Universidad Autónoma de Nuevo León

Licenciatura en Actuaría

Minería de datos

AVANCE I PROYECTO INTEGRADOR

Equipo 14

Suarez Martinez Tadeo Alejandro #1806069

Grupo: 002

7to Semestre

Octubre del 2019

Título de la base de datos:

Cervical Cancer Risk Classification

https://www.kaggle.com/loveall/cervical-cancer-risk-classification

Descripción de los datos:

Los datos vienen en una tabla con columnas, a continuación, en la siguiente tabla describo lo que cada columna se refiere.

Columna	Descripción
Age	Indica la edad de la mujer. Se expresa en términos de valores numéricos.
Number of sexual partners	Indica la cantidad total de parejas sexuales encontradas. Se expresa en términos de valores numéricos.
First sexual intercourse	Indica la edad de una mujer cuando tuvo su primera relación sexual. Se expresa en términos de recuento.
Num of pregnancies	Indica el número total de veces que la mujer quedó embarazada. Se expresa en términos del recuento total.
Smokes	Indica si la persona fuma o no. Se expresa en términos de ceros (no fuma) y unos (fuma).
Smokes (years)	Indica el número total de años durante los cuales la mujer fuma. Se expresa en términos de recuento total.
Smokes (packs/year)	Indica el número total de paquetes de cigarrillos por año que fuma la mujer. Se expresa en términos de números.
Hormonal Contraceptives	Indica si la paciente usa anticonceptivos hormonales o no.
Hormonal Contraceptives (years)	Indica que durante cuántos años se utilizó el método anticonceptivo. Se expresó en términos de número total de años.

IUD	Indica si se utilizó o no el dispositivo anticonceptivo intrauterino. Se expresó en términos de ceros (no usó DIU) y unos (usó DIU).
IUD (years)	Indicó cuántos años se utilizó el DIU. Se expresa en términos del número total de años.
STDs	Indica la presencia de enfermedades de transmisión sexual. Se expresa en términos de ceros y unos.
STDs (number)	Indica el número total de enfermedades de transmisión sexual presentes en el paciente. Se expresa en términos de números.
STDs:condylomatosis	Indica la presencia de condilomatosis con el paciente.
STDs:cervical condylomatosis	Indica la presencia de condilomatosis cervical.
STDs:vaginal condylomatosis	Indica la presencia de condilomatosis vaginal.
STDs:vulvo-perineal condylomatosis	Indica la presencia de condilomatosis vulvoperineal.
STDs:syphilis	Indica la presencia de sífilis.
STDs:pelvic inflammatory disease	Indica la presencia de enfermedad inflamatoria pélvica.
STDs:genital herpes	Indica la presencia de herpes genital.
STDs:molluscum contagiosum	Indica la presencia de molusco contagioso.
STDs:AIDS	Indica la presencia de SIDA en el paciente.
STDs:HIV	Indica la presencia de VIH en el paciente.
STDs:Hepatitis B	Indica la presencia de Hepatitis B en los pacientes.
STDs:HPV	Indica la presencia de VPH en los pacientes.
STDs: Number of diagnosis	Indica el número total de veces que se han diagnosticado las ETS
STDs: Time since first diagnosis	Indica el número total de años desde el primer

	diagnóstico.
STDs: Time since last diagnosis	Indica el número total de años transcurridos desde el último diagnóstico.
Dx:Cancer	Indica la presencia de cáncer después del diagnóstico.
Dx:CIN	Indica la presencia de neoplasia intraepitelial cervical
Dx:HPV	Indica la presencia de virus del papiloma humano.
Dx	Indica la presencia de cualquiera entre cáncer, CIN y VPH.
Hinselmann	También conocido como colposcopia, es un procedimiento de diagnóstico médico para examinar una vista ampliada e iluminada del cuello uterino, así como de la vagina y la vulva.
Schiller	Schiller es una prueba médica en la que se aplica una solución de yodo al cuello uterino para diagnosticar el cáncer de cuello uterino.
Cytology	También llamada prueba de PaP, ayuda a detectar células anormales en el cuello uterino, que pueden convertirse en cáncer.
Biopsy	Resultado de la biopsia.

Justificación del uso de datos:

Me decidí a trabajar con esta base de datos porque me quería centrar en alguna enfermedad mortal, para ayudar con el diagnostico de esta, ya que se pueden emplear muchas técnicas, busqué en varias bases de datos y esta fue la que vi que tenía información más útil y concreta.

Cuenta con información sólida y los datos están muy completos, confió plenamente en que puedo hacer un buen trabajo de predicción con estos, ya que tiene toda la información necesaria acerca del tumor de cada paciente.

Planteamiento del problema:

Cada año se diagnostican alrededor de 11.000 nuevos casos de cáncer de cuello uterino invasivo en los EE. UU, el cáncer de cuello uterino mata a unas 4.000 mujeres en los EE. UU. y a unas 300.000 mujeres en todo el mundo. Numerosos estudios informan que los altos niveles de pobreza están relacionados con bajas tasas de detección. Además, la falta de seguro médico, el transporte limitado y las dificultades del idioma dificultan el acceso de una mujer pobre a los servicios de detección, es una enfermedad sumamente grave, pero se puede curar si se detecta a tiempo, es por eso que este modelo quiere ayudar con esta problemática del mundo de la medicina, este programa ayudara a cualquier hospital, clínica o grupo médico.

Objetivo final:

El objetivo final es predecir si la mujer tiene esta enfermedad base a sus características anteriormente mencionadas como por ejemplo el tamaño de este, su perímetro entre otros, esto ayudara a la empresa que compre este programa en su diagnóstico para esta enfermedad.

Planeación de la herramienta a utilizar.

El modelo para predecir este problema puede ser varios, ya que es un problema de clasificación, pero en este momento tengo en mente los siguientes:

Árbol de decisión y bosque aleatorio:

El algoritmo que pienso implementar es el árbol de decisión, ya que es un algoritmo de clasificación muy útil para resolver esta clase de problemas, utilizaremos reglas de decisión en base a las variables tomadas y no podía faltar el bosque aleatorio también uno de los algoritmos más usados para la clasificación el cual al igual que el árbol lo implementare y comprobare que variables tienen mayor “exactitud” para usarlas como nodos.

.