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.

En las columnas de diagnostico 1 significa positivo y 0 negativo

Columna	Descripción
Age	Edad del paciente.
Number of sexual partners	Numero de parejas sexuales que ha tenido.
First sexual intercourse	Primera relación sexual
Num of pregnancies	Numero de embarazos
Smokes	Fuma
Smokes (years)	Años fumando
Smokes (packs/year)	Numero de cajetillas por año
Hormonal Contraceptives	Anticonceptivos hormonales
Hormonal Contraceptives (years)	Años que se ha usado los anticonceptivos hormonales
IUD	Anticonceptivo DIU
IUD (years)	Años con el anticonceptivo DIU
STDs	Infección de transmisión sexual
STDs (number)	Numero de infecciones de transmisión sexual
STDs:condylomatosis	Condilomatosis
STDs:cervical condylomatosis	Condilomatosis cervical
STDs:vaginal condylomatosis	Condilomatosis vaginal
STDs:vulvo-perineal condylomatosis	Condilomatosis vulvoperineal

STDs:syphilis	Sífilis
STDs:pelvic inflammatory disease	enfermedad inflamatoria pélvica
STDs:genital herpes	Herpes Genital
STDs:molluscum contagiosum	Molusco contagioso
STDs:AIDS	SIDA
STDs:HIV	VIH
STDs:Hepatitis B	Hepatitis B
STDs:HPV	VPH
STDs: Number of diagnosis	Numero de diagnosticos
STDs: Time since first diagnosis	Tiempo desde el primer diagnóstico
STDs: Time since last diagnosis	Tiempo desde el ultimo diagnóstico
Dx:Cancer	Cancer
Dx:CIN	Neoplasia intraepitelial cervical
Dx:HPV	Mayor valor medio más grande de la simetria
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.

.