|  |  |
| --- | --- |
| A picture of a winding road and trees  Proyecto final  Minería de datos | Sobre el trabajo  Este documento muestra la investigación, recolección y análisis de datos sobre COVID-19.  Luis Vásquez Benjamyn Recinos  Wilson Nájera |

Índice

Introducción

Descripción del proyecto

El proyecto consta del análisis de una base de datos sobre contagios de COVID-19 a nivel mundial.

La primera parte del análisis está enfocado al incremento de muertes por covid-19 con relación al tiempo transcurrido, esto quiere decir que, con la base de datos y con algoritmos como k-vecinos se podrá comparar el número de muertes con las fechas que se programen en el algoritmo, para luego ser comparadas con una fecha distinta, poder hacer un análisis de incremento o disminución de muertes. Así como la creación de gráficas sobre las muertes sobre diferentes fechas establecidas.

La segunda parte del proyecto consta de la creación e interpretación de un árbol de decisiones que permita visualizar de manera clara y gráfica la información sobre muertes en algún determinado país o región. Para esto es necesario utilizar el algoritmo de árbol de decisiones.

La tercera parte del proyecto consta de un modelo de predicción a través de regresión lineal que permita ver el comportamiento de cantidad de muertes en relación a los casos confirmados de COVID-19. También, la creación de una gráfica de dispersión con la cual pueda ser más fácil entender el modelo.

El archivo necesario para la construcción de este proyecto es el de la base de datos de covid19 a nivel mundial del 2020. Esta base de datos cuenta con los datos necesarios para trabajar la comparación de las muertes correspondientes a fechas y otras variables.

La entrega consta de lo siguiente:

Documentación: Se entregará un documento con las especificaciones del proyecto

Capturas: Capturas de la implementación de los algoritmos de análisis de datos sobre la base de datos anteriormente mencionada.

Archivos: Archivos trabajados para el proyecto como: archivos de rapidMiner y NoteBook de Python.

Objetivos

General

Analizar la estadística de tasa de mortalidad por covid-19 con relación a diferentes variables con el uso de una base de datos de COVID-19 recabados en el año 2020.

Específicos

Crear una gráfica con relación a número de muertes en diferentes fechas.

Utilizar el algoritmo k-vecinos cercanos para comprobar la relación entre el número de muertes de una fecha establecida por países.

Utilizar Python y librerías para la creación de un modelo de regresión lineal.

Antecedentes

Página del gobierno de Guatemala

Esta es una página desarrollada por el gobierno de Guatemala para el análisis de los casos y muertes de COVID-19 en el país.

En esta se puede encontrar datos sobre casos y muertes a nivel nacional. Así como gráficas, porcentajes y tasas sobre municipios y departamentos mostrándolos por colores.

https://covid19.gob.gt/semaforo.html

A picture containing map

Description automatically generated

Herramientas por utilizar

RapidMiner: RapidMiner es un software dirigido hacia el análisis de datos, con el cual se podrá implementar cada uno de los algoritmos necesarios para la construcción del proyecto.

Python: Python es un lenguaje con el cual se puede hacer Data Minning sobre una base de datos. Esta herramienta servirá para la creación y ploteo de datos sobre muertes a causa del covid-19. El servidor que ejecute el programa será Jupyter Notebook.

Librerias:

Numpy: Esta es una librería para el uso de números

Pandas: Librería para análisis de datos, ML y Data Minning

Statsmodels: Librería para la creación de regresión lineal.

Matplotlib: Librería para gráficos.

Parte 1: Incremento de muertes entorno al tiempo

Parte 2: Árbol de decisión de muerte en relación a región

Parte 3: Regresión lineal casos-muertes

La tercera parte del proyecto consta de un modelo de predicción a través de regresión lineal que permita ver el comportamiento de cantidad de muertes en relación a los casos confirmados de COVID-19. También, la creación de una gráfica de dispersión con la cual pueda ser más fácil entender el modelo.

El modelo de predicción es desarrollado utilizando una regresión lineal entre las variables “cases” y “deaths” de la base de datos, los cuales son casos y muertes, respectivamente.

Para el desarrollo de la regresión lineal fue necesario la utilización del lenguaje de programación Python con las siguientes librerías:

* Pandas
* Numpy
* Matplotlib
* Stastsmodel

Desarrollo:

1. Importar las librerías y el archivo necesario

* Primero se importan las librerías necesarias y el archivo como base de datos

Text

Description automatically generated

1. Se crea la relación entre los casos y muertes

* Con el uso de la librería statsmodel se crea la relación entre las dos variables casos y muertes y con el uso de rsquared\_adj se crea la validación del modelo

Graphical user interface, text, application

Description automatically generated

1. Validación matemática

* El uso de la opción predict, se crea la validación matemática del incremento de muertes de la base de datos.

Graphical user interface, text

Description automatically generated

1. Creación de gráfica

* Con el uso de la librería matplotlib se crea una gráfica de dispersión para tener un mejor análisis de los datos.

Chart, scatter chart

Description automatically generated

Conclusión del modelo

La relación entre los casos confirmados y las muertes por COVID-19, según la regresión lineal, es de 0.55, esto puede ocurrir por la fecha en la que se confirmaron los casos, puesto que al principio de la pandemia es sabido que murió más gente de la que muere hoy en día.

Podemos incluir otra variable para ver si es la que determina la cantidad de muertes, o crear una regresión lineal múltiple para ver el comportamiento del modelo con diferentes variables en conjunto.

Conclusiones

Recomendaciones

Anexos

Enlace al proyecto en GitHub

https://github.com/LuiiVasquez/proyectoMdD

Base de datos covid-19

Table

Description automatically generated

Pagina que almacena la base de datos

Graphical user interface, text, application, email

Description automatically generated

Herramienta: Rapid Miner

Graphical user interface, text, application

Description automatically generated

Herramienta Jupyter Notebook Python

Graphical user interface, application, website

Description automatically generated