**DOCUMENTACIÓN PROYECTO FINAL.**

**Python para el análisis de datos**

****

**Juan Diego Mangones Ramirez**

**Sandy Paola Ramos correa**

**Ana pupo polo**

**Lacides Doria Sierra**

**Oscar Andrés Avila Morelo**

**Andrés Rivera Ortiz**

**Bladimir Correa Martinez**

**Universidad de Córdoba**

**Facultad de Ingeniería**

**Departamento de Ingeniería de Sistemas y Telecomunicaciones**

**Ingeniería de Sistemas**

**Febrero - 2022**

**Montería**

**CONTENIDO.**

[**1.** **JUSTIFICACION.** 3](#_Toc122142112)

[**2.** **DESCRIPCION FUNCIONAL DEL SISTEMA.** 4](#_Toc122142113)

[**3.** **METODOLOGIA DE DESARROLLO** 5](#_Toc122142114)

[**3.1.** **Planificación del sprint.** 6](#_Toc122142117)

[**3.2.** **Planificación diría del proyecto.** 6](#_Toc122142118)

[**3.3.** **Etapa de desarrollo.** 6](#_Toc122142119)

[**3.4.** **Revisión del sprint.** 6](#_Toc122142120)

[**3.5.** **Retroalimentación.** 7](#_Toc122142121)

[**4.** **ARQUITECTURA DEL SISTEMA.** 8](#_Toc122142122)

[**5.** **ANALISIS Y DISEÑO DEL SISTEMA.** 9](#_Toc122142123)

[**5.1.** **Preprocesamiento De Los Datos.** 18](#_Toc122142124)

[**5.2.** **Modelamiento.** 19](#_Toc122142125)

[**6.** **ESPECIFICACIONES DE REQUISITOS FUNCIONALES / NO FUNCIONALES** 22](#_Toc122142126)

[**6.1.** **Requisitos Funcionales.** 22](#_Toc122142127)

[**6.2.** **Requisitos No Funcionales.** 22](#_Toc122142128)

[**7.** **OBJETIVOS DEL SISTEMA.** 23](#_Toc122142129)

[**7.1.** **Objetivo General.** 23](#_Toc122142130)

[**7.2.** **Objetivos Específicos.** 23](#_Toc122142131)

[**8.** **ANEXO.** 24](#_Toc122142132)

[**8.1.** **MANUAL DE USUARIO.** 24](#_Toc122142133)

[**REFERENCIAS.** 32](#_Toc122142134)

# **JUSTIFICACION.**

Python es un lenguaje de programación multiplataforma de código abierto, el cual soporta parcialmente la orientación a objetos, programación imperativa y la programación funcional, lo cual nos permite crear todo tipo de aplicaciones. Gracias a que es un lenguaje sencillo de leer y escribir, aspecto que lo ha hecho ganar adeptos y posibilidades de expansión como en la inteligencia artificial, big data, machine Learning y data science.

Ha tenido un impacto impresionante, tanto así que las empresas han empezado a implementar este lenguaje en sus aplicaciones ya que es muy efectivo y rápido. Permitiendo así tener mayor productividad y rendimiento en sus tareas. En la actualidad nos vemos cada día más sumergidos en la famosa Inteligencia Artificial, Blockchain y el Machine Learning. Por tal razón nos vemos obligados según la demanda que hay en el mundo laboral a aprender este lenguaje, que es y será de mucha relevancia en la actualidad y en el futuro.

Python dispone de un intérprete por línea de comandos en el que se pueden introducir sentencias. Cada sentencia se ejecuta y produce un resultado visible, que puede ayudarnos Machine Learning es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. Es una tecnología que permite hacer automáticas una serie de operaciones con el fin de reducir la necesidad de que intervengan los seres humanos. Esto puede suponer una gran ventaja a la hora de controlar una ingente cantidad de información de un modo mucho más efectivo.

El lenguaje Python juega un papel fundamental en el contexto de la analítica de datos y Big Data, ya que dispone de las herramientas para casi todos los aspectos relacionados con la computación científica. Hasta ahora, entre los lenguajes que más se usaban para realizar análisis y visualización de datos se encontraban Matlab y R. Sin embargo, en los últimos años Python se ha hecho muy popular entre los desarrolladores de aplicaciones y analistas de datos.

# **DESCRIPCION FUNCIONAL DEL SISTEMA.**

EL aplicativo web creado como proyecto final para el diplomado “Python para el análisis de datos”, consta de dos partes principales, la barra lateral y el cuerpo principal

-**Barra lateral**

En esta podemos hallar todas las páginas que se utilizaron para este aplicativo destinados para que el usuario final interactúe con el dashboard, en ella se encuentran algunas gráficas, un mapa interactivo y selectores del tipo slider y selectbox, que interactúan de forma directa con la predicción que crea el api por fuera del dashboard.

-**Cuerpo principal**

Dentro del contenedor principal se encuentran los visualizadores de datos, dependiendo en que página esté, se encuentran las predicciones, los datos utilizados para modelo de Maching Learning, las gráficas y el mapa interactivo.

# **METODOLOGIA DE DESARROLLO**

La metodología ágil es un tema candente y se escucha cada vez más en la ingeniería de software, que es un monopolio muy fuerte. Prueba de ello es que han formado un campo importante en muchos congresos y seminarios que se han celebrado en los últimos años. También es un campo que ha aparecido en las revistas internacionales más prestigiosas. En los círculos de ingeniería de software, existe un amargo debate entre los defensores de los métodos tradicionales (llamados despectivamente "pesados") y los defensores de las ideas creadas por el "Manifiesto Ágil". Las metodologías ágiles están especialmente orientadas para proyectos pequeños, estas constituyen una solución a medida para ese entorno, aportando una elevada simplificación que a pesar de ello no renuncia a las prácticas esenciales para asegurar la calidad del producto. (Penadés & Torres, 2006).

**Metodología Scrum:** Este enfoque es por fases, tenemos fases (análisis, desarrollo y pruebas) para hacer funcionar la empresa. Aquí resumimos cómo hacer este trabajo, cuántos ciclos y cuánto tiempo lleva. Es necesario utilizar métodos flexibles para desarrollar las prácticas propuestas. Debido a la adaptación a la realidad tecnológica y su constante desarrollo, se llegó a un acuerdo para desarrollar la aplicación en un plazo mínimo de uno a seis meses con diversas actualizaciones, introduciendo nuevas funcionalidades. (Martin, 2019) .

Por lo tanto, este enfoque ágil se utilizará en el desarrollo de este proyecto, ya que permite una implementación más rápida de proyectos a corto plazo y es más adaptable a los resultados. Scrum define y guía el desarrollo de este proyecto a través de 5 fases o hitos, son:



## **Planificación del sprint.**

La primera fase definirá aspectos como funciones, objetivos, riesgos y ciclo de entrega necesarios para el desarrollo del proyecto. Luego hay una reunión entre los equipos para explicar cómo se ha desarrollado cada punto de alcance. Los cambios, decisiones, mejoras y más se consideran aquí.

## **Planificación diría del proyecto.**

En esta u otra fase, las actividades creadas en el proyecto se planifican durante 15 minutos al comienzo del día y se elabora un plan de acción para las próximas horas. El equipo de desarrollo utilizará la línea de tiempo diaria para medir el progreso hacia el Sprint Goal y la dirección del progreso hacia la finalización del trabajo.

## **Etapa de desarrollo.**

En este punto del trabajo de sprint, el gerente se asegurará de que no se realicen cambios de última hora que puedan afectar los objetivos del proyecto, como requisitos de última hora o cambios de movimiento. Además, se asegura el cumplimiento de los plazos establecidos para la culminación exitosa del proyecto.

## **Revisión del sprint.**

Al final del desarrollo previsto, en este punto será posible analizar y evaluar los resultados y verificar que cumplan con todos los requisitos. Si es necesario, el equipo trabajará en conjunto para determinar qué se debe cambiar.

## **Retroalimentación.**

En este momento o etapa, se pueden obtener comentarios sobre los resultados alcanzables no solo de los expertos del proyecto, sino también de aquellos que usarán la aplicación directamente. La experiencia adquirida en esta etapa hará que el próximo enemigo sea más eficiente y ágil.

Con este enfoque, se asegura una operación más dinámica, lo que naturalmente aumenta la eficiencia y la calidad del producto, factores que las empresas buscan hoy, sus productos sean flexibles, independientes y eficientes, lo que significa menores costos de producción. Sin embargo, tiene importantes implicaciones comerciales.

# **ARQUITECTURA DEL SISTEMA.**

Interfaz de usuario gráfica

Descripción generada automáticamente

Para la arquitectura y despliegue de este aplicativo se tuvo en cuenta trabajar con dos repositorios en Github uno donde se encuentra alojado la API la cual está conectada al modelo de predicción, así como el mismo modelo y otro repositorio donde se encuentra alojado el Dashboard.

# **ANALISIS Y DISEÑO DEL SISTEMA.**

En este segmento, se hablará sobre el análisis de los datos encontrados en el dataset con el nombre Colombianos\_registrados\_en\_el\_exterior.csv correspondiente a nuestro grupo, que, según la investigación realizada, el dataset proviene de una plataforma de datos del gobierno colombiano llamada datos.gov.co.

En este caso lo primero que haremos será cargar los datos.

Captura de pantalla de un celular

Descripción generada automáticamente

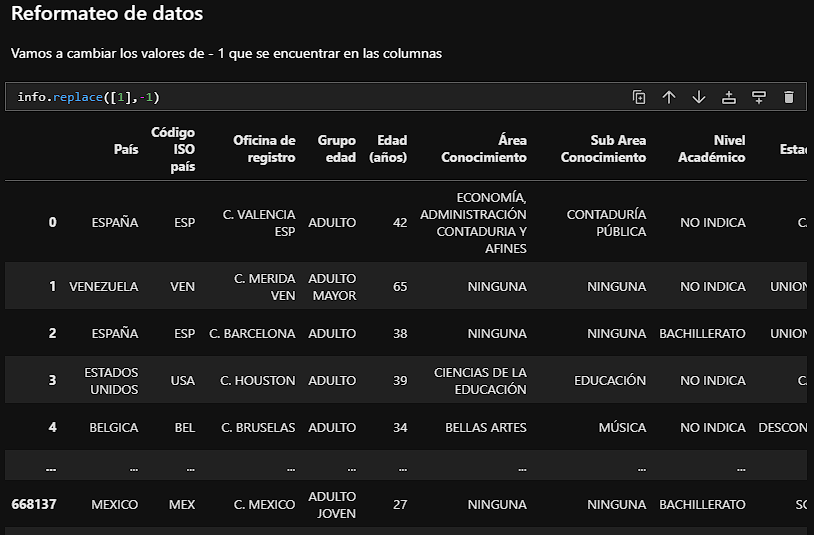
Procedemos a verificar si en el dataset se encuentran presentes datos NaN.

Texto

Descripción generada automáticamente

Se notan que no tenemos datos nulos en nuestro dataset.

Formateamos los datos de valores -1 a 1.



Analizamos que al quitar las filas que la edad mínima de 6 años y la estatura mínima en cm es de 1, puesto que 1 centímetro no es una estatura adecuada y puede que afecte nuestro análisis, entonces procedemos filtrar por unas tallas mínimas promedio, que corresponda adecuadamente con la edad mínima promedio entre 6 y 7 años, la cual estaría entre 90 cm bajo el supuesto de que la persona tenga problemas estatura.

Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Listo ya tenemos nuestro dataset limpio con respecto a las edades mínimas, ahora hay que revisar las edades y estaturas máximas con respecto a las edades, tenemos que **"el límite de la longevidad humana está en 115,7 años para las mujeres y 114,1 años para los hombres".** Sacado de: [Aquí](https://www.mundodeportivo.com/elotromundo/bienestar/20170906/431090171774/edad-maxima-ser-humano-anos-vida.html#:~:text=As%C3%AD%2C%20el%20l%C3%ADmite%20de%20la,en%20el%20momento%20de%20fallecer.). Para la estatura tenemos la siguiente información: **"El hombre más alto vivo en Estados Unidos es Igor Vovkovinskiy, quien mide 2,20 metros de altura. El hombre vivo más alto del mundo es Sultan Kösen, quien mide 2,52 metros. Pero el hombre más alto del mundo en toda la historia fue Robert Wadlow, quien murió en 1940 a los 22 años, y medía 2,70 metros de altura."** [ver pagina aquí](https://www.univision.com/explora/que-tan-alta-puede-ser-una-persona) (con base a esta información podemos decidir que la talla máxima que utilizaremos será de 220 cm y para la edad usaremos 116 años como mucho).

Captura de pantalla de computadora

Descripción generada automáticamente con confianza media

Se procede a verificar si las variables categóricas están bien escritas y verificar las variables binarias y se borran columnas que no sirven para este análisis.

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

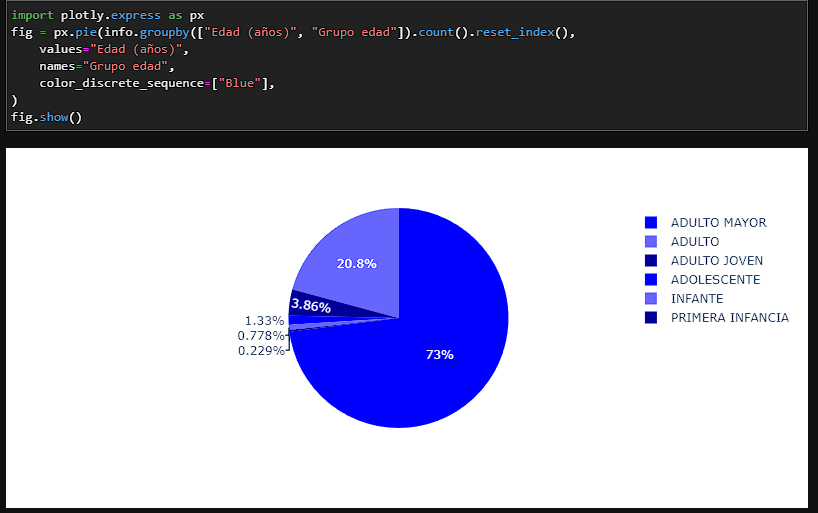
Texto

Descripción generada automáticamente

Teniendo en cuenta el análisis anterior profundizaremos en algunas columnas para verificar la lógica que tiene el dataset. Se revisará a mayor profundidad los datos para detectar ciertas interacciones y guiar la futura interpretación de los modelos.

Para ello, se revisarán interacciones, agrupaciones y visualizaciones de los datos, y se generarán ideas de las algunas interacciones interesantes que se detecten.

Usando plotly.express graficamos a continuación la cantidad de personas por su grupo de edad.



Se observa que el mayor porcentaje de colombianos en el extranjero son adultos mayores con 73%.

Se comparan por medio de un gráfico de barras las áreas de conocimientos.

Interfaz de usuario gráfica, Sitio web

Descripción generada automáticamente

Podemos observar que el conocimiento de los colombianos en el exterior varia bastante y que la mayoría de las personas no puso su área de conocimiento.

Con respecto al nivel académico de los colombianos se puede variar, así como el área de conocimiento.

Interfaz de usuario gráfica

Descripción generada automáticamente

Así como con el grafico del área de conocimiento, podemos observar que hay una gran variedad de títulos profesionales para los colombianos en el extranjero, pero de igual forma la mayoría de estos en su encuesta realizada no indicaron su nivel académico.

Pasaremos ahora a revisar la correlacionalidad que tiene que tienen las variables con respecto a las otras.

Procedemos a hacer el heatmap para la correlación.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

Con esto podemos concluir que:

* La variable a la cual se llevará a predecir será el país en base a algunas variables sociodemográficas.

## **Preprocesamiento De Los Datos.**

Se aplica cambios a los datos originales, donde tenemos tanto variables tanto numéricas como categorías donde se probarán inicialmente los modelos:

* RandomForestClassifier
* LogisticRegression

Para las pruebas iniciales, se hará un split 80-20 para Entrenamiento y Prueba, se tomarán más muestras de la clase minoritaria (upsampling) para manejar la clase no balanceada y se tomará la versión "por defecto" de los modelos en sklearn con todos los features del set de datos.

Teniendo en cuenta lo descubierto en el análisis exploratorio, partiremos el dataset para sacar las columnas que requerimos.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

## **Modelamiento.**

Se usarán funciones para partir los datos y usar Onehot para los datos categóricos.

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Aplicaremos estas funciones a los datos y usamos el primer modelo establecido que es el LogisticRegression.

Texto

Descripción generada automáticamente

Dado estos modelos y sus resultados, se escogió el RandomForestClasifier. Después de revisar los modelos usándolos en la API dispuesta para ello, se acordó para que nuestro caso, se usara el modelo del RandomForestClasifier sin tuneo, ya que da un resultado más acercado al esperado. Por ende, se procede a guardar el modelo.

Texto

Descripción generada automáticamente con confianza media

# **ESPECIFICACIONES DE REQUISITOS FUNCIONALES / NO FUNCIONALES**

## **Requisitos Funcionales.**

El dashboard debe poder mostrar graficas estadísticas aparte pueda realizar predicciones usando variables ya establecidas por el modelo de predicción, dicha predicción será país con esas características de las variables.

Para su uso de manera local de los archivos, se debe instalar las librerías que se mencionan en los **requirements.txt** encontrados en sus respectivos repositorios de Github.

## **Requisitos No Funcionales.**

* **Desarrollo:** Lenguaje de programación a usar, patrones de diseño y entorno de desarrollo.
* **Operaciones:** procedimientos operativos de cómo usar el software.
* **Éticos:** aseguran que el sistema sea fiable tanto para el usuario.

# **OBJETIVOS DEL SISTEMA.**

## **Objetivo General.**

Crear un aplicativo web utilizando las librerías de Python tales como Streamlit, Plotly Express, Pandas y Numpy que permita interactuar con modelos de aprendizaje automático previamente entrenados y luego interactuar con datos de un conjunto de datos determinado, siguiendo estrictamente las mejores prácticas para la comunicación y la visualización efectiva de datos.

## **Objetivos Específicos.**

* Implementar librerías y tecnologías que permitan la expresión de información de big data utilizando lenguajes de alto nivel como Python, posibilitando un análisis de datos eficiente.
* Desarrollar un modelo de Machine Learning que permita predecir, dado unas variables, el país para migrar.
* Establecer una conexión por medio de una API, el modelo con el dashboard.
* Diseñar con la librería Streamlit un dashboard funcional como parte de este proyecto.

# **ANEXO.**

## **MANUAL DE USUARIO.**

1. **INTRODUCCIÓN.**

Este documento está destinado a ayudar a los usuarios con el uso final del proyecto, enseñar el uso correcto y garantizar que nuestra documentación sea realmente útil; así el ambiente será de lo más familiar.

1. **DESCRIPCIÓN.**

A menudo, los usuarios de programas no utilizan estos programas correctamente, o más bien no utilizaron todas las herramientas disponibles para su trabajo; este es el por qué Se necesita una guía para guiarlos.

1. **CONTENIDO.**

A continuación, se presenta una introducción al sistema en donde se describirá su funcionamiento, permitiendo determinar la utilidad y usabilidad de este; luego en la sección de ingreso al sistema se presentarán los pasos determinados y adecuados que permitirán acceder a este.

En los componentes de interfaz, se evidencian las herramientas que pertenecen a este y la

descripción del uso de cada una de estas; posteriormente se hará referencia al uso del sistema para que el usuario conozca cada parte de este, con sus respectivas pantallas.

1. **INTRODUCCIÓN AL SISTEMA.**

El aplicativo permite facilitar el uso de un modelo de predicción para el precio de referencia a un vehículo ya usado.

**5. INGRESO AL SISTEMA.**

Para acceder a la aplicación, el usuario debe descargar el repositorio, instalar las librerías requeridas y correr en la terminal el comando:

streamlit run Bienvenido.py

Texto

Descripción generada automáticamente

**6. COMPONENTES DE INTERFAZ.**

**6.1. Inicio.**

Una vez iniciado se haya accedido por medio del link aparecerá la vista de Inicio en la que se dispone:

1. **Menú (Sidebar):** Contiene las diferentes páginas del aplicativo.

2. **Cuerpo:** Muestra los gráficos donde se muestran los distintos análisis basados en los

datos almacenados en el sistema más un mensaje de Bienvenida dependiendo de en qué página est.

Texto

Descripción generada automáticamente

**7. USO DEL SISTEMA**

**7.1. Predicción.**

Podemos dar clic en la página llamada Predicción.

Texto

Descripción generada automáticamente

Captura de pantalla de un celular

Descripción generada automáticamente

La predicción del país puede cambiar utilizando los selectores que se encuentran más abajo.

Patrón de fondo

Descripción generada automáticamente

**7.2. Gráficas:**

Para el apartado de gráficas podemos irnos ya sea a la página Pie o Barras.

En la pagina de Pie, podemos ver un gráfico de pie.

Gráfico, Gráfico circular

Descripción generada automáticamente con confianza media

En la página de Barras podemos ver los gráficos de barras sobre los datos.

Interfaz de usuario gráfica, Gráfico, Aplicación

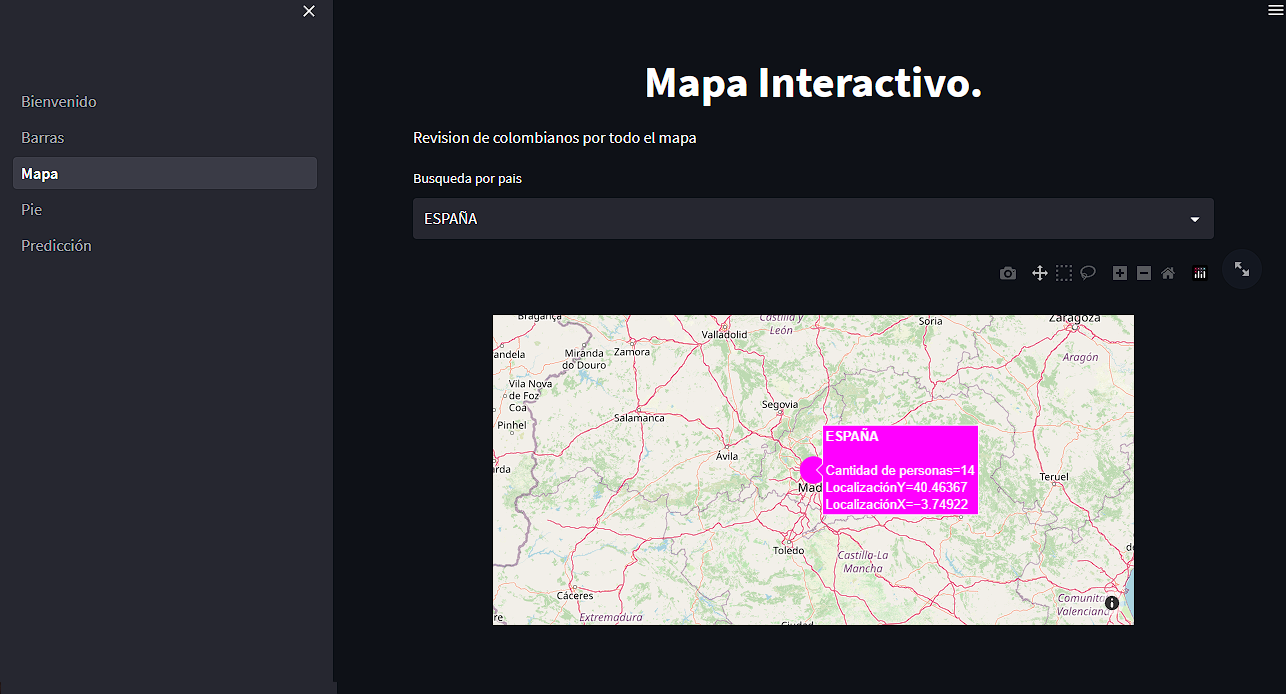
Descripción generada automáticamente

Interfaz de usuario gráfica, Gráfico

Descripción generada automáticamente

Se puede cambiar la marca con el selector, en donde aparecen todas las marcas.

Y por ultima tenemos la página para el mapa interactivo, la cual cambia de posición del mapa con el Selector.



Se puede cambiar el país en el selector, y el gráfico se actualizará con los valores nuevos.

Interfaz de usuario gráfica

Descripción generada automáticamente

Captura de pantalla de computadora

Descripción generada automáticamente

**8. CONCLUSIÓN.**

El anterior manual de usuario se espera que explique el funcionamiento del aplicativo presentado en este proyecto.

# **REFERENCIAS.**

Datamedia. (2012). *Datamedia*. Obtenido de Datamedia: https://datademia.es/blog/que-es-python

Martin, F. (03 de 12 de 2019). *martinFowler.com*. Obtenido de The New Methodology: https://martinfowler.com/articles/newMethodology.html

Penadés, C., & Torres, P. O. (2006). Métodologías ágiles para el desarrollo de software: eXtreme Programming (XP). *Técnica administrativa*, 1666-1680.