**Relatório de mineração de dados - Pré-processamento**

**Silas Leme Silvério – BI1760343**

O objetivo das técnicas de pré-processamento de dados é preparar os dados brutos para serem analisados.

**Informações de Data Set**

O data set escolhido se chama HCC Survival Data Set possui 49 colunas, contendo 165 instancias no total. A característica de seus atributos é de números inteiros e reais, associando os dados de modo a classificá-los.

O conjunto de dados de HCC foi obtido em um Hospital Universitário em Portugal e contém vários dados demográficos, fatores de risco, dados laboratoriais e de sobrevida global de 165 pacientes reais com diagnóstico de HCC. O conjunto de dados contém 49 recursos selecionados de acordo com as Diretrizes de Prática Clínica da EASL-EORTC (Associação Europeia para o Estudo do Fígado - Organização Europeia para Pesquisa e Tratamento do Câncer), que são o estado da arte atual sobre o gerenciamento de HCC.

Trata-se de um conjunto de dados heterogêneo, com 23 variáveis ​​quantitativas e 26 variáveis ​​qualitativas. No geral, os dados ausentes representam 10,22% de todo o conjunto de dados e apenas oito pacientes têm informações completas em todos os campos (4,85%). A variável alvo é a sobrevivência em 1 ano, e foi codificada como uma variável binária: 0 (morre) e 1 (vive). Um certo grau de desequilíbrio de classes também está presente

**Informações das colunas**

Gender: nominal

Symptoms: nominal

Alcohol: nominal

Hepatitis B Surface Antigen: nominal

Hepatitis B e Antigen: nominal

Hepatitis B Core Antibody: nominal

Hepatitis C Virus Antibody: nominal

Cirrhosis : nominal

Endemic Countries: nominal

Smoking: nominal

Diabetes: nominal

Obesity: nominal

Hemochromatosis: nominal

Arterial Hypertension: nominal

Chronic Renal Insufficiency: nominal

Human Immunodeficiency Virus: nominal

Nonalcoholic Steatohepatitis: nominal

Esophageal Varices: nominal

Splenomegaly: nominal

Portal Hypertension: nominal

Portal Vein Thrombosis: nominal

Liver Metastasis: nominal

Radiological Hallmark: nominal

Age at diagnosis: integer

Grams of Alcohol per day: continuous

Packs of cigarets per year: continuous

Performance Status: ordinal

Encefalopathy degree: ordinal

Ascites degree: ordinal

International Normalised Ratio: continuous

Alpha-Fetoprotein (ng/mL): continuous

Haemoglobin (g/dL): continuous

Mean Corpuscular Volume (fl): continuous

Leukocytes(G/L): continuous

Platelets (G/L): continuous

Albumin (mg/dL): continuous

Total Bilirubin(mg/dL): continuous

Alanine transaminase (U/L): continuous

Aspartate transaminase (U/L): continuous

Gamma glutamyl transferase (U/L): continuous

Alkaline phosphatase (U/L): continuous

Total Proteins (g/dL): continuous

Creatinine (mg/dL): continuous

Number of Nodules: integer

Major dimension of nodule (cm): continuous

Direct Bilirubin (mg/dL): continuous

Iron (mcg/dL): continuous

Oxygen Saturation (%): continuous

Ferritin (ng/mL): continuous

Class: nominal (1 se o paciente sobreviveu, 0 se o paciente morreu)

**Informações do Código**

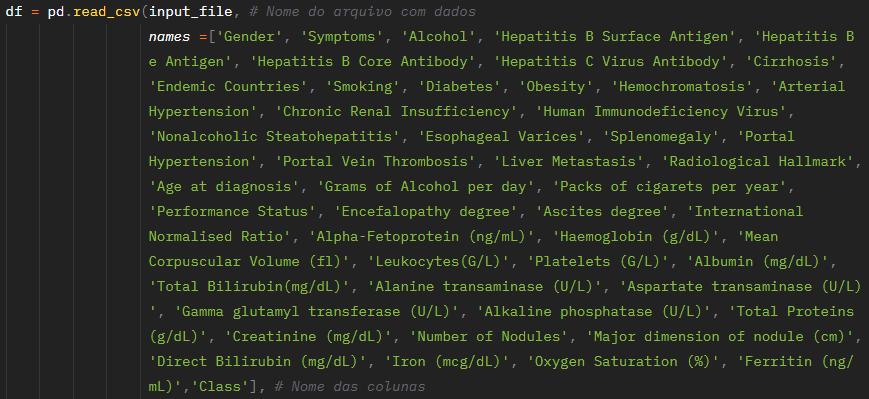
**Data Cleaning**

Para essa etapa do pré-processamento, constituída da limpeza dos dados foi utilizado como base o código disponibilizado pelo professor.

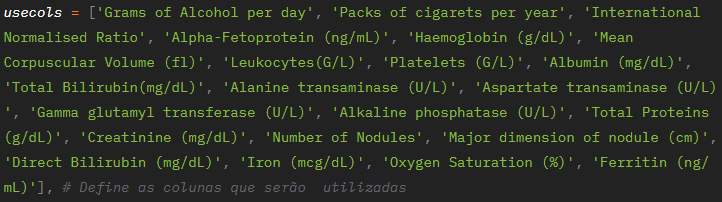
Algumas alterações foram feitas para que o algoritmo realizasse o tratamento de forma correta:

- Especificação do caminho para acesso ao arquivo com os dados.



- Input do nome de todas as colunas do Data Set.

- Seleção das tabelas que seriam utilizadas para o processo de limpeza.



Foram selecionadas essas colunas devido ao fato de haver nelas maior quantidade de valores ausentes, além de que muitas colunas possuírem valores nominais (1 - sim e 0 - não). Essas não foram consideradas para a limpeza, apenas as colunas com valores contínuos.

- Definição do caractere que sinaliza um valor ausente no arquivo com os dados.

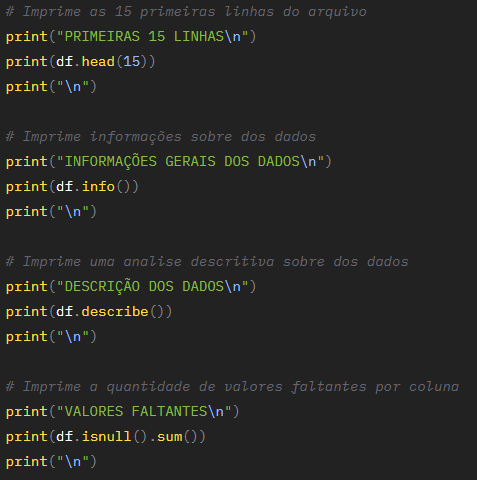


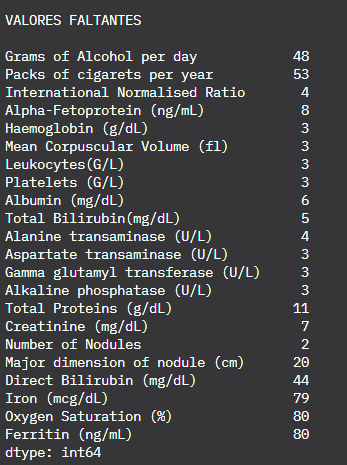
Como descrito acima essa data set possui diversos valores faltantes (10,22%), definidos com “?” no arquivo com os dados.

Para a configuração dos arquivos foram selecionados como arquivo de entrada o Data set original obtido no repositório de Data Sets e como arquivo de saída foi criado documento de texto chamado hcc-dataClear onde foram armazenados os dados depois do pré-processamento com as colunas preenchidas e sem dados faltantes.

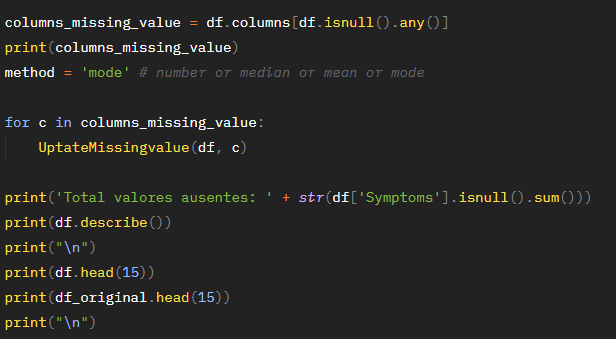


Algumas das funções iniciais não foram alteradas. Essas são funções secundárias, mas que auxiliam na visualização e estudo dos dados.



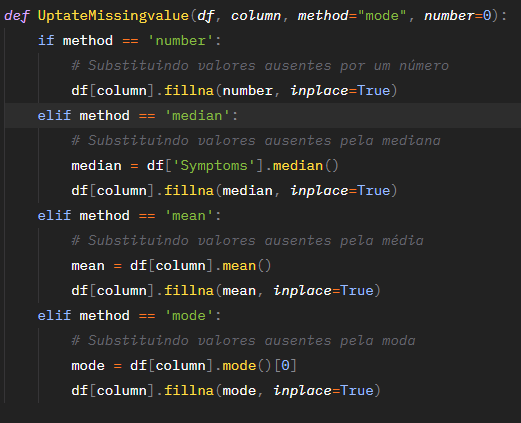


A função principal dessa etapa de pré-processamento é chamada de “*UpdateMissingvalue”*. Essa função é chamada dentro de um laço de repetição “for” que percorre cada coluna do Data set onde existem valores ausentes armazenados em uma variável chamada de “*columns\_missing\_values*”.



Na função UpdateMissingvalues são passados como parametro o DataFrame (df) que contem os dados, o indice da coluna (column) onde existem valores faltantes, o método (number, median, mean ou mode) que será utilizado para realozação do preenhimento dessas lacunas e o numero que será usado csa o método number seja escolhido.

Dentro da função está uma sequencia de laços condicionais que verificam qual a opção de método passada como parametro para que assim realize uma determinada ação conforme a escolha.



O método utilizado nessa atividade foi o de moda, pois diversas colunas possuem valores discretos definidos como 1 ou 0 por isso seria interessante utilizar aqueles que mais se repetem dada as certas condições, além de se tratar de dados sobre uma doença.

**Data Normalization**

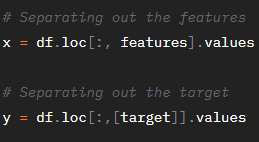
Para realização dessa etapa de pré-processamento foi utilizado como arquivo de entrada o documento gerado como saida da etapa anterior de limpeza dos dados, (hcc-dataClear.txt)



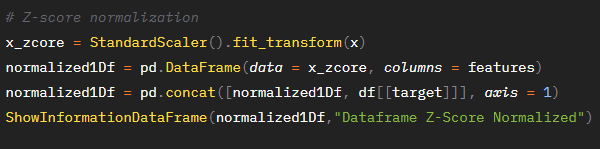
Da mesma maneira que na etapa anterios foram informados os nomes das colunas (names), bem como aquelas que seriam utilizadas na normalização (features).

Uma diferença realizada nesse código foi a identificação do *target*, isso é, aquela coluna que representa uma “resposta” ou que na analize das demais colunas nos da uma informação, no caso do Data set apresentado a coluna *Class* representa o target, pois informa se o paciente sobreviveu ou se ele morreu através dos valores 1 e 0, respectivamente.

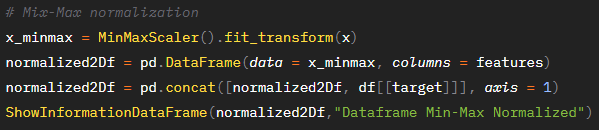
Após isso foi realizada a separação dos valores das *features* e do *target*, armazenando esses dados nas variéveis x e y, respectivamente.



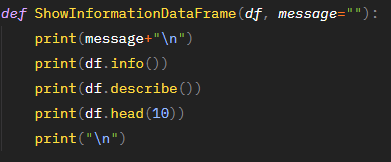
Foi realizada então a normalização Z (Z-Score) também conhecida como desvio padrão, que nos permite dizer o quão distante os valores estão da média. Em termos mais tecnicos é a medida de quantos desvios acima ou abaixo de uma pupulação média um dado cru está.



Além da *Z-score normalization* também foi realizada a normalização de máximos e mínimos (*Min-Max Normalization*) que é uma das mais comuns. Consiste em definir, para cada uma das *feature*, um valor mínimo que é transformado em 0 e o valor máximo que é transformado em 1, dessa forma todos os outros valores são transformados em valores decimais entre 1 e 0.



É possível realizar a comparação dos dados através da visualização dos dados do *DataFrame* original.

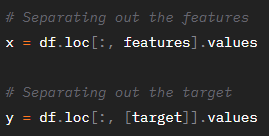


**Data Reduction**

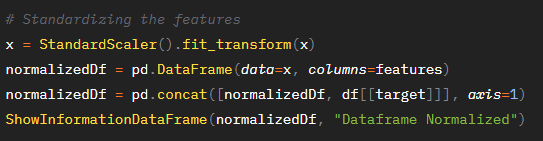
Para realização dessa etapa de pré-processamento foi utilizado como arquivo de entrada o documento gerado como saida da etapa anterior de limpeza dos dados, (hcc-dataClear.txt)



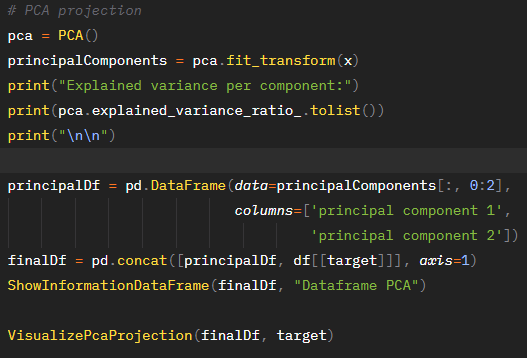
Da mesma maneira que nas etapas anteriores foram informados os nomes das colunas (names), bem como aquelas que seriam utilizadas na normalização (features), target e a separação desses dados atraves dos comandos



Após isso os as colunas foram padronizadas através do bloco de comandos:



E foi realizada a projeção PCA (Principal Component Analysis) que é responsável por reduzir a dimensão dos dados, ou seja, um DataFrame com n número de colunas pode ser projetado em um subespaço de um numero menor de colunas, mantendo a essência dos dados.



Além disso os dados também foram plotados, através da função VisualizePcaProjection, para uma melhor análise e avaliação dos dados.

