UNIVERSIDAD DE ALCALÁ DE HENARES

Anteproyecto de Trabajo Final de Grado

Plataforma basada en Python para transformación e integración de datos abiertos (open data) sobre servicios de análisis y visualización

Marcos Navarro Juan

Grado en Ingeniería Informática

Escuela Politecnica Superior

marcos.navarrroj@edu.uah.es

Tutor: Manuel de Buenaga Rodríguez

manuel.buenaga@uah.es

1. Introducción

La idea del proyecto nace de la importancia que existe en la actualidad de la información, más concretamente, en el tratamiento de datos abiertos. Se considera una parte fundamental a la hora de tener una mejor percepción del mundo que nos rodea. Todo tipo de agentes utilizan múltiples técnicas de tratamiento de datos y de visualización de estos para obtener una serie de conclusiones esenciales para el campo en el que se este trabajando. Obtener unas conclusiones de calidad puede suponer un gran avance en los campos en los que se esté trabajando.

Este tipo de tratamientos de datos y visualización se pueden aplicar a infinidad de campos, algunas de una importancia notable para la sociedad, alguno de estos ejemplos son campo de la Medicina, de salud y medioambiente y financieros.

En este proyecto se plantea trabajar con la información recogida de diferentes datos abierto (*open data [1]*). Este concepto de open data es fundamental para obtener diferentes perspectivas de múltiples áreas de trabajo. Poder trabajar con la información, centrándose en la manera en la que se analizan los datos. Concretamente, se utilizarán diferentes bases de datos abiertas suministradas desde internet, las cuales deben de tener una infraestructura que garantice la calidad y validación de la información.

Tanto la información recogida como las conclusiones que nos aporten los datos, deben de ser tratadas para que, a la hora de visualizar la información, tenga un formato simple e intuitivo para el usuario. Para ello, en el proyecto se utilizará la plataforma de *Microsoft PowerBi [1]*, para aportar servicios de análisis y de visualización.

En el proyecto una parte fundamental será la aportación de nuevas funcionalidades al sistema de PowerBi a partir de *Python [3]*. Se escoge Python como el lenguaje de programación del proyecto debido a que es el lenguaje mas utilizado en la actualidad, por la cantidad de funcionalidades extra que podemos introducir, a través de las diferentes bibliotecas y por su destacada conectividad con el software de PowerBi.

2. Objetivos y campo de aplicación.

El objetivo principal de este proyecto es trabajar sobre un conjunto fuentes de datos abiertos que están relacionados con proyectos de investigación en los que ha participado el departamento (SkillsMatch o Smacite), aplicando nuevas funcionalidades, basadas en Python, aportadas al Software de PowerBi para el tratamiento de datos y la visualización de ellos. Este objetivo principal está compuesto por subobjetivos más concretos. Estos subjetivos son los siguientes:

- La utilización de bases de datos abiertas (Open Data) de dominios relacionados con las ciudades inteligentes (Smacite) y con la salud y el medioambiente. Las administraciones públicas proporcionan una serie de bases de datos abiertas de múltiples campos de información. Algunos de estos campos pueden ser de información relacionada con accidentes de tráfico, la distribución de mobiliario urbano, valores climatológicos, …
- Relación al proyecto internacional "Eugloh", realizado por universidades europeas, incluida la Universidad de Alcalá de Henares. Eugloh es un proyecto relacionado con la salud global y Smart Cities.
- La utilización de bases de datos abiertas (Open Data) de dominios relacionados con el proyecto "SkillMatch". Consiste en identificar las habilidades personales orientadas al

trabajo para la recomendación de puestos de trabajos en toda Europa y cursos online para potenciar dichas habilidades.

- Realización de mejoras al actual proceso de extracción, transformación y cargar de datos (*ETL [4]*) por parte de PowerBi, están nuevas mejoras estarán basadas en Python.
- La implementación de técnicas avanzadas basadas en analítica de datos, desarrolladas en Python, sobre las bases de datos abiertas anteriormente descritas y su integración en PowerBi. A partir de la analítica de datos se obtienen conclusiones que aportan funcionalidad
- Visualizar la información recabada tras la implementación de técnicas avanzadas basadas en analítica de datos, de una forma intuitiva y compresible de procesar para el usuario en cuestión.

3. Descripción del trabajo.

En primer lugar, será necesario realizar un estudio en profundidad del funcionamiento del software a utilizar, de PowerBi. Es necesario tener unos conocimientos solidos para realizar los objetivos anteriormente descritos. Será preciso ver que funcionalidades implementa, aprender a manejarlas y estudiar que funcionalidades se pueden implementar.

Una vez que se tiene claro cómo funciona el software base, es necesario obtener los datos sobre los que se va a trabajar. Se seleccionan una serie de bases de datos abiertas, los dominios van a estar relacionados con el concepto de Smart Cities y Eugloh, de salud y medioambiente y con el concepto de SkillMatch. En esta selección es necesario analizar los atributos del conjunto de datos, examinar si existen posibles errores o inconsistencias.

Con la información con la que vamos a trabajar ya está definida, esta información se carga en PowerBi a través del proceso ETL que esta implementado en el software. El proceso ETL de PowerBi, permite la carga de la información representada de múltiples formas como pueden ser bases de datos SQL, archivos de Excel, incluso documentos PDF. Si alguna información, con la que se trabaje, está representada de una forma que el proceso ETL no es capaz de realizarla de una manera valida, será necesario una nueva implementación para llevar esta función a cabo.

A través de la información recogida en las bases de datos abiertas (Open Data) que se han escogido, se realizaran técnicas avanzadas en analítica de datos. Estas técnicas aportaran nueva funcionalidad de gran importancia a PowerBi. Las técnicas estarán desarrolladas en el lenguaje de programación Python. Python proporciona múltiples librerías de código abierto para el análisis de datos, como pueden ser TensorFlow, PyTorch y *Scikit-learn [7]*.

Python además es utilizada por incontables usuarios y entidades a todos los niveles profesionales, por lo que supone un fácil acceso a nueva información de gran ayuda a la hora de mejorar la implementación, en la resolución de dudas y determinación de errores. La conectividad que ofrece PowerBi con Python es de gran importancia, ya que permite ejecutar Scripts desde la propia interfaz del software. Además, Python tiene una proyección exponencial laboral.

Una vez se hayan desarrollado las técnicas de avanzadas en analítica de datos, y se hayan probado con la información procedente de las bases de datos abiertas, se obtendrán una serie de resultados. Estos resultados se analizarán en detalle, que proporcionarán una serie de conclusiones de vital importancia.

Estas conclusiones obtenidas a través del análisis de datos serán visualizadas en diferentes formas a través de PowerBi.

El proceso de realización del proyecto esquematizado es el siguiente:

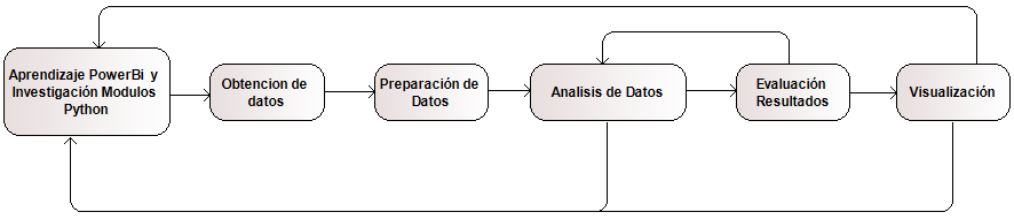

Figura 1: Diagrama de Bloques

Como se puede apreciar en el diagrama de bloques, en todo momento se aprenderán nuevas funcionalidades de PowerBi y se investigarán las diferentes bibliotecas de Python. Aunque en el momento inicial es donde más tiempo se invertirá en el estudio del software. Durante la evaluación de los resultados es posible que se obtengan conclusiones que pueden ser de vital importancia a la hora de como se ha enfocado el análisis de datos, por lo que, es posible que sea necesario volver al análisis de datos con la evaluación de unos primeros resultados completados.

4. Metodología y plan de trabajo.

Para el trabajo, se utilizará la metodología de *CRISP-DM*. Consiste en una metodología que divide el trabajo en diferentes fases y que es adecua perfectamente al proyecto. A continuación, se enumeran las 6 fases de *CRISP-DM*, junto con sus respectivas subtareas.

1. **Comprensión del Negocio**: Consiste en elaborar un plan para el proyecto, a partir de los objetivos identificados.
2. **Compresión de los datos**: Consiste en la recolección de datos y la compresión de estos. Se deberá realizar un análisis de los datos a utilizar para detectar los posibles problemas potencias a la hora de tratar con ellos, detectar valores atípicos y analizar la distribución de los datos.
3. **Preparación de los datos**: Consiste en limpiar los datos seleccionados, construir el formato de los datos y formar un conjunto de datos.
4. **Modelado**: Consiste en determinar la técnica avanzada de análisis de datos que se va a utilizar en el conjunto de datos, anteriormente preparado.
5. **Evaluación**: Consiste en evaluar y comparar resultados, para determinar que sucesos están implícitos en la información.
6. **Despliegue**: Organización y presentación de las conclusiones obtenidas a partir de la evaluación.

Una ventaja notoria de utilizar la metodología *CRISP-DM [8]*, es que permite moverse entre las distintas fases, cuando el proyecto lo requiere. Para facilitar la viabilidad se fija un hito, en mitad del tiempo destinado al desarrollo, donde se haya avanzado de forma de significado en todas las fases del proyecto

Las fechas del proyecto estarán basadas en el siguiente diagrama:

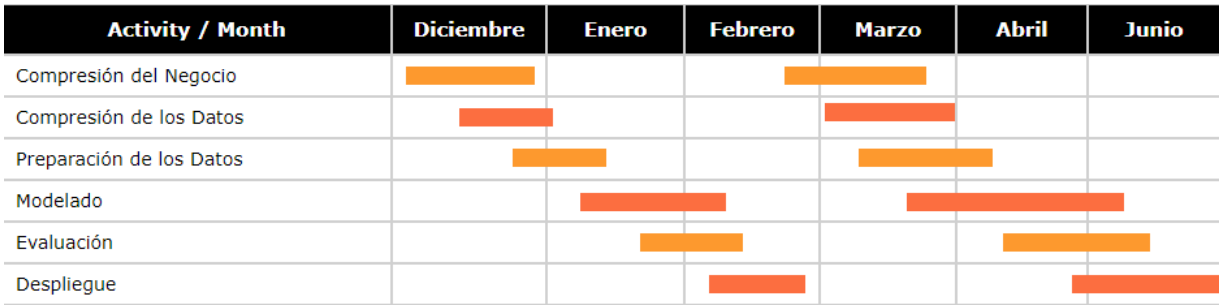

Figura 2: Diagrama de Gantt

5. Medios

Los medios que se van a utilizar en la realización del proyecto son los siguientes:

- Lenguaje de programación Python.
- Entorno de desarrollo Visual Studio Code.
- Software de visualización PowerBi.
- Conjunto de herramientas auxiliares para el tratamiento de datos Excel.
- Software de documentación Word.

6. Bibliografía

[1] Efor, Internet y Technologic, "Expertos en PowerBi", efor.es. [Online]. Available: https://www.efor.es/servicios/soluciones-de-software/power-bi.html?gclid=Cj0KCQiApb2bBhDYARIsAChHC9uw3mU6j3WDs4gRVzlLDG1rhruAxpxJ6bfo8FFcpBMcADZHq2P6B9waAsGIEALw_wcB%20s.f.

[2] AEMET OpenData, "centro de descargas", opendata.aemet.es. [Online]. Available: https://opendata.aemet.es/centrodedescargas/inicio

[3] AWS Amazon, "what is python", aws.amazon.com. [Online]. Available: https://aws.amazon.com/es/what-is/python/

[4] Gravitar, "etl-etl", gravitar.biz. [Online]. Available: https://gravitar.biz/bi/etl-elt/

[5] Jure Leskovec (Stanford Univ.), Anand Rajaraman (Milliway Labs) and Jeffrey D. Ullman (Stanford Univ), "Mining of Massive Datasets", Second Edition, pp. 241- 280. [Online]. Available: http://infolab.stanford.edu/~ullman/mmds/book.pdf

[6] Learn Microsoft, "explore analyze data with python", learn.microsoft.com. [Online]. Available: https://learn.microsoft.com/es-es/training/modules/explore-analyze-data-with-python/1-introduction

[7] Scikit-learn, "scikit-learn User Guide", scikit-learn.org. [Online]. Available: https://scikit-learn.org/stable/user_guide.html

[8] SNGULAR, "data science crisp dm metodologia", sngular.com. [Online]. Available: https://www.sngular.com/es/data-science-crisp-dm-metodologia/