Cleveland: 282 ejemplos limpios.
Hungría: 294 ejemplos limpios.
Long Beach VA: 200 ejemplos limpios.
Suiza: 123 ejemplos datos.

Atributos
1 id: patient identification number
Se considera quitar este atributo, ya que el mismo no aporta para predicciones futuras, por ser algo propio de cada paciente.
Acción: Quitar

2 ccf: social security number (I replaced this with a dummy value of 0)
Se considera quitar este atributo, ya que el mismo no aporta para predicciones futuras, por ser algo propio de cada paciente. Además de que ha sido modificado de manera randómica.
Acción: Quitar

3 age: age in years
Se considera dejar este atributo, ya que el mismo está presente en los datos de cada paciente en cada data set, consideramos que es un factor de incidencia en la posibilidad de sufrir un infarto. Se puede apreciar en los historgramas de los dataset como en edades entre los 50 y 65 se concentra la gran mayoría de casos.
Acción: Dejar

4 sex: sex (1 = male; 0 = female)
Se considera dejar este atributo, ya que el mismo está presente en los datos de cada paciente en cada data set, consideramos que es un factor de incidencia en la posibilidad de sufrir un infarto. Se puede apreciar en los historgramas de los dataset como los hombres son más propensos que las mujeres a tener infartos.
Acción: Dejar

5 painloc: chest pain location (1 = substernal; 0 = otherwise)
Este atributo es un atributo, que se probará que resultados se obtiene sin el y cuales con el mismo, ya que no está presente en el dataset de Cleveland.
Se puede notar una tendencia que hay predisposición al infarto para aquellos que están en la clase 1.
Acción: Probar

6 painexer (1 = provoked by exertion; 0 = otherwise)

Este atributo es un atributo, que se probará que resultados se obtiene sin el y cuales con el mismo, ya que no está presente en el dataset de Cleveland.
Se puede notar una tendencia que hay predisposición al infarto para aquellos que están en la clase 1 para los dataset de Suiza y Long Beach. Sin embargo en el caso de Hungría ambas clases se pueden apreciar bien representadas, siendo la 0 en la que se encuentran más tendencia a infartos.
Acción: Probar

7 relrest (1 = relieved after rest; 0 = otherwise)
Este atributo es un atributo, que se probará que resultados se obtiene sin el y cuales con el mismo, ya que no está presente en el dataset de Cleveland y ni tampoco en de los 4 registros de Long Beach.
Al igual que el atributo anterior existe una tendencia hacia la clase 1 en los dataset de Suiza y LongBeach. Pero en el caso de Hungría, ambas clases están representadas de manera bastante equitativa.
Acción: Probar

8 pncaden (sum of 5, 6, and 7)
Por lo que se puede entender de la descripción es la suma de los 3 atributos anteriores, pero la misma no está realizada en ningún dataset.
Acción: Quitar

9 cp: chest pain type
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic
Se considera dejar este atributo, ya que el mismo, se encuentra presente todos los registros de todos los dataset.
En el caso de Suiza, Long Beach y Cleveland se puede apreciar una tendencia hacia los casos 3 y 4. Sin embargo el dataset de Hungría existe una gran cantidad de casos de la clase 2, siendo este número de casos mayor a los de la clase 3. Situación que no es así en los otros dataset.
Acción: Dejar

10 trestbps: resting blood pressure (in mm Hg on admission to the hospital)
Este atributo, tiene un promedio que ronda los 130 mm, con una media que está entorno a los 20 mm, entre todos los dataset. Es un dato que está presente en casi todos los registros de los dataset, pero en el caso del dataset de Long Beach faltan 56 registros con este dato, mientras que en los otros son 2 o 4. Por lo que se analizará si dejar el atributo y completar con un el promedio o media y ver que se resultados se obtiene.

Acción: Completar datos faltantes y probar.

11 htn
Este atributo tiene 30 faltantes en el dataset de Suiza, 3 y 1 en los casos de Hungría y Long Beach. Se probará que sucede sin o con este atributo, complementando con la media correspondiente de su dataset, valores >0.5 igualarlos a 1 caso contrario a 0. Por otra parte se puede ver que cada clase está baste pareja en cuanto a su distribución, por lo que no consideramos que sea un dato que tenga alta correlación la probabilidad de tener un infarto.
Acción: Quitar

12 chol: serum cholestoral in mg/dl
Es un atributo que en el caso de Suiza no está presente. Y en el caso de Long Beach son 7 los faltantes, mientras que en Hungría son 23. La media para Cleveland y Hungría está entorno a los 250 con un desvío promedio de 59 entre ambos. No obstante Long Beach está más alejado de la media de los otros dataset con una media de 178 y un desvío de 114.

Acción: Quitar

13 smoke: I believe this is 1 = yes; 0 = no (is or is not a smoker)
Consideramos quitar este atributo dado el alto número de faltantes en los dataset de Suiza, Hungría y Cleveland.
Acción: Quitar

14 cigs (cigarettes per day)
Situación similar al atributo anterior.
Acción: Quitar

15 years (number of years as a smoker)
Situación similar al atributo anterior.
Acción: Quitar

16 fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
De este atributo faltan datos 75 de Suiza, 7 de Long Beach, 8 de Hungría. Este datos parece ser importante ya que existe una tendencia hacia la clase 0.
Acción: Completar datos faltantes y probar.

17 dm (1 = history of diabetes; 0 = no such history)
Dado que este atributo solo está presente en el dataset de Hungría, y además en el mismo hay faltantes, no se tomará en cuenta.

Acción: Quitar

18 famhist: family history of coronary artery disease (1 = yes; 0 = no)
Dado que este atributo no está presente en el dataset de Suiza, hay un alto número de faltantes en el caso de Hungría y en los otros dos dataset no hay una clara tendencia hacia una de las clases no se tomará en cuenta.
Acción: Quitar

19 restecg: resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Solamente en el caso de Suiza y Hungría hay una tupla con este valor faltante por lo que se dejará el mismo. Se completará con la clase mayoría o se quitarán las tuplas. En el caso de Suiza y Hungría existe una clara tendencia hacia la clase 0.
Acción: Dejar

20 ekgmo (month of exercise ECG reading)
Entendemos que el mes en que se realizan el estudio no influye.
Acción: Quitar

21 ekgday(day of exercise ECG reading)
Se debería quitar este atributo, ya que es demasiado preciso cómo para poder predecir a futuro.
Acción: Quitar

22 ekgyr (year of exercise ECG reading)
Se debería quitar este atributo, ya que la intención es predecir a futuro.
Representa una variable específica del estudio realizado
Acción: Quitar

23 dig (digitalis used during exercise ECG: 1 = yes; 0 = no)
Parece una variable muy condicionante en cuanto a lo que se realizó en el electrocardiograma. Se puede apreciar tendencia hacia la clase 0.
Acción: Completar datos faltantes y probar.

24 prop (Beta blocker used during exercise ECG: 1 = yes; 0 = no)
Representa una variable específica del estudio realizado
Acción: Completar datos faltantes y probar.

25 nitr (nitrates used during exercise ECG: 1 = yes; 0 = no)
Representa una variable específica del estudio realizado
Acción: Completar datos faltantes y probar.

26 pro (calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)
Representa una variable específica del estudio realizado
Acción: Completar datos faltantes y probar.

27 diuretic (diuretic used used during exercise ECG: 1 = yes; 0 = no)
Representa una variable específica del estudio realizado
Acción: Completar datos faltantes y probar.

28 proto: exercise protocol
1 = Bruce
2 = Kottus
3 = McHenry
4 = fast Balke
5 = Balke
6 = Noughton
7 = bike 150 kpa min/min (Not sure if "kpa min/min" is what was written!)
8 = bike 125 kpa min/min
9 = bike 100 kpa min/min
10 = bike 75 kpa min/min
11 = bike 50 kpa min/min
12 = arm ergometer
Representa una variable específica del estudio realizado
Acción: Quitar

29 thaldur: duration of exercise test in minutes
Representa una variable específica del estudio realizado
Acción: Quitar

30 thaltime: time when ST measure depression was noted
Representa una variable específica del estudio realizado. A su vez, faltan muchos datos.
Acción: Quitar

31 met: mets achieved
Falta de valores y diferencias en las medias.
Acción: Quitar

32 thalach: maximum heart rate achieved
Faltan atributos, pero se puede notar una tendencia hacia valores que rondan los 125.
Acción: Completar y probar

33 thalrest: resting heart rate
Faltan atributos, pero se puede notar una tendencia hacia valores que rondan los 70.
Acción: Completar y probar

34 tpeakbps: peak exercise blood pressure (first of 2 parts)
Faltan atributos, pero se puede notar una tendencia hacia valores que rondan los 170.
Acción: Completar y probar

35 tpeakbpd: peak exercise blood pressure (second of 2 parts)
Faltan atributos, pero se puede notar una tendencia hacia valores que rondan los 85.
Acción: Completar y probar

36 dummy
Por falta de conocimiento sobre lo que representa el atributo
Acción: Quitar

37 trestbpd: resting blood pressure
Faltan atributos, pero se puede notar una tendencia hacia valores que rondan los 82.
Acción: Completar y probar

38 exang: exercise induced angina (1 = yes; 0 = no)
Puede influir en el resultado del estudio
Acción: Probar

39 xhypo: (1 = yes; 0 = no)
Puede influir en el resultado del estudio
Acción: Probar

40 oldpeak = ST depression induced by exercise relative to rest
Completar datos con 0,8
Acción: Probar y completar

41 slope: the slope of the peak exercise ST segment
-- Value 1: upsloping
-- Value 2: flat
-- Value 3: downsloping
Quitar por falta de valores.
Acción: Quitar

42 rldv5: height at rest
Quitar por falta de valores.
Acción: Quitar

43 rldv5e: height at peak exercise
Quitar por falta de valores y altos desvios en la media.
Acción: Quitar

44 ca: number of major vessels (0-3) colored by flourosopy
Quitar por falta de valores.
Acción: Quitar

45 restckm: irrelevant
Acción: Quitar

46 exerckm: irrelevant
Acción: Quitar

47 restef: rest raidonuclid (sp?) ejection fraction
Quitar por falta de valores.
Acción: Quitar

48 restwm: rest wall (sp?) motion abnormality
0 = none
1 = mild or moderate
2 = moderate or severe
3 = akinesis or dyskmem (sp?)
Quitar por falta de valores.
Acción: Quitar

49 exeref: exercise radinalid (sp?) ejection fraction
Quitar por falta de valores.
Acción: Quitar

50 exerwm: exercise wall (sp?) motion

Quitar por falta de valores.
Acción: Quitar

51 thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
Quitar por falta de valores.
Acción: Quitar

52 thalsev: not used
Quitar por falta de valores.
Acción: Quitar

53 thalpul: not used
Quitar por falta de valores.
Acción: Quitar

54 earlobe: not used
Quitar por falta de valores.
Acción: Quitar

55 cmo: month of cardiac cath (sp?) (perhaps "call")
Se considera que este atributo puede ser de utilidad si se considera que existe cierta concentración de infartos en los primeros meses del año y final del año, especialmente en los dataset de Suiza y Cleveland. Por lo que si se ingresa el mes que la persona que se está estudiando podría aumentarse o disminuir su chance de tener un infarto.
Acción: Dejar y Probar

56 cday: day of cardiac cath (sp?)
Se considera que este atributo es demasiado preciso cómo para preveer cuando será el próximo infarto de un paciente.
Acción: Quitar

57 cyr: year of cardiac cath (sp?)
Este atributo, no es de utilidad si se quieren predecir a futuro. Ya que mucha de la información fue recolectada en la década de 1980.
Acción: Quitar

58 num: diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing
-- Value 1: > 50% diameter narrowing
(in any major vessel: attributes 59 through 68 are vessels)
Por no faltan datos de este atributo

Acción: Dejar

59 lmt
Faltan 275 valores del atributo en el dataset de Hungría.
Acción: Quitar

60 ladprox
Faltan 276 valores del atributo en el dataset de Hungría.
Acción: Quitar

61 laddist
Faltan 246 valores del atributo en el dataset de Hungría.
Acción: Quitar

62 diag
Faltan 276 valores del atributo en el dataset de Hungría, 282 de Cleveland.
Acción: Quitar

63 cxmain
Faltan 235 valores del atributo en el dataset de Hungría.
Acción: Quitar

64 ramus
Faltan 285 valores del atributo en el dataset de Hungría, 282 de Cleveland.
Acción: Quitar

65 om1
Faltan 271 valores del atributo en el dataset de Hungría.
Acción: Quitar

66 om2
Faltan 289 valores del atributo en el dataset de Hungría, 282 de Cleveland.
Acción: Quitar

67 rcaprox
Faltan 244 valores del atributo en el dataset de Hungría.
Acción: Quitar

68 rcadist
Falta el 269 valores del atributo en el dataset de Hungría.
Acción: Quitar

69 lvx1: not used
Dado que no queda claro que es lo que expresa esta variable la misma no se utilizará.
Acción: Quitar

70 lvx2: not used
Dado que no queda claro que es lo que expresa esta variable la misma no se utilizará.
Acción: Quitar

71 lvx3: not used
Dado que no queda claro que es lo que expresa esta variable la misma no se utilizará.
Acción: Quitar

72 lvx4: not used
Dado que no queda claro que es lo que expresa esta variable la misma no se utilizará.
Acción: Quitar

73 lvf: not used
LVF significa “Left Ventricular Failure”, pero dado que no queda claro que es lo que se expresa de ello con estos números, a suponer clases se descartara.
Acción: Quitar

74 cathef: not used
Dado que no queda claro que es lo que expresa esta variable la misma no se utilizará. Además existen muchos faltantes de este atributo.
Acción: Quitar

75 junk: not used
Dado que no queda claro que es lo que expresa esta variable la misma no se utilizará.
Acción: Quitar

76 name: last name of patient (I replaced this with the dummy string "name")
Este atributo se quitará, ya que es igual para todas las tuplas de todos los dataset, y fue puesto en reemplazo al apellido del paciente.

Acción: Quitar