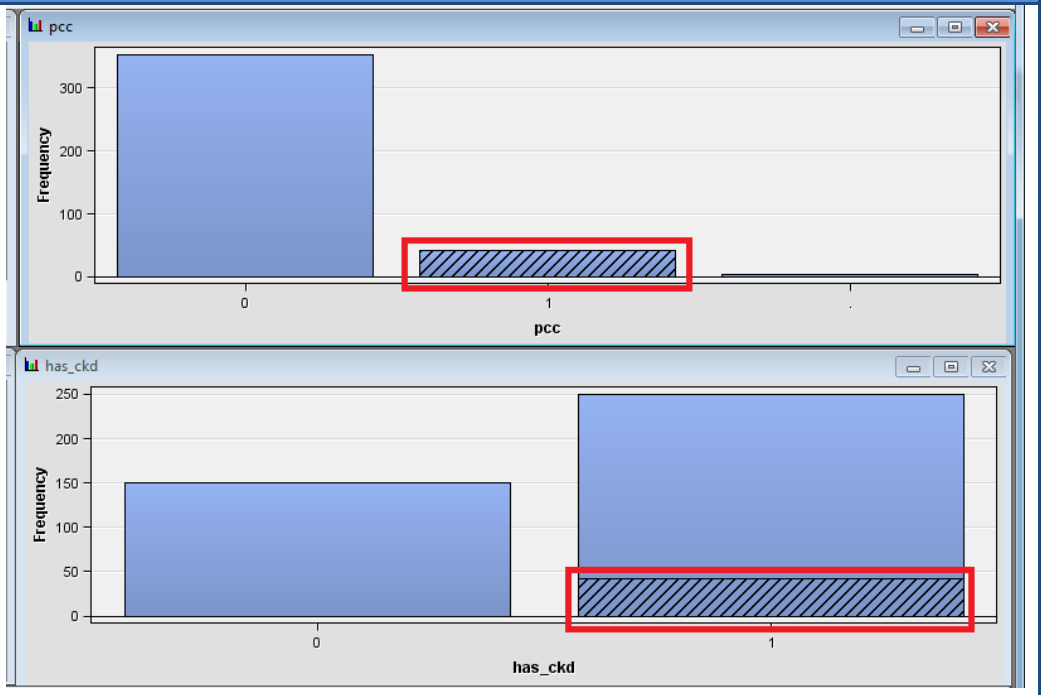
pc	
Descrição	Variável binária, indicativa da presença de piúria, ou seja, leucócitos na urina.
Histograma	
Situação	Pela análise do gráfico, foi possível verificar que existe um elevado número de pacientes (66) para os quais não foi indicado nenhum valor para a presença de piúria.
Decisão	Comparando os valores com o *target*, foi possível concluir que todos os pacientes com piúria apresentam insuficiência renal:  No entanto, relativamente aos valores omissos, não foi possível tirar concluões, uma vez que estes estão distribuídos entre os pacientes com e sem insuficência renal. Existe uma maior percentagem de pacientes que apresentam insuficiência renal, mas não é suficientemente significativa permitir tomar uma decisão relativamente aos valores omissos.

pc	
	

Aglomerados de leucócitos (pcc)

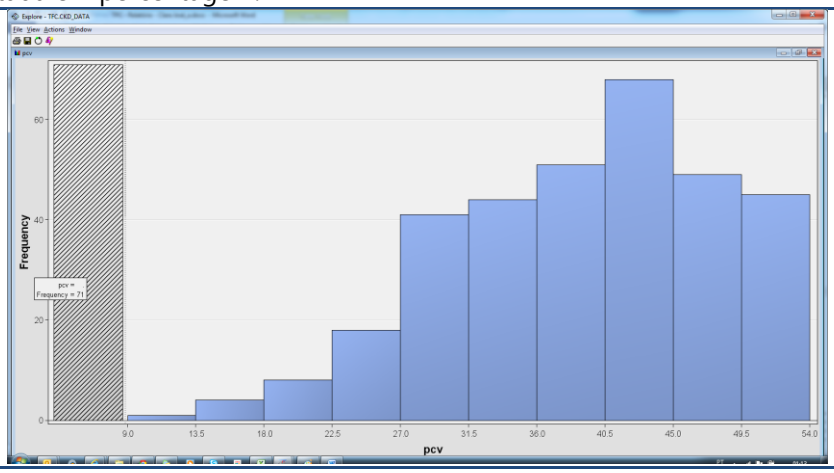
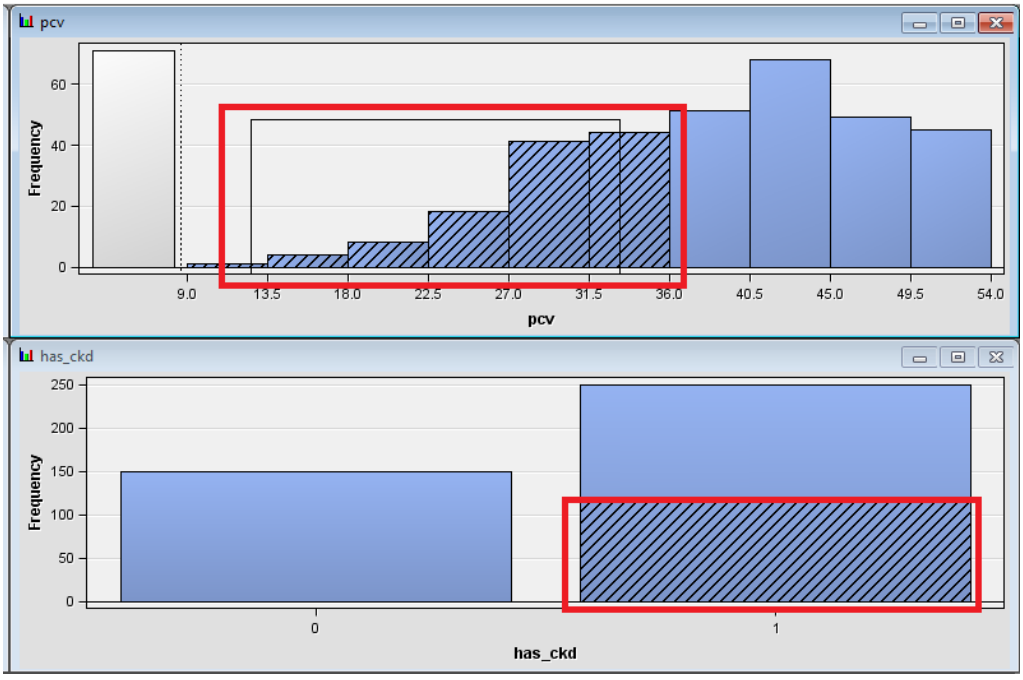
pcc	
Descrição	Variável binária indicativa da presença de aglomerados de leucócitos na urina. Os aglomerados de leucócitos são indicativos de uma infecção renal bacteriana grave.
Histograma	
Situação	Como se pode ver pelo histograma, a maior parte dos pacientes não apresenta aglomerados de leucócitos Existem 4 pacientes sem valor para esta variável.
Decisão	Pela comparação com o *target* foi possível concluir que todos os pacientes que apresentam aglomerados de leucócitos, também apresentam insuficiência renal, o que configura um quadro clínico bastante grave

pcc

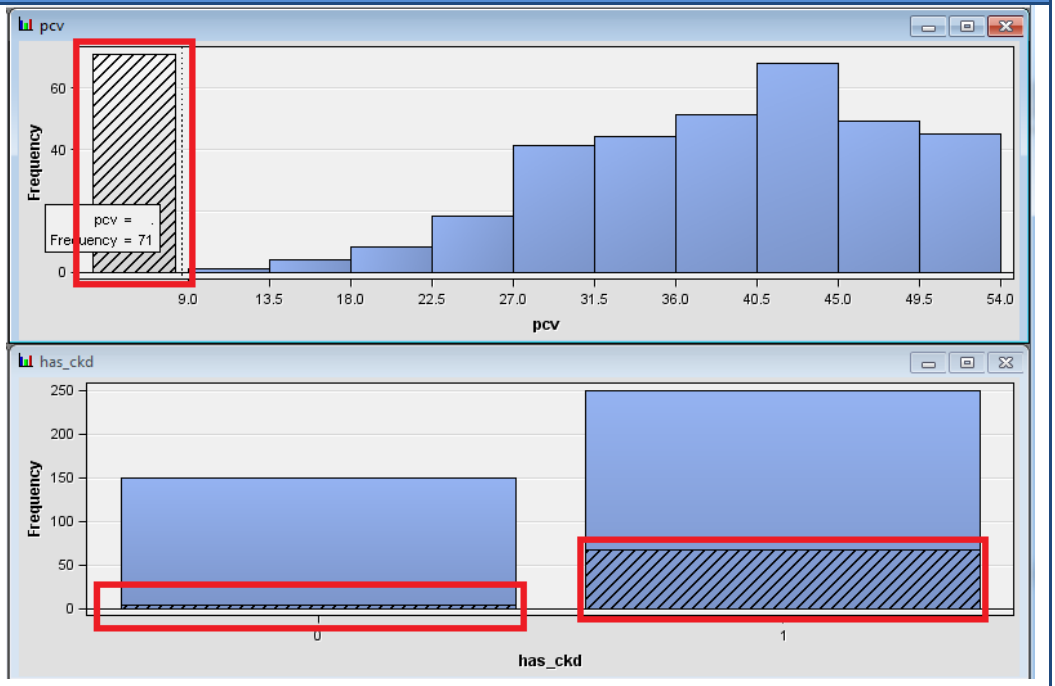


Assim, e uma vez que os pacientes sem valor para esta variável não são insuficientes renais, optei por considerar que não apresentavam aglomerados de leucócitos na urina, atribuindo o valor 0 a esta variável.

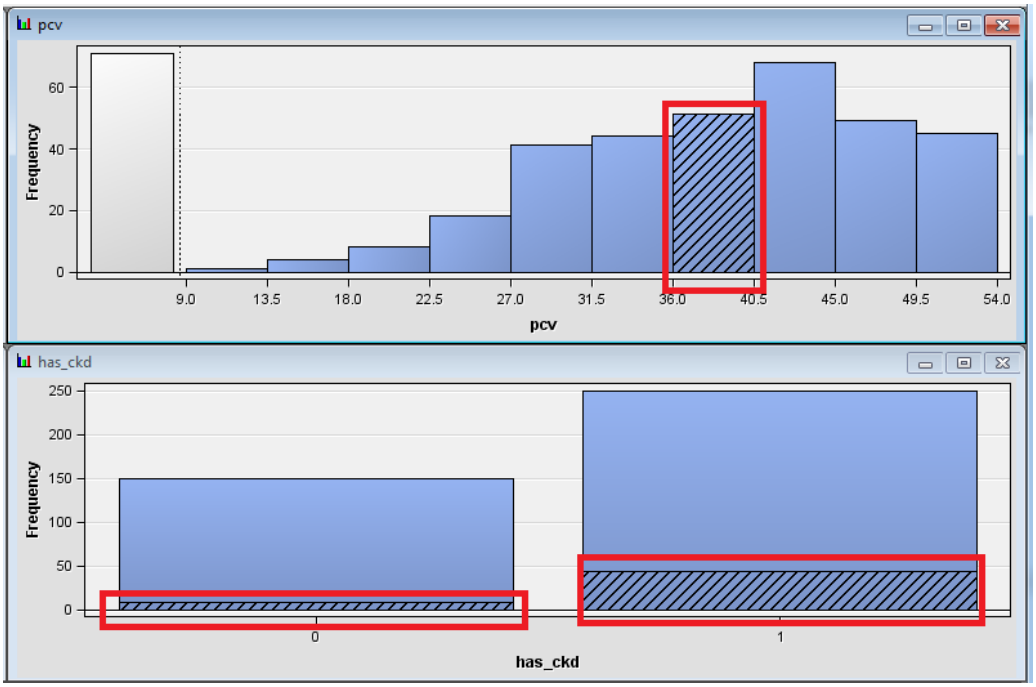
Volume globular médio (pcv)

pcv	
Descrição	Variável numérica, contínua relativa ao tamanho dos glóbulos vermelhos no sangue, apresentada em percentagem.
Histograma	
Situação	Pela análise da distribuição dos valores, podemos verificar que existem 71 casos sem valor para esta variável. Pela comparação com o *target* verifica-se que para valores conhecidos até 36 % (que é considerado um valor abaixo do normal) os pacientes apresentam insuficiência renal. 
Decisão	Por comparação com o target, a maior parte dos pacientes sem valor nesta variavel apresentam insuficiencia renal:

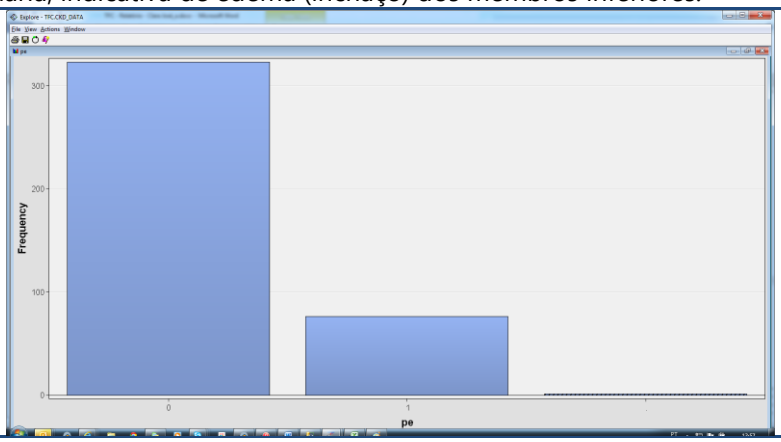
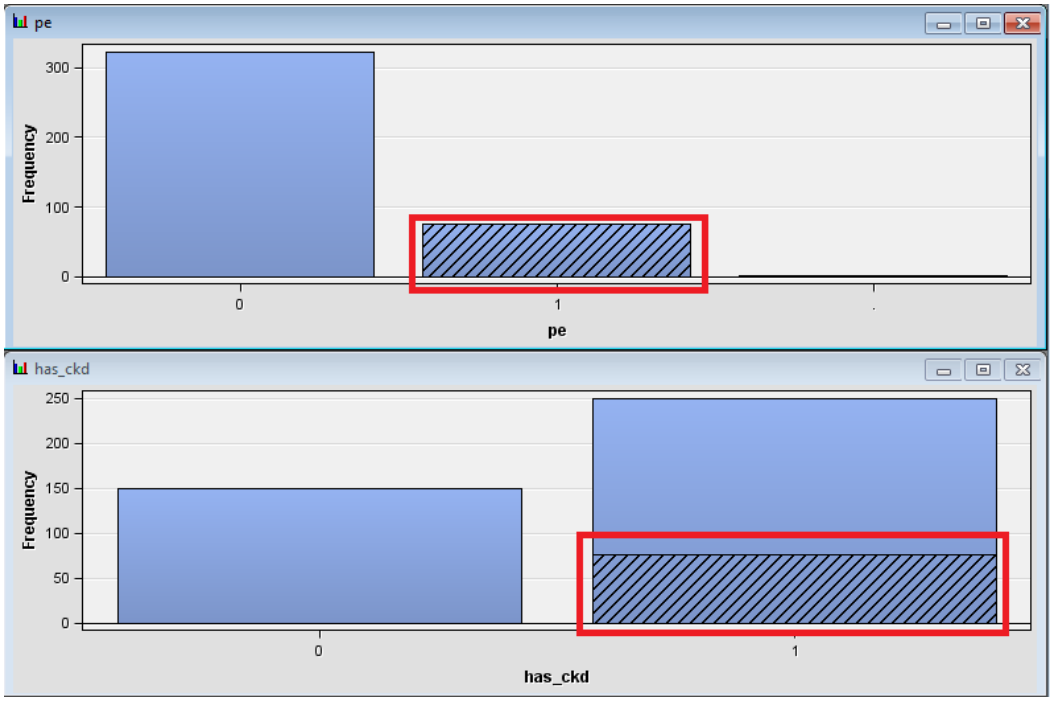
pcv



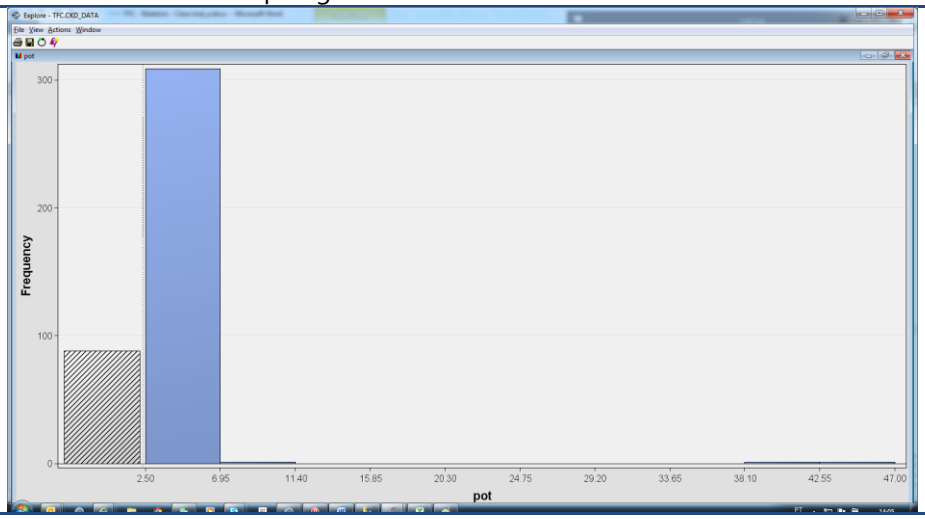
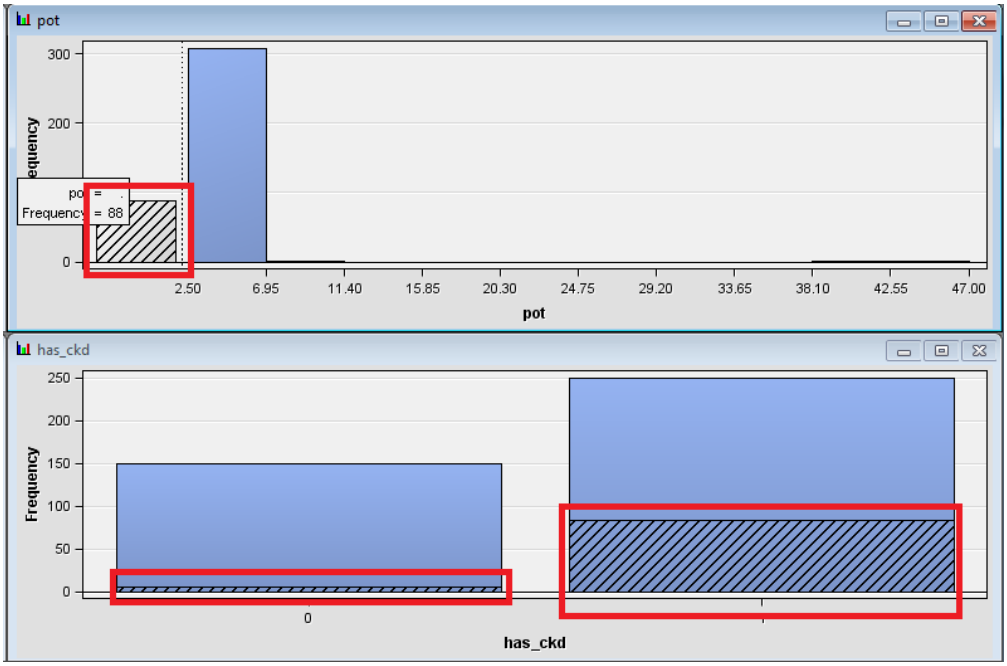
Ainda assim, existem alguns que não apresentam a patologia, pelo decidi preencher os valores omissos com o valor médio que ronda 39 %. Desta forma, os casos vão estar no intervalo [36; 40,5[que apresenta uma distribuição de casos mais de acordo a dos valores omissos. Na figura seguinte pode ver-se que neste intervalo existem pacientes sem insuficiência renal crónica:

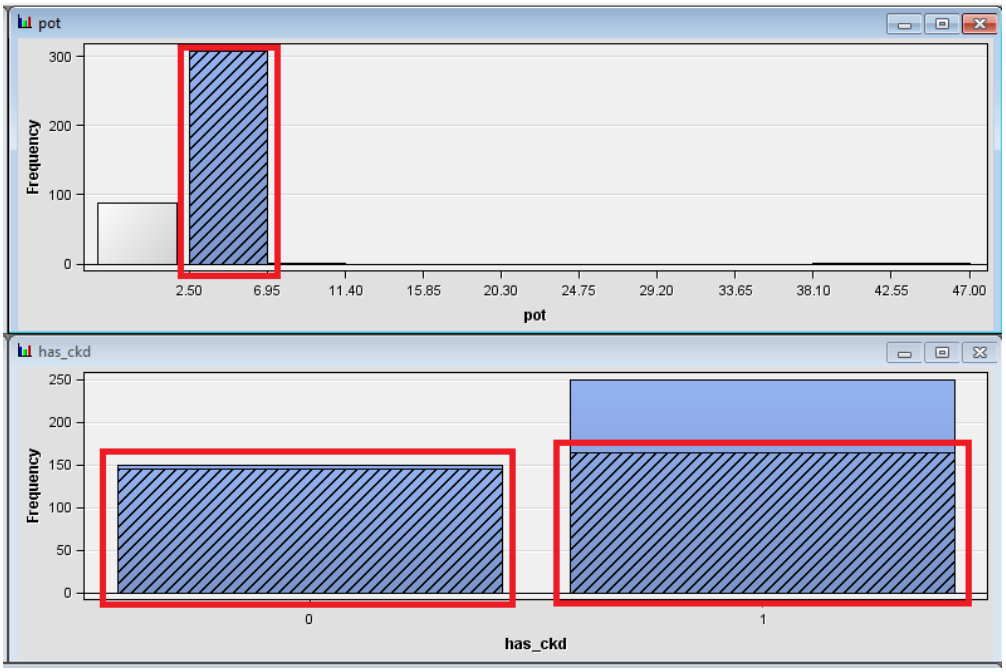


Edema dos membros inferiores (pe)

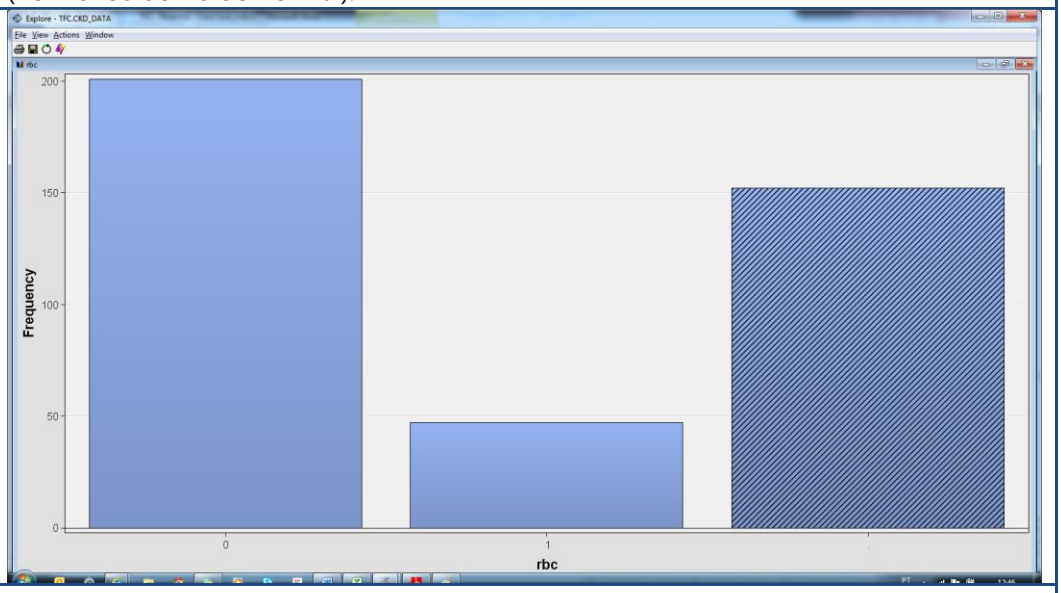
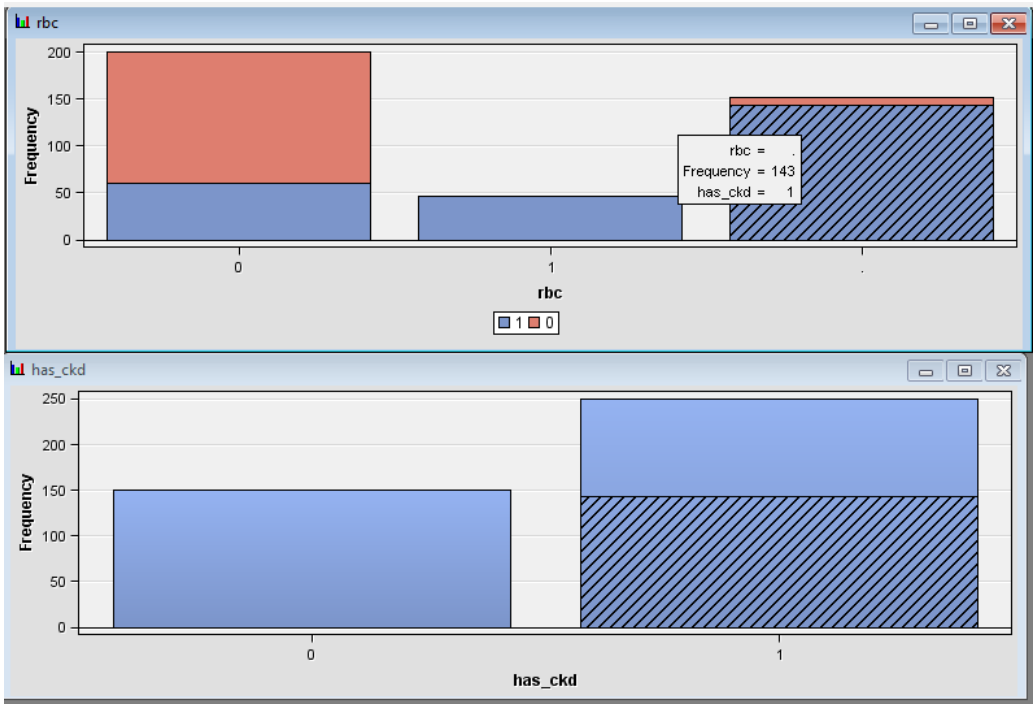
pe	
Descrição	Variável binária, indicativa de edema (inchaço) dos membros inferiores.
Histograma	
Situação	Para esta variável existe um único paciente em relação ao qual não conhecemos a situação.
Decisão	Por comparação com o target todos os pacientes com edemo dos membros inferiores, apresentam insuficiência renal crónica.  Assim, e uma vez que este paciente não apresenta insuficnência renal, optei por considerar que o valor para esta variável é 0, ou seja, sem edema.

Potássio (pot)

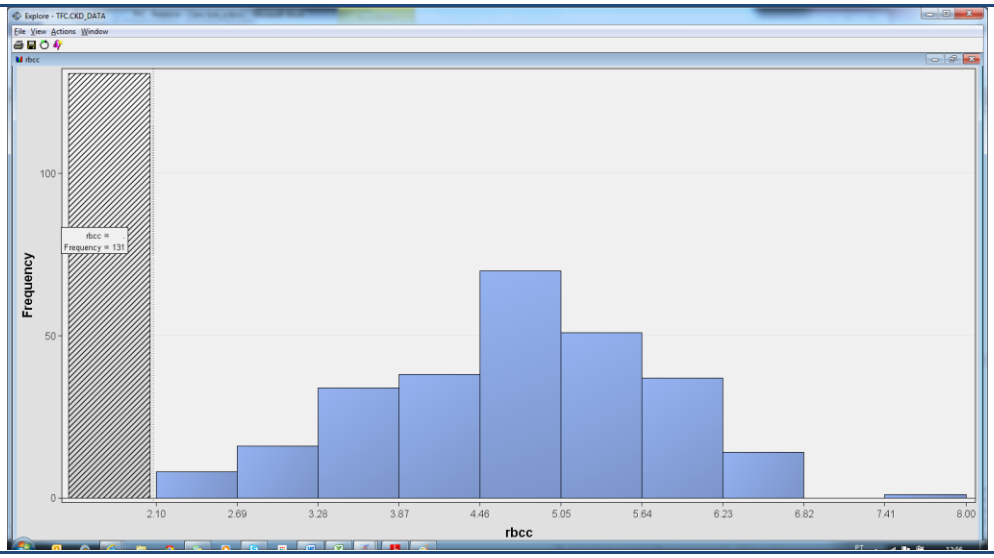
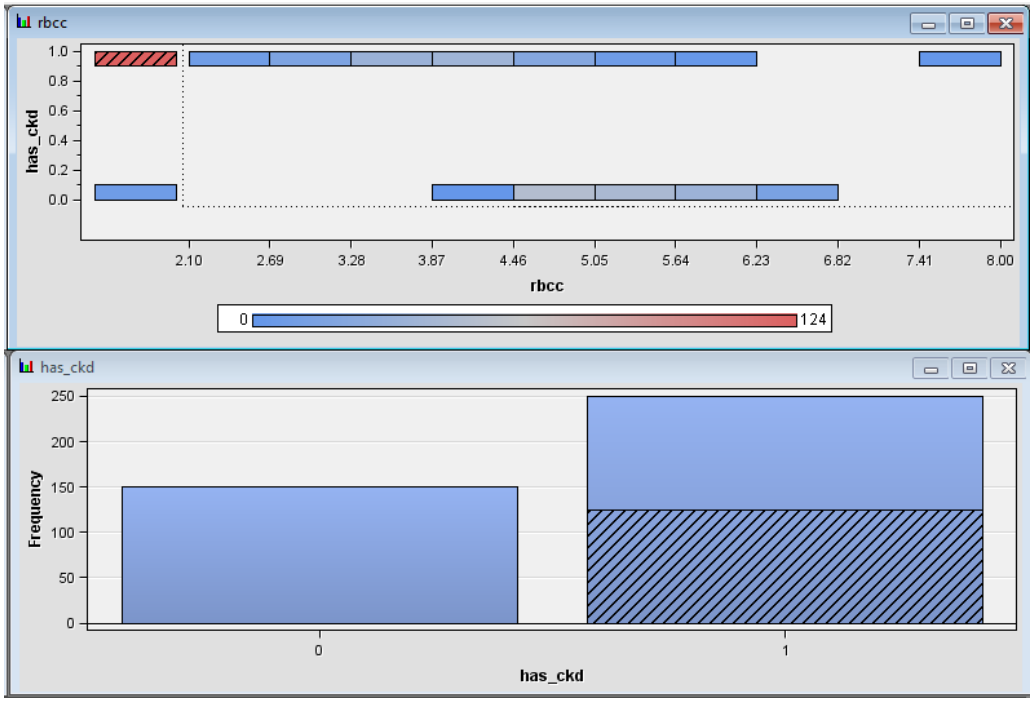
pot	
Descrição	Quantidade de potássio presente no sangue, medido em miliequivalentes por litro, correspondente à concentração de potássio no sangue. O potássio em excesso é excretado pelos rins, mas nas situações em que a função renal está comprometida, o potássio acumula-se no organismo e torna-se tóxico, podendo até levar a uma paragem cardíaca.
Histograma	
Situação	Pela análise do histograma, existem 88 pacientes para os quais não sabemos o valor de potássio que apresentam nas análises:  Podemos ver que a maior parte dos pacientes sem valor para esta variável, apresenta insuficiência renal.
Decisão	Para a maior parte dos pacientes (77%) o seus valores de potássio situam-se no intervalo entre 2,50 e 6,9 mEq/l, ou seja, dentro dos valores de referência considerados normais.

pot	
	Existem 3 com valores fora deste intervalo e dois deles apresentam valores bastante elevados, que parecem corresponder a erros de registo. Por comparação com a prevalência de insuficiência renal, no intervalo com maior número de pacientes, podemos ver que a distribuição dos pacientes é linear:  Assim, para este caso, optámos por escolher a média do valor de potássio para preenchimento dos valores omissos (~4,63 mEq/l)

Glóbulos vermelhos na urina (rbc)

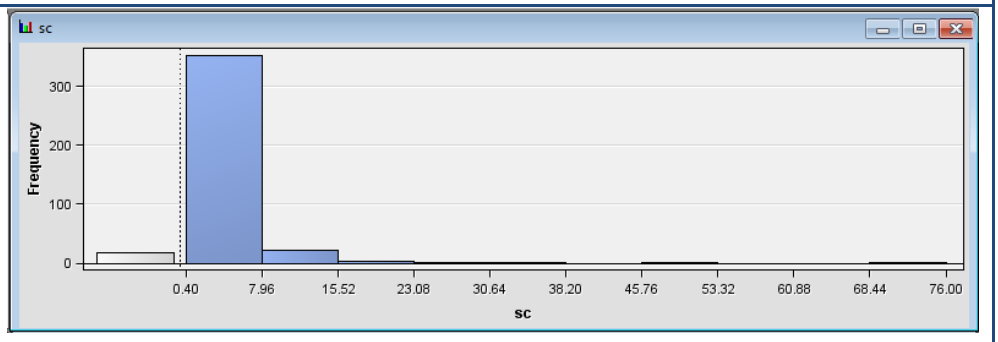
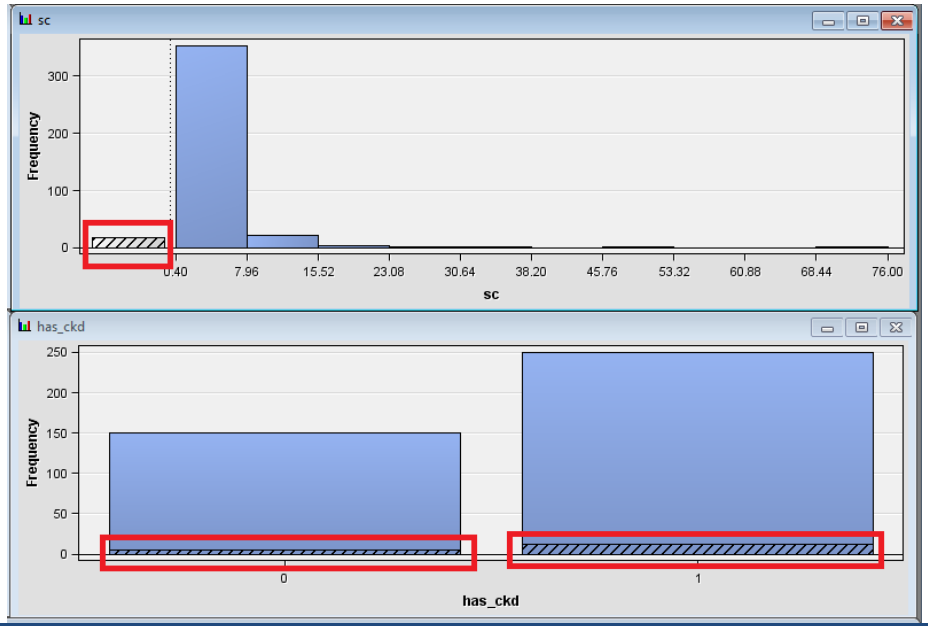
rbc	
Descrição	Variável binária indicativa da quantidade glóbulos vermelhos presentes na urina (normal ou acima do normal).
Histograma	
Situação	Para esta variável existem 152 pacientes para os quais não sabemos a situação, porque aparecem sem valor.
Decisão	Por comparação com a prevalência da patologia, pode-se verificar que a maior parte das situações apresenta insuficência renal:  Ou seja, dos 152 pacientes, 143 apresentam insuficiência renal. Assim, para este caso, decidi preencher os valores omissos com o valor máximo da variável, ou seja, 1, porque todos os que apresentam glóbulos vermelhos na urina em quantidade acima do normal, apresentam também insuficiência renal.

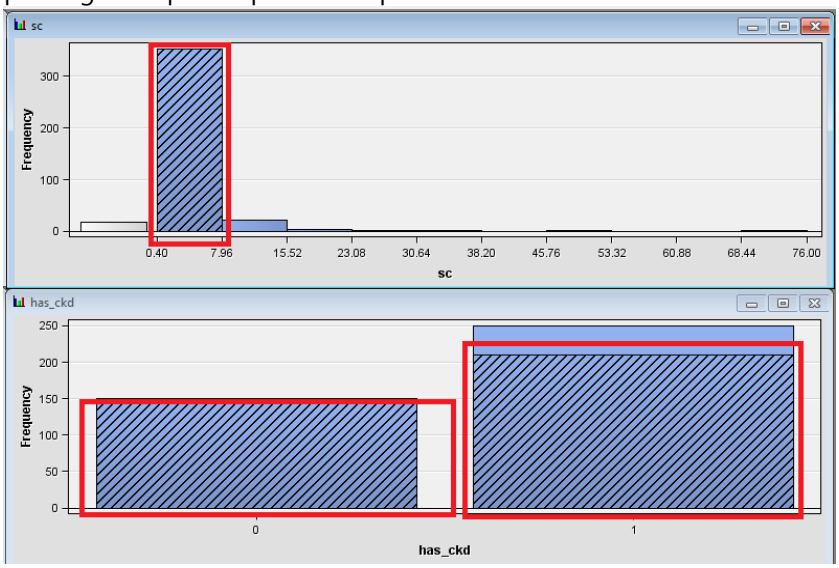
Contagem de glóbulos vermelhos (rbcc)

rbcc	
Descrição	Variável numérica, contínua correspondente ao número de glóbulos vermelhos presentes no sangue, medida em milhões por cm^3.
Histograma	
Situação	Para esta variável existem 131 paciente para os quais não sabemos o valor dos glóbulos vermelhos
Decisão	Por comparação com a prevalência da patologia, foi opossivel determinar que para valores muito baixos (até ao limite inferior do intervalo de maior frequência), os pacientes apresentam insuficiência renal. Por comparação, dos 131 pacientes pacientes foi possível determinar que 124 apresentam insuficência renal e 7 não apresentam.  Assim, em termos do valor a usar para preenchimento dos valores omissos, optei por

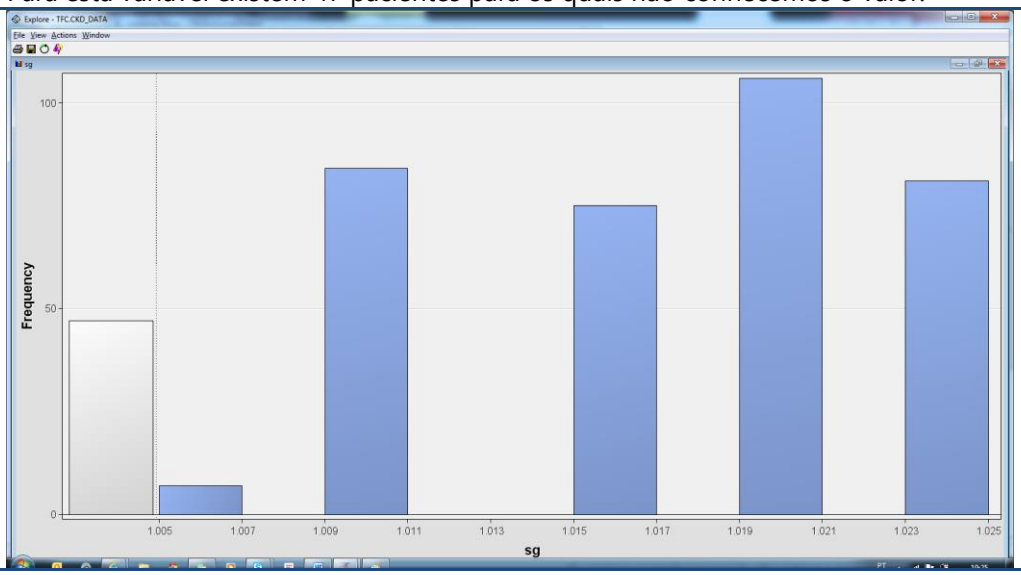
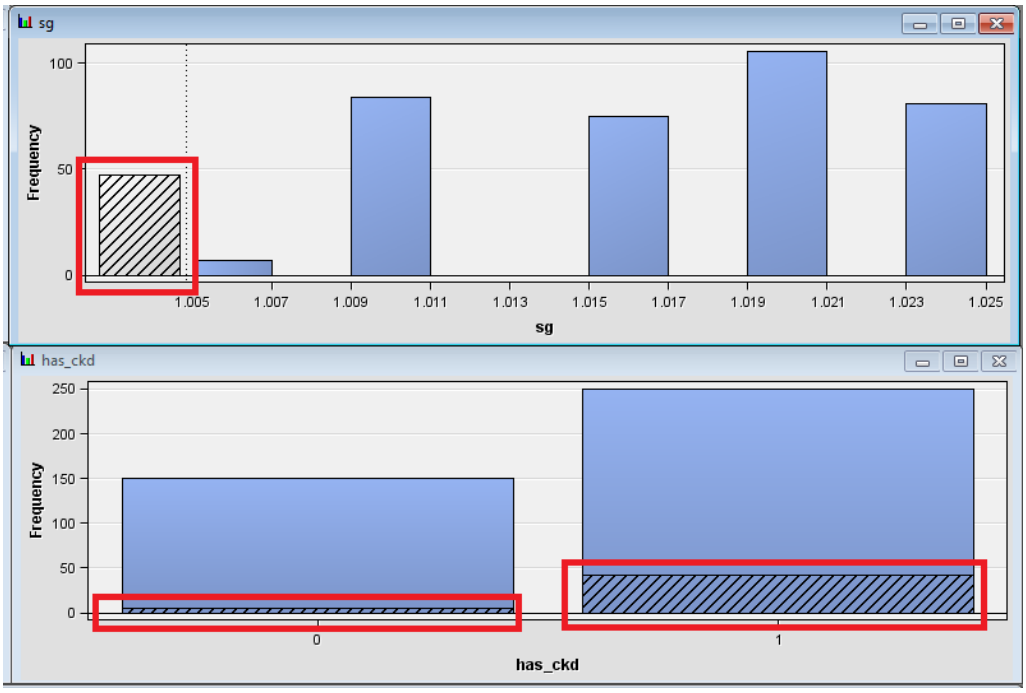
rbcc	
	preencher com o valor mínimo, dado que a maior parte dos pacientes com valores omissos apresentam a patologia.

Creatinina (sc)

sc	
Descrição	Variável numérica, contínua indicativa da quantidade de creatinina presente no sangue. A creatinina é um subproduto do consumo de creatina pelos músculos, o que significa que está diretamente relacionada com a massa muscular e a atividade física. No entanto, os valores de referência considerados normais para estão entre 0,6 mg/dl e 1,3 mg/dl.
Histograma	
Situação	Para esta variável existem 17 pacientes para os quais não sabemos o valor e existem alguns casos, com valores anormalmente altos, fora dos parâmetros, que devem corresponder a erros de registo.
Decisão	Por compração com a variável target foi possível concluir que os pacientes sem valor para esta variável estão proporcionalmente distribuídos nos dois grupos (com e sem insuficiência renal): 

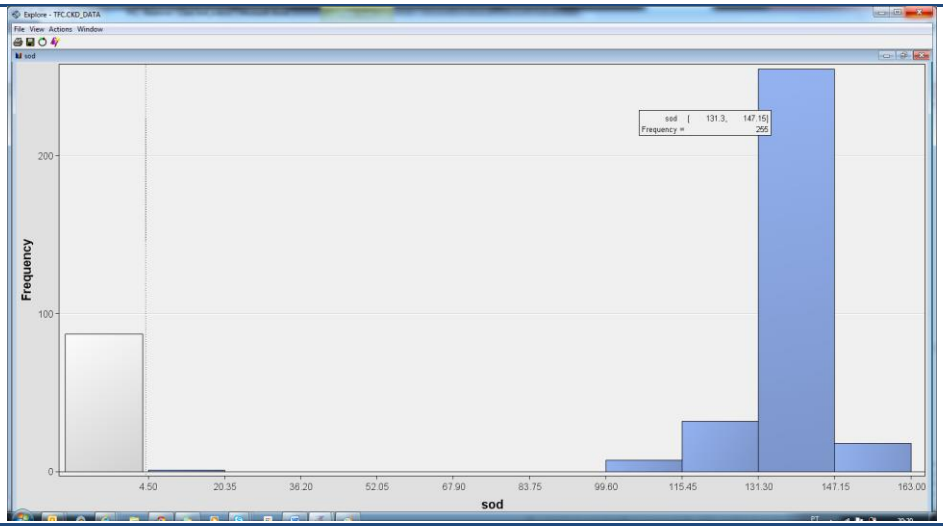
sc	
	Ao comparar a maior parte dos resultados com o *target* foi possível concluir que para valores médios, não há uma clara diferença entre os que apresentam a patologia e aqueles que não a apresentam.  Assim, para este caso, decidi preencher os valores omissos com o valor médio: ~3.07 mg/l.

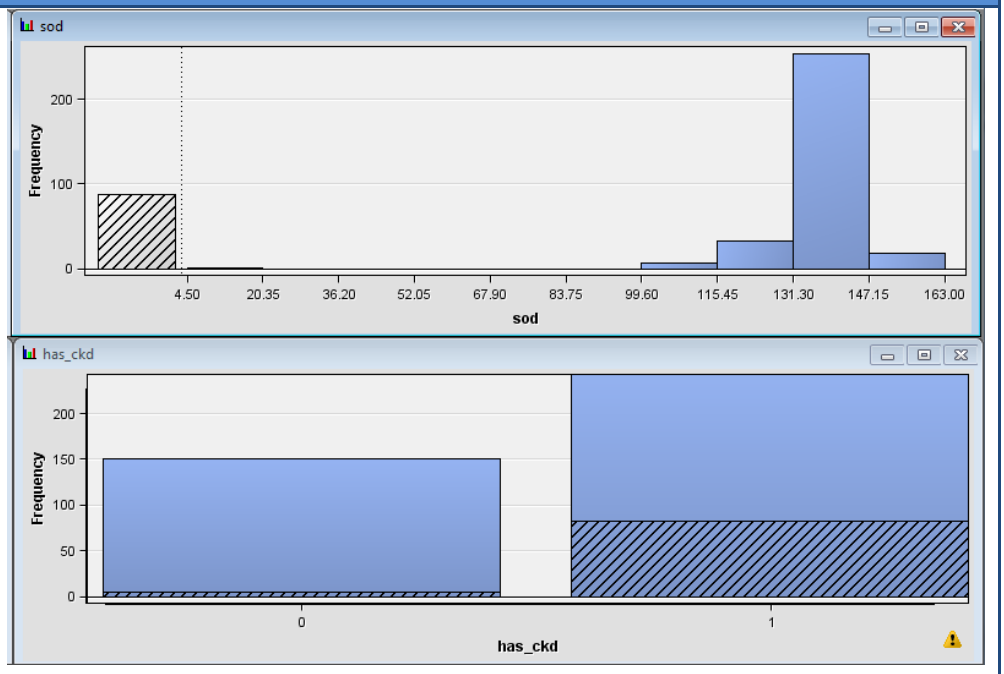
Densidade da urina (sg)

sg	
Descrição	Variável nominal, com valores específicos, indicativa da densidade da urina.
Situação	Para esta variável existem 47 pacientes para os quais não conhecemos o valor.
Histograma	
Decisão	Ao comparar os valroes omissos com a variável target foi possível eterminar que dos 47 pacientes para os quais não conhecemos o valor desta variável 42 apresentam insuficiência renal.  Assim, para este caso foi possível determinar que dos 47 pacientes, 42 apresentam a patologia enquanto que 5 deles não a apresentam. Ao comparar os intervalos de valroes também foi possivel perceber que para valroes menores de densidad da urina, os pacientes paresentam todos insuficiência renal, enquanto que para valores maiores a maior parte não apresenta a patologia. Assim, para este caso, optei por preencher com o valor médio (~1.07), embora isso

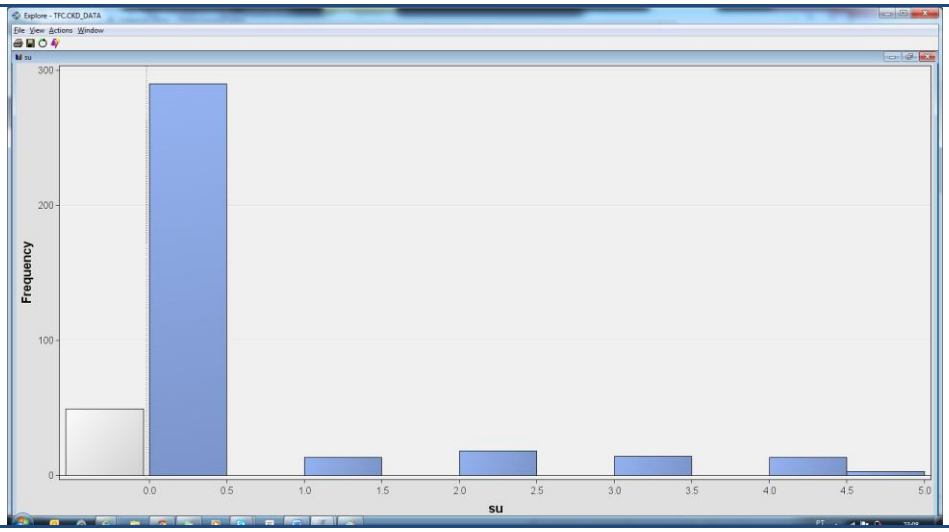
sg	
	significique que os 5 pacientes que não apresentam a patologia fiquem mal classificados.

Sódio (sod)

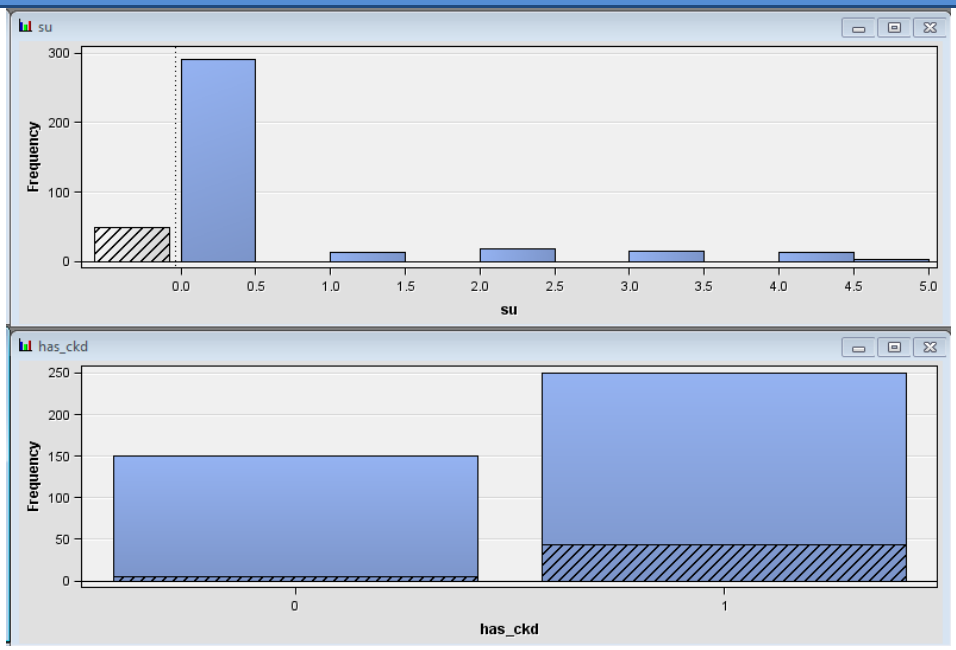
sod	
Descrição	Variável numérica, contínua, medida em mEq/l (miliequivalentes por litro) indicativa da concentração de sódio no sangue. Os valores considerados normais situam-se no intervalo entre 135 – 145 mEq/l.
Histograma	
Situação	Relativamente a esta variável, existem 87 pacientes para os quais não conhecemos os valores de sódio no organismo.
Decisão	Por comparação com a variável target foi possível determinar que 82 destes pacientes apresentam insuficiência renal enquanto que 5 deles não apresentam a patologia. Foi ainda possível que os pacientes com valores médios desta variável, os pacientes estão proporcionalmente distribuídos entre os que apresentam a patologia e os que não a apresentam.

sod	
	 Assim, para este caso, decidi preencher os valores omissos com o valor médio(~137.52 mEq/l).

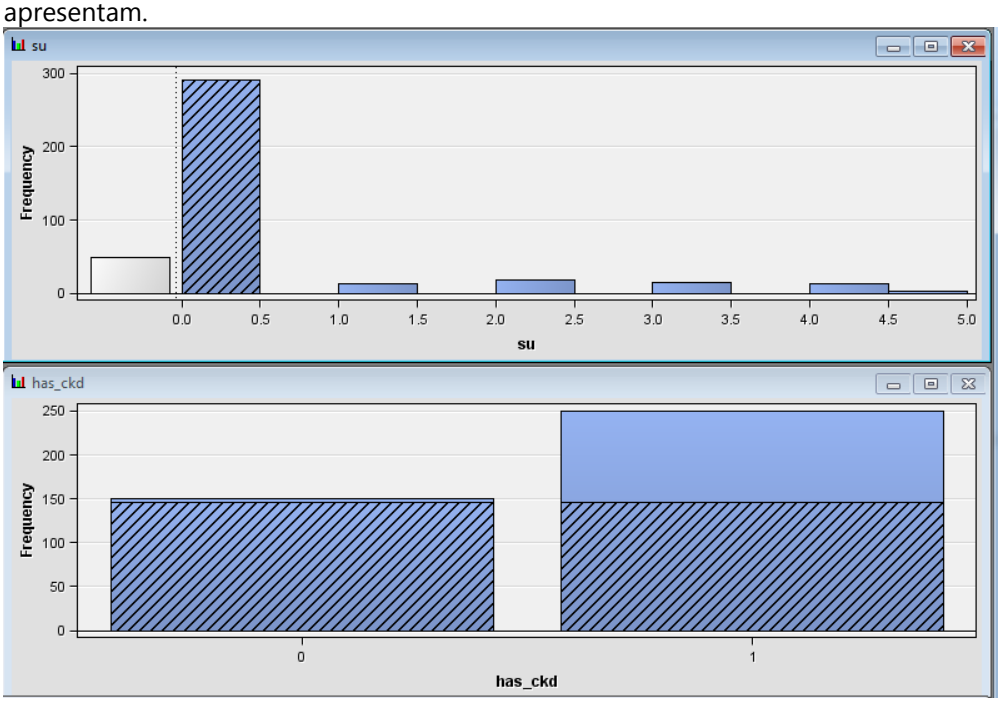
Glicosúria (su)

su	
Descrição	Variável nominal, indicativa da presença de açúcar na urina.
Histograma	
Situação	Para esta variável existem 49 pacientes para os quais não sabemos os valores de açúcar no sangue.
Decisão	Por comparação com a variável *target*, foi possível determinar que a maior parte dos pacientes com valores omissos também apresentam insuficiência renal.

su

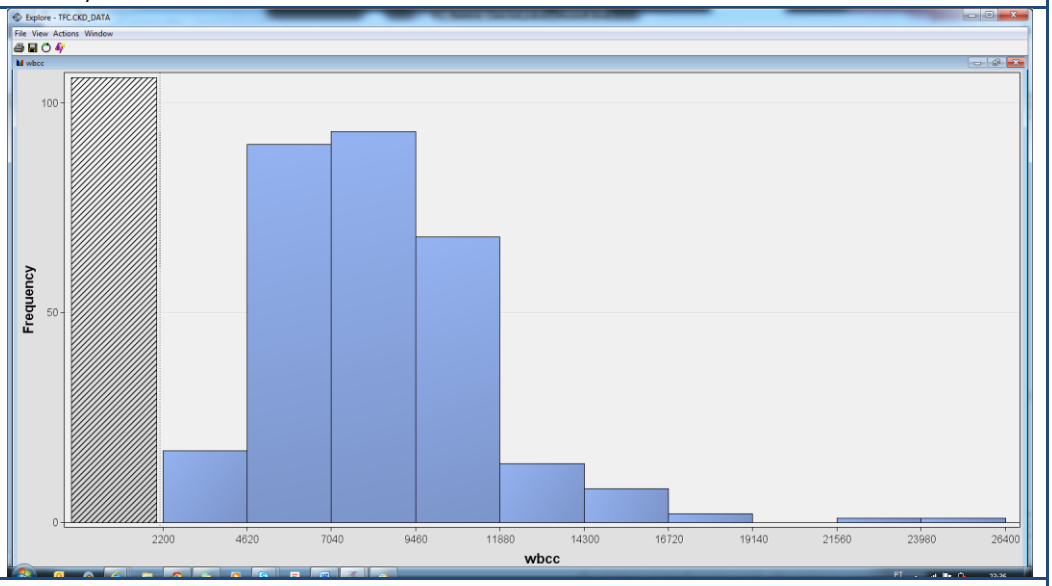


Por comparação com o valor médio, verificou-se que os pacientes estão uniformemente distribuídos entre os que apresentam a patologia e os que não a apresentam.



Assim, decidi preencher os valores omissos com a média: 0.45.

Contagem de glóbulos brancos (wbcc)

wbcc	
Descrição	Variável numérica, contínua indicativa da quantidade de leucócitos no sangue e é medida em número de células por mm^3 de sangue. Os valores de referência considerados normais para esta variável situam-se no intervalo: 4.000 – 12.000 células/mm^3
Histograma	
Situação	Para esta variável, existem 106 pacientes para os quais não sabemos o valor dos leucócitos presentes no sangue.
Decisão	Dado o elevado número de valores omissos face à amostra, optei por não usar esta variável.

Anexo III – Diagrama do caso do estudo

Os passos seguidos na construção dos modelos no **EMiner**™ estão representados no esquema da figura seguinte:

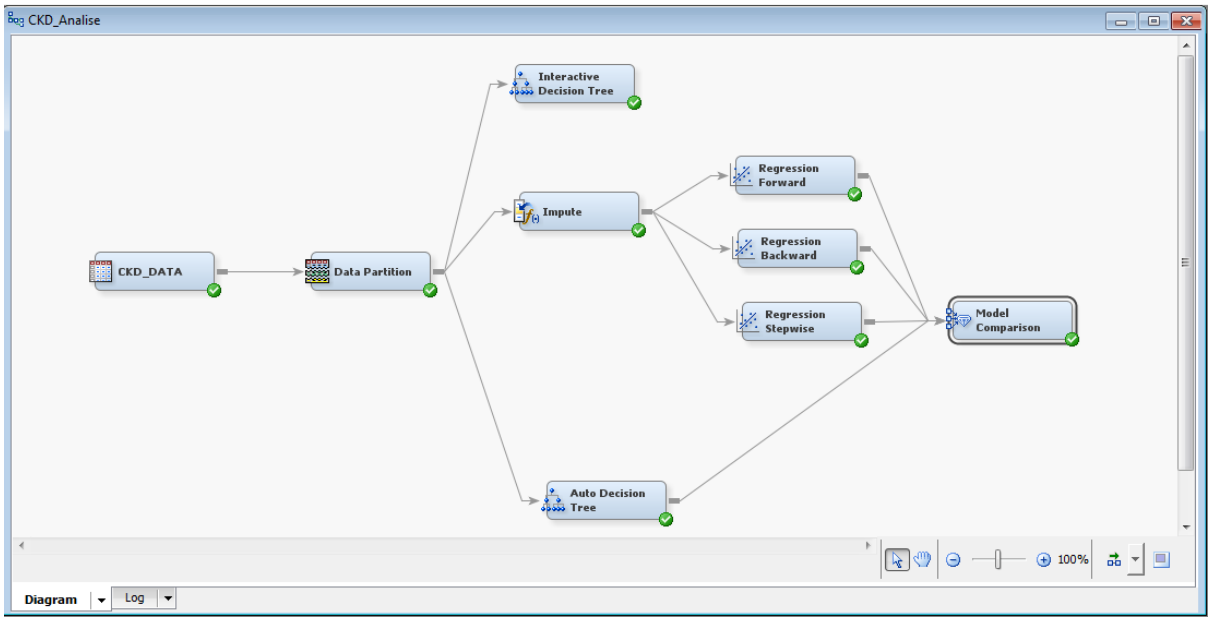


Figura 18 – Diagrama EMiner™