Breast Cancer Wisconsin (Diagnostic) Data Set

Arthur Biscaino Fruch 164373
José Carlos Cieni Júnior 170859

Base de dados

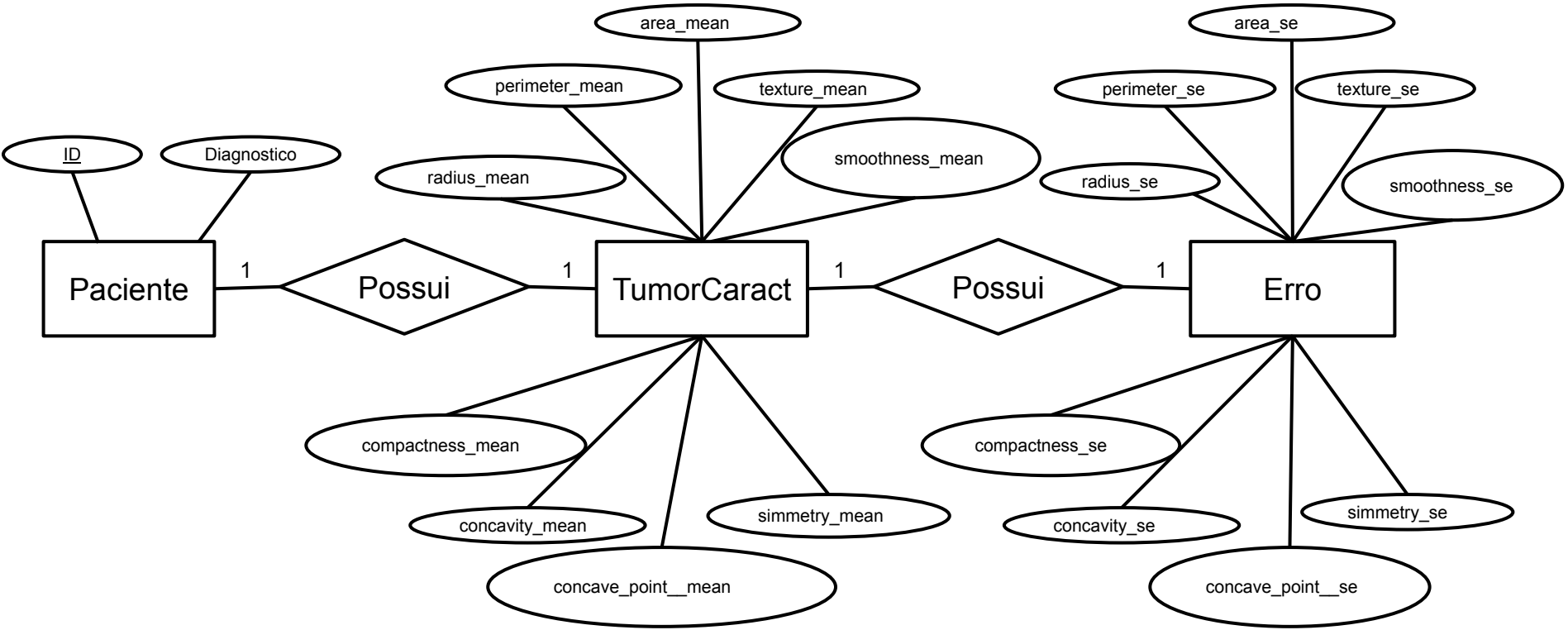
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

A base contém informações sobre diversos atributos físicos observados no núcleo celular de amostras de tumores na região da mama.

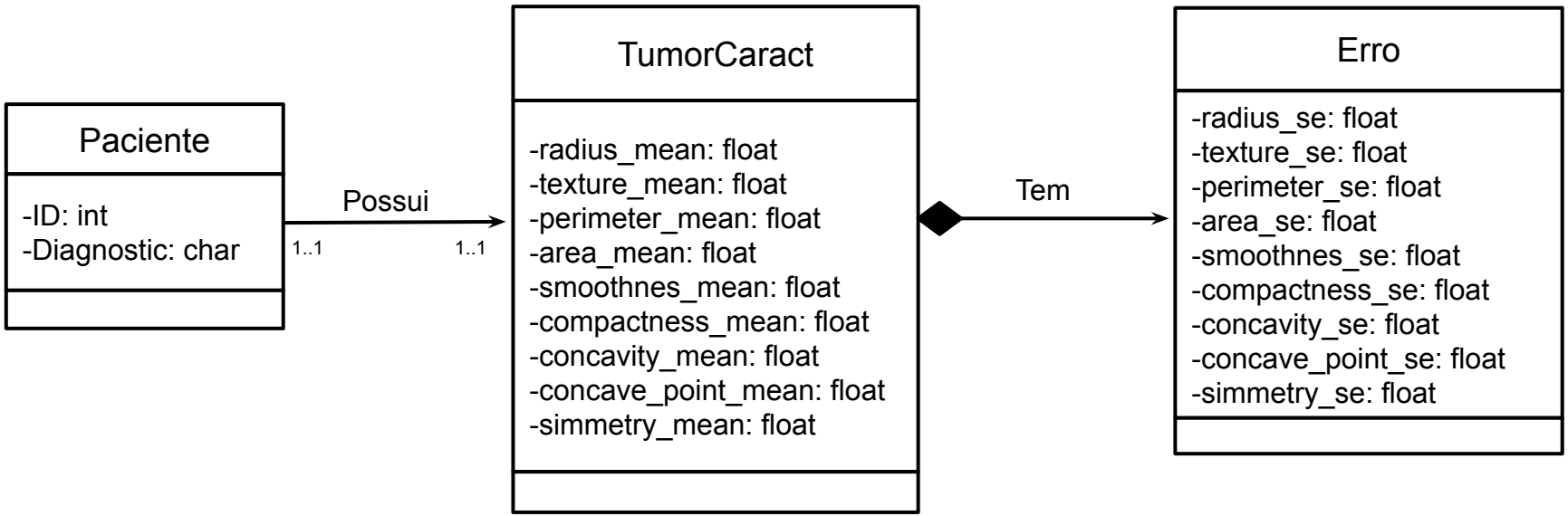
As informações foram agrupadas por paciente, contendo o diagnóstico (M = maligno, B = benigno) e as características da célula: raio (média da distância do centro da célula até o os pontos do perímetro), textura (desvio padrão dos valores em escala de cinza), perímetro, área, suavidade (variação local em raios diferentes), compacidade (perímetro²/área - 1), concavidade (grau de concavidade em pontos do contorno), pontos côncavos (total no contorno), simetria e dimensão fractal (aproximação do contorno - 1). Cada registro possui o valor médio, erro padrão e valor para o pior caso observado para cada paciente.

Nossa análise (no notebook) leva em consideração apenas o valor médio e o erro padrão.

Modelo conceitual - ER



Modelo conceitual - UML



Resultados

Através de consultas em SQL, foram comparadas as características obtidas de forma a tentar identificar quais são maiores indicativos de um tumor maligno.

Com base nos valores médios das amostras malignas e benignas foram feitas algumas comparações para determinarmos qual destes atributos eram mais relevantes para predizer justamente se uma amostra seria maligna ou benigna. Após algumas consultas foi identificado que os atributos mais relevantes para predição seriam a área, perímetro, compacidade, concavidade e os pontos côncavos.

Foi calculada a correlação entre a área e as demais características (identificamos que a área pode ser um bom ponto de partida, pois há diferenciação mais perceptível entre diagnósticos benignos e malignos - o que também faz sentido biologicamente).

Encontramos correspondência maior entre área x raio (0,987), área x perímetro (0,986) e área x n de pontos côncavos (0,823), o que era de se esperar, já

RAIO_MEDIA	12.14652380952381
TEXTURA_MEDIA	17.914761904761903
PERIMETRO_MEDIA	78.07540616246499
AREA_MEDIA	462.7901960784314
SUAVIDADE_MEDIA	0.0924776470588235
COMPACIDADE_MEDIA	0.0800846218487394
CONCAVIDADE_MEDIA	0.0460576210084033
PONTOSCONCAVOS_MEDIA	0.0257174061624649
SIMETRIA_MEDIA	0.1741859943977591
DIMENSAOFRACTAL_MEDIA	0.0628673949579832

Figura 1 - médias dos valores atribuídos às características observadas em pacientes com diagnóstico benigno

RAIO_MEDIA	17.462830188679245
TEXTURA_MEDIA	21.60490566037736
PERIMETRO_MEDIA	115.36537735849056
AREA_MEDIA	978.3764150943397
SUAVIDADE_MEDIA	0.10289849056603774
COMPACIDADE_MEDIA	0.14518778301886792
CONCAVIDADE_MEDIA	0.16077471698113208
PONTOSCONCAVOS_MEDIA	0.08799
SIMETRIA_MEDIA	0.19290896226415094
DIMENSAOFRACTAL_MEDIA	0.06268009433962264

Figura 2 - médias dos valores atribuídos às características observadas em pacientes com diagnóstico maligno