ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

COMPARACIÓN DE EXACTITUD ENTRE LOS MODELOS DE CADENA DE MÁRKOV Y DE REGRESIÓN LINEAL PARA REALIZAR PREDICCIONES UTILIZANDO FUENTES DE DATOS INTEGRADAS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

CUMBICOS MIZHQUERO NATHALY FERNANDA
nathaly.cumbicos@epn.edu.ec

DIRECTOR: EDISON FERNANDO LOZA AGUIRRE
edison.loza@epn.edu.ec

Quito, Diciembre 2018

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Nathaly Cumbicos, bajo mi supervisión.

Edison Fernando Loza Aguirre
DIRECTOR DE PROYECTO

DECLARACIÓN DE AUTORÍA

Yo, Nathaly Cumbicos, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

_____

Nathaly Fernanda Cumbicos Mizhquero

DEDICATORIA

A mi mami Bachita y a mi papi Félix.

Por todo el tiempo, fortaleza y amor que supieron darme. Por la paciencia, por aceptarme tal cual soy, por enseñarme a trabajar para conseguir mis sueños, por enseñarme a ser mejor persona por sobre todas las cosas, por todo esto y muchísimo más. Esto es, enteramente por y para ustedes. Los amo con mi vida.

A mi hermana Martha y mis hermosos sobrinos.

Por ser mi heroína, por demostrarme que todo se puede con esfuerzo, dedicación y por amor a la familia. Cada paso que estás dando, te acerca más a todo eso que quieres lograr, estoy segura de que continuarás siendo mi ejemplo y el de tus hijos, si alguna vez piensas en decaer, recuerda cuan fuerte has sido y todo lo que tú y tu fortaleza han logrado. Aquí estoy para ti. Te amo ñaña.

AGRADECIMIENTO

Agradezco todo lo errado, todo lo acertado, todo lo aprendido. Agradezco a la Escuela Politécnica Nacional, porque en sus aulas adquirí el conocimiento necesario para desempeñarme como una profesional competente, porque me permitió conocer personas que marcaron mi vida, porque aprendí a valorar lo realmente importante, porque pude crecer como persona. Gracias.

Sin su apoyo, si ustedes no hubieran creído en mí, aún más de lo que yo misma, no lo habría logrado. Se los agradezco tanto, Bachita y Félix. Todos mis logros han sido y seguirán siendo de ustedes.

Gracias por motivarme, por ser paciente, por ayudarme tanto, por brindarme todo el tiempo que pudiste, tus esfuerzos contribuyeron en todo el desarrollo de este proyecto, todo valió la pena, muchas gracias Eli.

Agradecimientos especiales a usted Edison, gracias por integrarse al proyecto y ser mi guía en esta etapa final de mi carrera profesional universitaria.

ÍNDICE DE CONTENIDO

CERTIFICACIÓN.....I
DECLARACIÓN DE AUTORÍA II
DEDICATORIA III
AGRADECIMIENTO IV
ÍNDICE DE CONTENIDO V
RESUMEN..... VII
ABSTRACT VIII
1. INTRODUCCIÓN 1
1.1 Motivación 1
1.2 Iniciativa 2
2. METODOLOGÍA 4
2.1 Análisis predictivo..... 5
2.1.1 Requisitos de los datos..... 5
2.1.2 Recolección de los datos 9
2.1.3 Procesamiento de los datos 17
2.1.4 Limpieza de los datos 22
2.1.5 Modelado y Algoritmos 24
2.1.6 Comunicación 47
2.2 Desarrollo de software 49
2.2.1 Prototipo de interfaz..... 50
2.2.2 Prototipo vertical 53
3. RESULTADOS Y DISCUSIÓN 56
3.1 Resultados 56
3.1.1 Comportamiento de los datos con el modelo de Cadenas de Márkov 62

3.1.2 Comportamiento de los datos con el modelo de Regresión Lineal 65
3.2 Discusión 68
4. CONCLUSIONES 69
5. REFERENCIAS 71
6. ANEXOS 74
6.1 Tabla de figuras 74
6.2 Tabla de tablas 75
6.3 Anexo 1. Código para extracción de datos numéricos de indicadores 76
6.4 Anexo 2. Lista de países contemplados en la base de datos del Banco Mundial 78
6.5 Anexo 3. Organización y procesamiento de valores de temática 85
6.6 Anexo 4. Organización y procesamiento de valores numéricos de indicadores 86
6.7 Anexo 5.Organización y procesamiento de descripción de indicadores 88
6.8 Anexo 6. Limpieza de valores numéricos de indicadores 89
6.9 Anexo 7. Implementación de Regresión Lineal 92
6.10 Anexo 8. Implementación de Cadenas de Márkov 93
ORDEN DE EMPASTADO 97

RESUMEN

Realizar un análisis predictivo permite a una organización tomar decisiones estratégicas e informadas aprovechando sus propias colecciones de datos para realizar sus predicciones. Sin embargo, encontrar un modelo de predicción que ofrezca resultados óptimos, en cuanto a la exactitud de la predicción, puede representar un problema ya que el modelo puede no ajustarse al tipo de datos disponible o no ajustarse a las necesidades de predicción de la organización.

El objetivo de este proyecto es comparar las predicciones de valores continuos de indicadores de empleo en Ecuador para el año 2017 realizadas por dos modelos: Regresión Lineal y Cadenas de *Márkov*. Esto permitirá determinar cuál de estos modelos, ofrece mejores resultados en la exactitud de su predicción.

Para llevar a cabo el proyecto y obtener las predicciones de los modelos propuestos, se desarrollan las fases de la analítica de datos como metodología de trabajo para el análisis. Además, para la comunicación de los resultados del análisis y la predicción de los dos modelos se construirá un sistema web basado en la metodología de prototipado.

De nuestros resultados, verificamos que el modelo de Cadenas de *Márkov* es ligeramente más preciso que el modelo de Regresión Lineal para el conjunto de datos analizado. Además, se encontró que, del conjunto de datos analizado, el modelo de Cadenas de *Márkov* obtuvo mejores resultados, en cuanto a exactitud, en datos con tendencial no lineal.

Palabras Clave: análisis predictivo, Regresión lineal, Cadenas de Márkov, exactitud, predicción.

ABSTRACT

Performing a predictive analysis allows an organization to make strategic and informed decisions as they take advantage of their own collections of data. However, finding a prediction model that offers optimal results, in terms of prediction accuracy, can be a problem since the model may not fit the type of data available or may not fit the organization's prediction needs.

The objective of this project is to compare the predictions of continuous values of employment indicators in Ecuador for 2017 using two models: Linear Regression and Markov Chains. We aim to determine which of these models, offers more accurate predictions.

To carry out the project and obtain the predictions of the proposed models, we followed the phases of data analytics as work methodology for the analysis. In addition, for the communication of the results, a web system based on the prototyping methodology is built.

After analyzing our data, we concluded that the predictions made using Markov Chains model are slightly more exact than those of the Linear Regression model. In addition, it was found that, from the analyzed data set, the Markov Chains model obtained better results, in terms of accuracy, in data with a non-linear trend.

Key Words: Markov Chains, Linear Regression, predictive analysis, accuracy, prediction.

1. INTRODUCCIÓN

1.1 Motivación

Desde la recomendación de compras basadas en los artículos que los clientes escogieron en línea, hasta identificar a los pacientes hospitalarios que tienen un mayor riesgo de readmisión; hoy en día, el uso de herramientas y técnicas de análisis predictivo permiten a las organizaciones aprovechar sus colecciones de datos para predecir resultados futuros [1]. Estas predicciones, acompañadas de métodos avanzados de procesamiento estadístico y matemático de datos, enriquecen y permiten una toma de decisiones estratégica e informada. Actualmente, una empresa no solo requiere resumir su pasado (análisis descriptivo), sino también, establecer relaciones y comparaciones entre variables para hacer un análisis predictivo, que permita extraer conocimiento de su información como tendencias y patrones de comportamiento, contribuyendo de esta forma al proceso de toma de decisiones dentro de la organización [2].

En la actualidad, existen varios métodos que permiten realizar un análisis predictivo, sin embargo, esto a pesar de ser una ventaja, puede constituir también un problema dado que en ocasiones se desconoce cuál de todos los métodos disponibles ofrece mejores resultados de acuerdo con el tipo de datos disponibles, o incluso si estos modelos se ajustan o no a las necesidades de predicción de una organización [3]. Estudiar cómo realizar dichas predicciones con mayor exactitud es la problemática que pretende resolver el presente proyecto.

Para esto, se selecciona una base de datos *Linked Open Data*, datos libres y sin restricción de copyright provenientes de datos enlazados mediante tecnologías de la Web Semántica [4] para la obtención de los datos a predecir. En específico, se utilizaron datos referentes a la temática de Protección Social y Empleo en Ecuador. Se utilizan además dos modelos predictivos: Regresión Lineal y Cadenas de *Márkov*. La comparativa de los resultados de la implementación de ambos modelos sobre los datos, las predicciones, permitirán escoger la mejor opción a la hora de predecir y con ello una toma de decisiones más acertada.

1.2 Iniciativa

El proyecto pretende comparar la exactitud obtenida en las predicciones de dos modelos: Regresión Lineal y Cadenas de *Márkov* sobre datos de la temática de Protección Social y Empleo en Ecuador, para determinar cuál de los modelos ofrece mejores resultados en la exactitud de su predicción.

Para el desarrollo del proyecto se empleó la metodología de analítica de datos propuesta por [5], cuyas fases son:

- Requisitos de los datos
- Recolección de los datos
- Procesamiento de los datos
- Limpieza de los datos
- Análisis exploratorio de datos
- Modelamiento y Algoritmos
- Comunicación

En la fase de requisitos de los datos se identificaron los recursos necesarios para llevar a cabo el proyecto. Se identificó a una base de datos *Linked Open Data*, específicamente de la base de datos del Banco Mundial para recolectar datos continuos de indicadores de empleo en Ecuador para el periodo 2003-2017 [6]. Los datos luego de pasar por las etapas de procesamiento y limpieza fueron dispuestos en los modelos de Regresión Lineal [7] y Cadenas de *Márkov* [8] para realizar una predicción para el año 2017. Las predicciones fueron comparadas con los datos reales del 2017 para comparar la exactitud de estas predicciones.

En la fase de comunicación se desarrolla un sistema web empleando la metodología de prototipado en el desarrollo del software. En el sistema web se visualizan los resultados del análisis y la predicción de los dos modelos a manera de gráficos de líneas y tabulares

El trabajo final tiene como estructura: Metodología, Resultados y Discusión, Conclusiones, Referencias y Anexos.

La Metodología describe cada una de las fases de la metodología empleada para el análisis predictivo de los indicadores de Protección Social y Empleo en Ecuador, mencionadas previamente. En Resultados y Discusión se exponen los resultados obtenidos del análisis predictivo a manera de comparación entre los resultados de la predicción obtenidos por cada

modelo: Regresión Lineal y Cadenas de *Márkov*. En Conclusiones, se establece el modelo que ofrece mejores características de rendimiento en cuanto a exactitud de la predicción. Y finalmente, se adjuntan las Referencias del trabajo citado y los Anexos desarrollados en el proyecto.

2. METODOLOGÍA

La metodología propuesta es una adaptación de las fases del ciclo de vida de la analítica de datos, según [5]. Se excluye la fase de análisis exploratorio de datos, en la que se aplican técnicas para indagar sobre los datos tratados y entender mensajes contenidos en ellos [9], ya que se tiene certeza de la naturaleza de los datos y de lo que se desea predecir: datos de tipo numérico de indicadores de empleo en Ecuador para el año 2017.

A continuación, en la Figura 1, se presenta la arquitectura del sistema para predecir el valor de indicadores de empleo en Ecuador, adoptando las fases presentadas, en el apartado anterior.

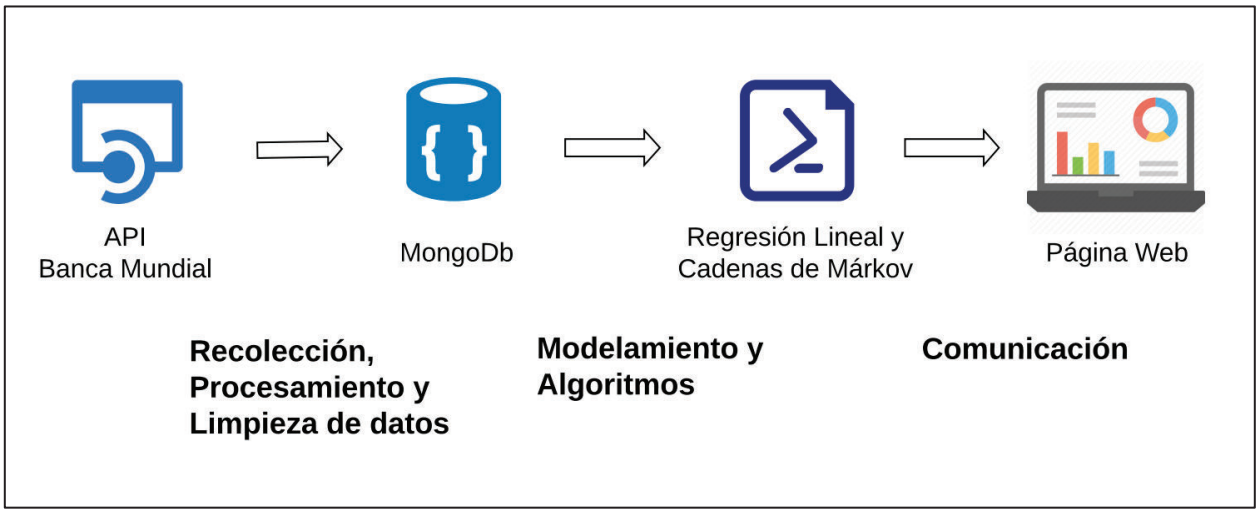

Figura 1. Arquitectura del sistema web para predicción de valores

Se definieron los requisitos de los datos a través de la búsqueda de datos confiables en fuentes integradas, escogiendo a la API del Banco Mundial por sus características de confiabilidad y completitud. Los datos se recolectan, procesan y limpian para ser almacenados en una base de datos no relacional (MongoDb). Después, los datos son guardados para evitar acceder constantemente a su fuente original ya que esto requirió mayor tráfico en la red. Además, de requerir mayor tiempo dedicado a la descarga de datos de forma continua. La siguiente fase fue la creación de los modelos de Regresión Lineal y Cadenas de *Márkov* y se determina la exactitud de sus predicciones, utilizando scripts de Python. Y finalmente la fase de

comunicación en donde una aplicación web recoge los resultados obtenidos previamente y permite visualizarlos a manera de gráficos lineales y tabulares.

El desarrollo de este proyecto consta de las siguientes secciones: Requisitos de los datos, donde se definen las características de los datos necesarias para el desarrollo del proyecto, Recolección de los datos, en la cual los datos son recogidos de fuentes integradas para su posterior tratamiento. Procesamiento de los datos, en la que los datos recolectados son organizados dentro una base de datos noSQL para facilitar el análisis. Limpieza de los datos, en la cual se eliminan datos duplicados o incompletos. Modelado y Algoritmos, fase en donde se explican, implementan y evalúan cada modelo predictivo en cuanto a la exactitud obtenida en la predicción. Y finalmente Comunicación, fase de visualización de resultados a través de una interfaz web.

2.1 Análisis predictivo

2.1.1 Requisitos de los datos

En esta etapa se evalúan e identifican los recursos necesarios para respaldar el proyecto. En este contexto, los recursos incluyen fuentes y características de los datos, la tecnología y herramientas para su manejo [10]. Las fuentes de los datos, dependiendo del alcance del proyecto, pueden ser internas a la organización, como los almacenes de datos, o externas como proveedores de datos, blogs, sitios web basados en contenidos, entre otros [11].

Los datos necesarios para el análisis predictivo deben contar con características de veracidad y completitud, ya que de esta forma se asegura calidad en la evaluación de la exactitud de los modelos. Considerando estas características, se realizó una búsqueda en un repositorio de bases de datos de *Linked Open Data,* datos libres y sin restricción de copyright provenientes de datos enlazados mediante tecnologías de la Web Semántica [4]. Los repositorios de *Linked Open Data* clasifican sus bases de datos por temática, para facilitar la búsqueda (Ver Figura 2).

Figura 2. Temáticas de *Linked Open Data*

Dentro de la temática de Gobierno, podemos encontrar la base de datos del Banco Mundial que aborda temáticas de interés como: educación, desarrollo social, energía y minería, protección social y trabajo, entre otros. En todos los países del mundo y cuyos datos son recogidos de diversas fuentes. Las fuentes se muestran en la Figura 3.

La base de datos del Banco Mundial a través de su API, un conjunto de rutinas, protocolos y herramientas para crear aplicaciones de software [12], proporciona acceso a sus datos integrados provenientes de organizaciones como: UNICEF, Organización Internacional del trabajo, Organización Mundial de la Salud, Fondo de Naciones Unidas para la infancia, entre otras. La Figura 3 es un extracto de las organizaciones que recogen información periódicamente sobre temáticas de importancia mundial como: pobreza, salud, educación, población, empleo, entre otras. Por tanto, se consideran veraces los datos de dichas fuentes, ya que son avalados por organizaciones reconocidas y que trabajan a nivel mundial.

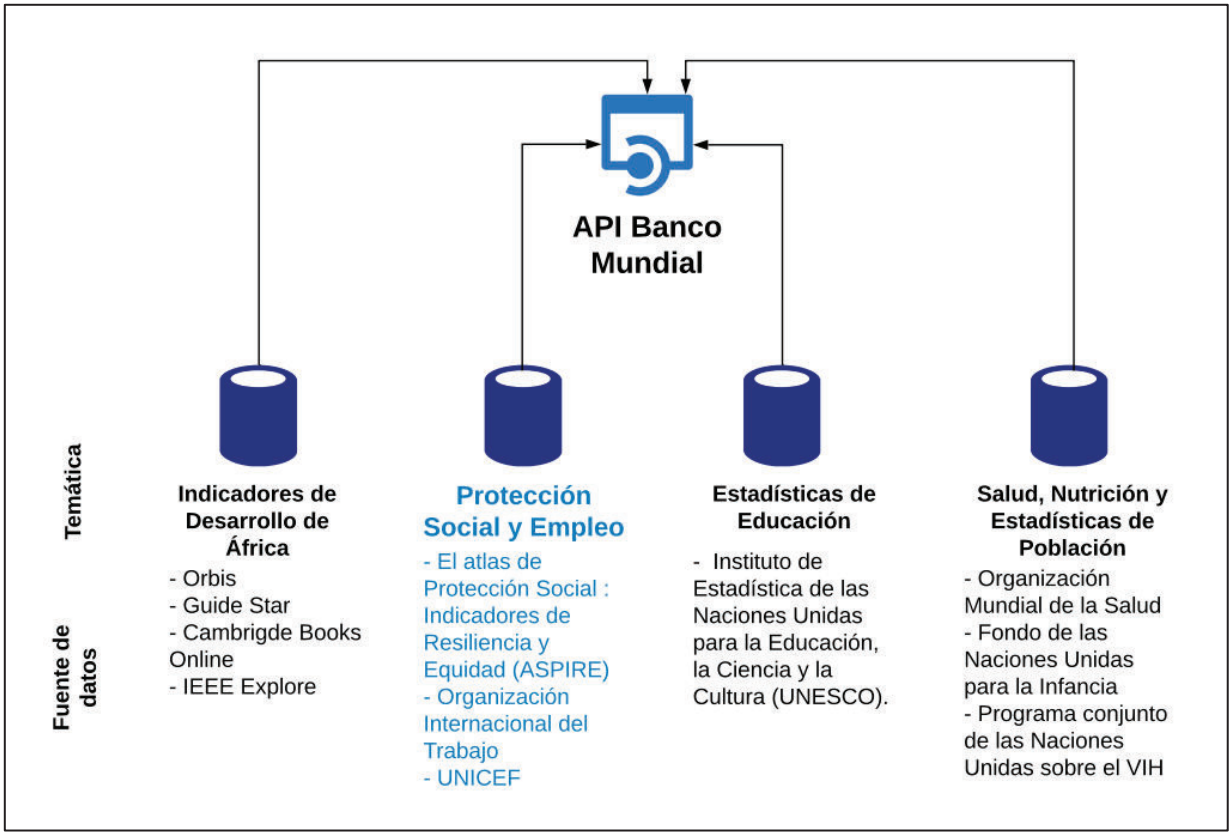

Figura 3. Elaboración propia basada en la referencia (World Bank, 2018)

Es posible utilizar los datos de cualquier temática contenida en la base de datos del Banco Mundial, ya que cada temática está descrita en base a un conjunto de indicadores y estos indicadores contienen datos continuos, característica indispensable para generar los modelos de Regresión Lineal y Cadenas de *Márkov*.

El empleo en Ecuador es una temática abordada por Protección Social y Empleo señalada en la Figura 3. Esta temática fue escogida por presentar el conjunto de datos más completo en el período 2003-2017, en comparación con otras temáticas tales como: pobreza y salud. Entendiendo como conjunto completo de datos, al conjunto en el que, para cada año dentro del periodo seleccionado, existe un valor numérico relacionado a un indicador de empleo. El período de tiempo se establece hasta el año 2017, ya que es hasta este año que se tienen datos para los indicadores en la base de datos de la Banca Mundial, datos con los que se

puede comparar a la predicción realizada por los modelos, de otro modo, la comparación no tendría sentido, pues no se tendría un dato real con cual realizar la comparación.

En la Tabla 1, se observa la lista de temáticas comprendidas en la base de datos del Banco Mundial, siendo *Social Protection & Labor* la que se refiere al empleo.

Tabla 1. Temáticas de la Base de datos del Banco Mundial

ID	Nombre Temática
1	Agriculture & Rural Development
2	Aid Effectiveness
3	Economy & Growth
4	Education
5	Energy & Mining
6	Enviroment
7	Financial Sector
8	Health
9	Infraestructure
10	**Social Protection & Labor**
11	Poverty
12	Private Sector
13	Public Sector
14	Science & Technology
15	Social Development
16	Urban Development
17	Gender
18	Millenium Developmente Goals
19	Climate Change
20	External Debt
21	Trade

2.1.2 Recolección de los datos

La recolección de datos implica la recopilación de los mismos, tomando en cuenta tanto las características como las fuentes de datos definidas en la etapa Requisitos de los datos. Para ello, se utilizaron herramientas como scripts de Python, R, SQL y una combinación de ellas [5].

En esta fase se utilizó, el API del Banco Mundial que proporciona una librería llamada wbdata, escrita en Python y que permite buscar y solicitar información de todas sus fuentes de datos integradas [13]. Los datos integrados del Banco Mundial están organizados por temáticas, en donde cada temática cuenta con su propio conjunto de indicadores que la describen y de los cuales se puede obtener su valor numérico, tomando al año como unidad de medida de tiempo.

2.1.2.1 Recolección de datos de Temática

La función utilizada para obtener información de la temática fue *get_topic*. La estructura de la función es la siguiente: ***get_topic(topic_id=None, display=None)*** de acuerdo con [13]**.**

La Tabla 2 muestra una descripción y el tipo de dato de acuerdo con el parámetro de la función *get_topic*:

Tabla 2. Descripción de parámetros de la función de *wbdata get_topic*

Parámetro	**Descripción**	**Tipo de Dato**
topic_id	Código de la temática de la cual se requiere información.	Int
display	Determina de qué forma se presentan los datos, si es *true* se muestra el id de la temática y su nombre, mientras que si es *false* se muestra un objeto JSON que añade una descripción extraída de la fuente de datos de la temática, además de los campos de id y su nombre.	Boolean

La Figura 4 muestra, cómo se recolectan los datos para la temática escogida, Protección Social y Empleo.

```
#Autor: Nathaly Cumbicos
#Fecha Creación: 15-abril-2018
#Archivo: recolectTopic.py

#Librerías
import wbdata

#Obteniendo datos de temática: Protección Social y Empleo
#Campos: id, value, sourcenote
#Temática-Código: Social Protection & Labor-10

topics = [] #Lista que guarda información de las temáticas
topicList = [] #Lista que extrae los datos del API
#función get_topic(codigo_tematica=10, mostrarsoloId=False)
topicList = wbdata.get_topic(10,display=False)

#Secuencial para id
cont = 0
#Recorriendo la lista de temáticas
for item in topicList:
    #Creando diccionario topic que guarda la data extraída
    topic = {}
    cont +=1
    topic ["id"] = cont
    topic ["id_topic"] = item["id"]
    topic ["name"] = item["value"]
    topic ["description"] = item["sourceNote"]
    #Guardando el diccionario topic en lista topics
    topics.append(topic)
```

Figura 4. Script en Python de Recolección de datos de temática Protección Social y Empleo

En la Figura 4 se presenta la implementación de la función *get_topic,* para obtener datos de id, nombre y descripción de la temática Protección Social y Empleo cuyo código es el 10, de acuerdo con la Tabla 1. Los campos de id, nombre y descripción fueron estructurados en un diccionario, para facilitar la organización y almacenamiento dentro de la base de datos de MongoDb, de la cual se hablará con más detalle en la etapa de Procesamiento de los datos. En la Figura 5, se muestra el resultado de la ejecución del código presentado en la Figura 4, un diccionario con los datos proporcionados por el API de la temática "*Social Protection & Labor*".

```
[
  {
    "id": 1,
    "id_topic": "10",
    "name": "Social Protection & Labor",
    "description": "The supply of labor available in an economy includes people who are employed,
      those who are unemployed but seeking work, and first-time job-seekers. Not everyone who
      works is included: unpaid workers, family workers, and students are often omitted, while
      some countries do not count members of the armed forces. Data on labor and employment are
      compiled by the International Labour Organization (ILO) from labor force surveys, censuses,
      establishment censuses and surveys, and administrative records such as employment exchange
      registers and unemployment insurance schemes."
  }
]
```


Figura 5. Contenido de la lista topics de la temática protección social y empleo

2.1.2.2 Recolección de datos de indicadores

Las características necesarias para los indicadores de la temática de Protección Social y Empleo son sus descripciones y su valor numérico medido por año. Las funciones utilizadas para esta fase fueron:

a) search_indicators
b) get_indicator
c) get_data

a) La función *search_indicators*, facilita la búsqueda de indicadores. Su estructura es la siguiente:

search_indicators(query, source=None, topic=None, display=None) de acuerdo con [13].

La Tabla 3 muestra una descripción y el tipo de dato de acuerdo con el parámetro de la función *search_indicators*:

Tabla 3. Descripción de parámetros de la función de wbdata search_indicators

Parámetro	Descripción	Tipo de Dato
Query	Palabra o nombre que se relacione al indicador como criterio de búsqueda.	String
source	Código de la fuente de datos del indicador.	String

topic	Código de la temática.	Int
display	Determina de qué forma se presentan los datos, si es true se muestra el id de indicador y su nombre, mientras que si es false se muestra un objeto JSON que añade una descripción extraída de la fuente de datos de la temática, además de los campos de id y su nombre.	Boolean

b) La función que permite obtener información de los indicadores en cuanto a su descripción es *get_indicator*, cuya estructura es:

get_indicator(indicator=None, source=None, topic=None, display=None) de acuerdo con [13]**.**

La Tabla 4 muestra una descripción y el tipo de dato de acuerdo con el parámetro de la función *get_indicator*:

Tabla 4. Descripción de parámetros de la función de wbdata get_indicator

Parámetro	**Descripción**	**Tipo de Dato**
indicator	Código del indicador del que se requiere información	String
source	Código de la fuente de datos del indicador	String
topic	Código de la temática	Int
display	Determina de qué forma se presentan los datos, si es true se muestra el id de indicador y su nombre, mientras que si es false se muestra un objeto JSON que añade una descripción extraída de la fuente de datos de la temática, además de los campos de id y su nombre.	Boolean

En la Figura 6 se presenta el script sobre cómo se recolectaron los datos de indicadores de la temática Protección Social y Empleo. Además, la codificación de la búsqueda de indicadores relacionados a la palabra "*employ*" y la temática Protección Social y Empleo a través de la función *search_indicators*. El resultado de esta búsqueda sirve como entrada para la función de extracción de la descripción del indicador mediante la función *get_indicator.* No todos los indicadores obtenidos en la búsqueda poseen datos de la descripción del indicador dentro del

API, por tal razón, esos datos se descartan y luego se almacenan en una lista de indicadores obtenidos con error denominada *indicatorFailing*.

```
#Autor: Nathaly Cumbicos
#Fecha Creación: 15-abril-2018
#Archivo: recolectIndicatorsData.py
#Librerías
import wbdata
import datetime

#Obteniendo datos de concepto de indicadores
#Temática: Protección Social y Empleo
#Campos: id, name, source, sourceNote,
#sourceOrganization, topic
#Temática-Código: Social Protection & Labor-10

#Buscando indicadores relacionados con empleo en el API
#función
#search_indicators(criteriobusqueda="employ",codigoTematica=10,mostrarsolo
Id=False)
employment = wbdata.search_indicators("employ",topic=10,display=False)
#Lista de Ids de indicadores obtenidos de la búsqueda
employmentList = []

#Recorriendo lista employment
for indicator_employment in employment:
    employmentList.append(indicator_employment["id"])#Almacenando
únicamente id

#Lista que contiene los datos de indicadores obtenidos de la búsqueda
indicatorList = []

#Lista con indicadores que no tienen datos o falla su extracción de datos
indicatorFailing = []

#Lista que ejecuta la función de extracción de datos del indicador
(get_indicator)
data = []

#Recorriendo lista employmentList
for indicatorID in employmentList:
    try:
        #Ejecución función get_indicator
#get_indicator(codigoIndicador=idsObtenidosBusqueda,codigoTematica="10",mo
strarsoloId=False)
        data =
wbdata.get_indicator(indicator=indicatorID,topic="10",display=False)
        if data is not None: #Datos no vacíos
            indicatorList.append(data) #Guardando los datos de indicadores
en indicatorList
    except Exception: #Datos con error
        indicatorFailing.append(indicatorID) #Guardando id de indicador en
indicatorFailing
```

Figura 6. Script en Python de recolección de datos de indicadores

En la Figura 7 se muestra un extracto del resultado de la ejecución del código mostrado en la Figura 6. El indicador que se muestra en la Figura 7, tiene el código (id): "SL.UEM.NEET.FE.ZS", *name* representa el nombre del indicador, *source* la fuente de datos de donde proviene, con su código (id) y su nombre (*value*), *sourceNote* es la descripción del indicador, *sourceOrganization* la organización que provee los datos y topics, la temática que contempla al indicador con su código (id) y su nombre (*value*).


```
[
  {
    "id": "SL.UEM.NEET.FE.ZS",
    "name": "Share of youth not in education, employment or training, female (% of female youth
      population)",
    "source": {
      "id": "2",
      "value": "World Development Indicators"
    },
    "sourceNote": "Share of youth not in education, employment or training (NEET) is the
      proportion of young people who are not in education, employment, or training to the
      population of the corresponding age group: youth (ages 15 to 24); persons ages 15 to 29; or
      both age groups.",
    "sourceOrganization": "International Labour Organization, ILOSTAT database. Data retrieved in
      September 2017.",
    "topics": [
      {
        "id": "10",
        "value": "Social Protection & Labor"
      }
    ]
  }
]
```


Figura 7. Descripción del indicador con código id: "SL.UEM.NEET.FE.ZS"

c) Los valores numéricos del indicador para cada año del periodo 2003-2017, se obtuvieron con la función *get_data*, cuya estructura es la siguiente:

get_data(indicator, country=u'all', data_date=None, convert_date=False, pandas=False, column_name=u'value', keep_levels=False) de acuerdo con [13]**.**

La Tabla 5 muestra una descripción y el tipo de dato de acuerdo con el parámetro de la función *get_data*:

Tabla 5. Descripción de parámetros de la función de wbdata get_data

Parámetro	Descripción	Tipo de Dato
indicator	Código del indicador del que se requiere información	String

Country	Código del país del indicador, descrito en 3 letras.	String
data_date	Periodo de tiempo en el cual se miden los indicadores	Date
convert_date	Si es true, se convierte al periodo de tiempo en meses y días.	Boolean
Pandas	Si es True, devuelve resultados como una serie de pandas. El índice será la fecha de los datos si solo se especifica un país, los países si solo se especifica una fecha, o un multi-índice de país y la fecha de lo contrario.	Boolean
column_name	Nombre deseado para la columna de pandas	String
keep_levels	Si el parámetro pandas es true y keep_levels también, entonces no se reducen el número de niveles si solo se obtiene una fecha o un país.	Boolean

La Figura 8 muestra la extracción de los datos numéricos de los indicadores de Protección Social y Empleo para el período entre 2003-2017 y para Ecuador. Se definen como criterios de búsqueda al periodo 2003-2017 y al país Ecuador, pudiendo escoger cualquier país listado en el Anexo 2. Lista de países contemplados en la base de datos del Banco Mundial, para la ejecución de la función *get_data.* Así, se obtienen los campos *id, value y date* que corresponden al código del indicador, el valor numérico del mismo y el año en que fue medido. Cabe recalcar que al igual que en la recolección de datos de descripciones, existen indicadores que no poseen datos en su descripción por lo que se descartan y se almacenan en la lista denominada *indicatorFailing*: `indicatorFailing.append(indicatorID).`

```
#Autor: Nathaly Cumbicos
#Fecha Creación: 10-mayo-2018
#Archivo: baseValueIndicators.py


#Ver Anexo 1.
…

#Definiendo criterios de búsqueda
#Período de tiempo: 2003-2016
#País: ECU

data_date = (datetime.datetime(2003, 1, 1), datetime.datetime(2017, 1, 1))
country = 'ECU'

#Lista que contiene los datos de indicadores obtenidos de la búsqueda
```

```
indicatorList = []


#Lista con indicadores que no tienen datos o falla su extracción de datos
indicatorFailing = []


#Lista que ejecuta la función de extracción de datos numéricos del
indicador (get_data)
data = []


#Recorriendo lista employmentList
for indicatorID in employmentList:
    try:
        #Ejecución función get_data

#get_data(codigoIndicador=idsObtenidosBusqueda,codigoPais=ECU,data_date=20
03-2016)
        data =
wbdata.get_data(indicatorID,country=country,data_date=data_date)
        if data is not None: #Datos no vacíos
            indicatorList.append(data) #Guardando los datos numéricos en
indicatorList
    except Exception:
        indicatorFailing.append(indicatorID) #Guardando id de indicador en
indicatorFailing
```

Figura 8. Extracto del Anexo 1, código para recolectar datos numéricos de indicadores

La Figura 9, muestra los valores para dos de los años dentro del periodo 2003-2017 para el indicador con código (id) “SL.SRV.EMPL.FE.ZS”, siendo esta la misma estructura con la que se almacenan los datos numéricos de los indicadores en la base de datos MongoDb.

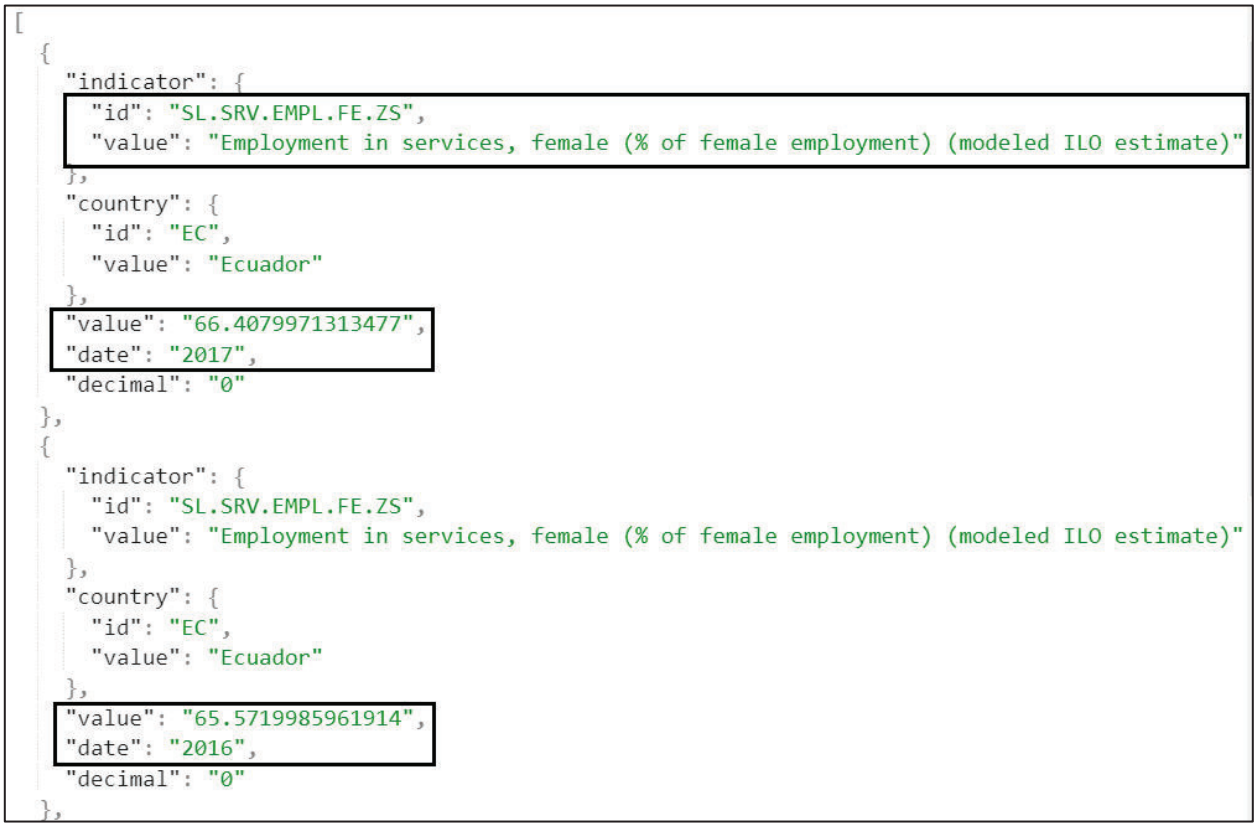

Figura 9. Extracto de los valores numéricos del indicador con id: “SL.SRV.EMPL.FE.ZS”

2.1.3 Procesamiento de los datos

Una vez recolectados los datos necesarios para el análisis, la siguiente fase corresponde al tratamiento y organización de los mismos para su análisis [5].

Para la descripción de la organización de los datos se ha divido a esta fase en 3 subsecciones: la subsección de **Descripción de la estructura de la base** que detalla lo utilizado para el almacenamiento de los datos, la subsección **Temática** que detalla la forma de almacenamiento de las temáticas en la base de datos y la subsección **Indicadores** que detalla la forma de almacenamiento de los datos de los indicadores (descripción y valor numérico).

2.1.3.1 Descripción de la estructura de la base

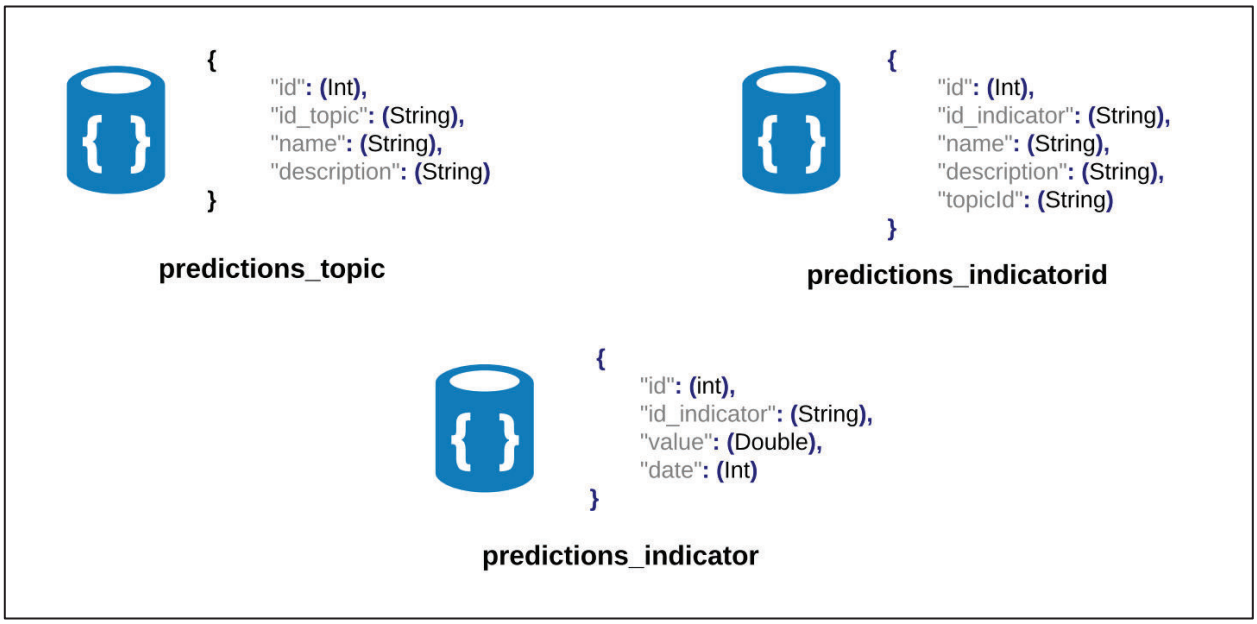

Figura 10. Colecciones de la base de datos *indicators*

Como se había indicado anteriormente, para la etapa de organización y almacenamiento se escogió la base de datos noSQL, MongoDb, una base de datos no relacional orientada a documentos, de fuente abierta, gratuita, escalable y flexible [14]. MongoDb almacena sus registros (denominados documentos) como objetos JSON dentro de sus colecciones (el equivalente de tabla en SQL). La estructura de la base de datos creada para almacenar los datos de temática e indicadores se muestra en la Figura 10.

A la base de datos en MongoDb de los indicadores se le llamó como ***Indicators***. *Indicators* contiene 3 colecciones:

- *predictions_topic*, en donde se almacenan los datos de la temática Protección Social y Empleo
- *predictions_indicatorid*, en donde se almacenan los datos de la descripción de los indicadores correspondiente a la misma temática
- *predictions_indicator*, que almacena los datos numéricos de los indicadores medidos por año.

Los campos de cada colección son descritos a continuación.

La Tabla 6 muestra una descripción de los campos de la colección *predictions_topic* y su tipo de dato.

Tabla 6. Descripción de campos de la colección *predictions_topic*

Campo	Descripción	Tipo de dato
Id	Secuencial de la temática	Int
Id_topic	Código de temática	String
Name	Nombre de temática	String
Description	Breve descripción de la temática	String

La Tabla 7 muestra una descripción de los campos de la colección *predictions_indicatorid* y su tipo de dato.

Tabla 7. Descripción de campos de la colección *predictions_indicatorid*

Campo	Descripción	Tipo de dato
Id	Secuencial del indicador	Int
id_indicator	Código del indicador	String
Name	Nombre del indicador	String
Description	Breve descripción del indicador	String
topicId	Código de la temática al que el indicador pertenece	String

La Tabla 8 muestra una descripción de los campos de la colección *predictions_indicator* y su tipo de dato.

Tabla 8. Descripción de campos de colección *predictions_indicator*

Campo	Descripción	Tipo de dato
Id	Secuencial de indicador	Int
Id_indicator	Código del indicador	String
Value	Valor numérico del indicador	Double
Date	Año	Int

2.1.3.2 Temática

Los scripts de Python utilizados en la **Recolección de los datos** tuvieron el fin de establecer una conexión con la base de datos *Indicators* para almacenar los datos de la temática de Protección Social y Empleo como documento con la instrucción *insert* en la colección *predictions_topic*, como se muestra en la Figura 11.

```
…
27  for item in topicList:
28      #Creando diccionario topic que guarda la data extraída
29      topic = {}
…
…
35      #Guardando el diccionario topic en lista topics
36      topics.insert(topic)
```

Figura 11. Extracto de Anexo 3. Organización y procesamiento de datos de temática.

En la Figura 12, se observa la temática de Protección Social y Empleo almacenada en la base de datos *Indicators*, de acuerdo con lo ilustrado en la Figura 11. La herramienta que se utilizó para visualizar las temáticas insertadas en la base de datos fue MongoDB Compass, una herramienta que permite a los usuarios analizar y comprender fácilmente los contenidos de sus colecciones de datos dentro de MongoDB y realizar consultas, sin necesidad de conocer la sintaxis de la consulta de MongoDB [14].

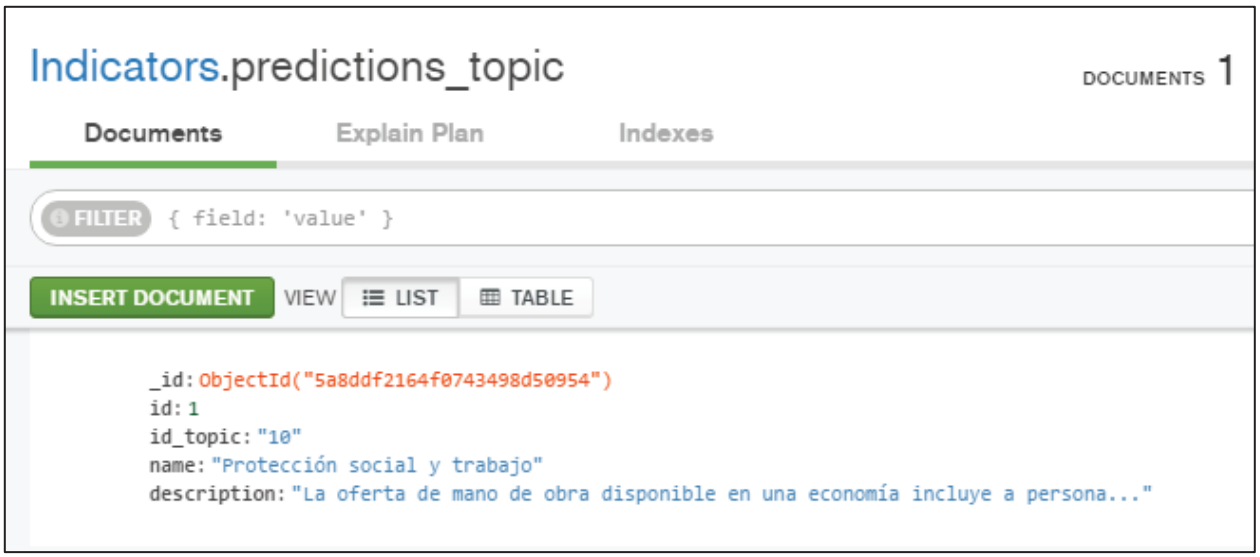

Figura 12. Colección *predictions_topic*

2.1.3.3 Indicadores

a) Valores numéricos

Los valores numéricos medidos por años de los indicadores, en la colección *predictions_indicator*, se redondean a dos decimales como se observa a continuación a través de la función *round*:

```
…
87      indicator ["value"] = round(float(valueIndicator["value"]),2)
…
```

Figura 13. Extracto de Anexo 4. Organización y procesamiento de valores numéricos de indicadores.

b) Descripción

Los datos de descripción de los indicadores de Protección Social y Empleo son almacenados en la colección *predictions_indicatorid*, de la misma forma que los datos de temática, utilizando la instrucción *insert.* Ver el Anexo 5.Organización y procesamiento de descripción de indicadores.

Los datos inicialmente se almacenaron en el idioma inglés. Al tener un conjunto de datos de 60 indicadores, se realizó una traducción mediante la herramienta de Google de traducción (*Google translate*). En la Figura 14, se muestran los datos de descripción de la temática de Protección Social y Empleo antes y después de la traducción. Los datos se muestran en la herramienta *MongoDB Compass*.

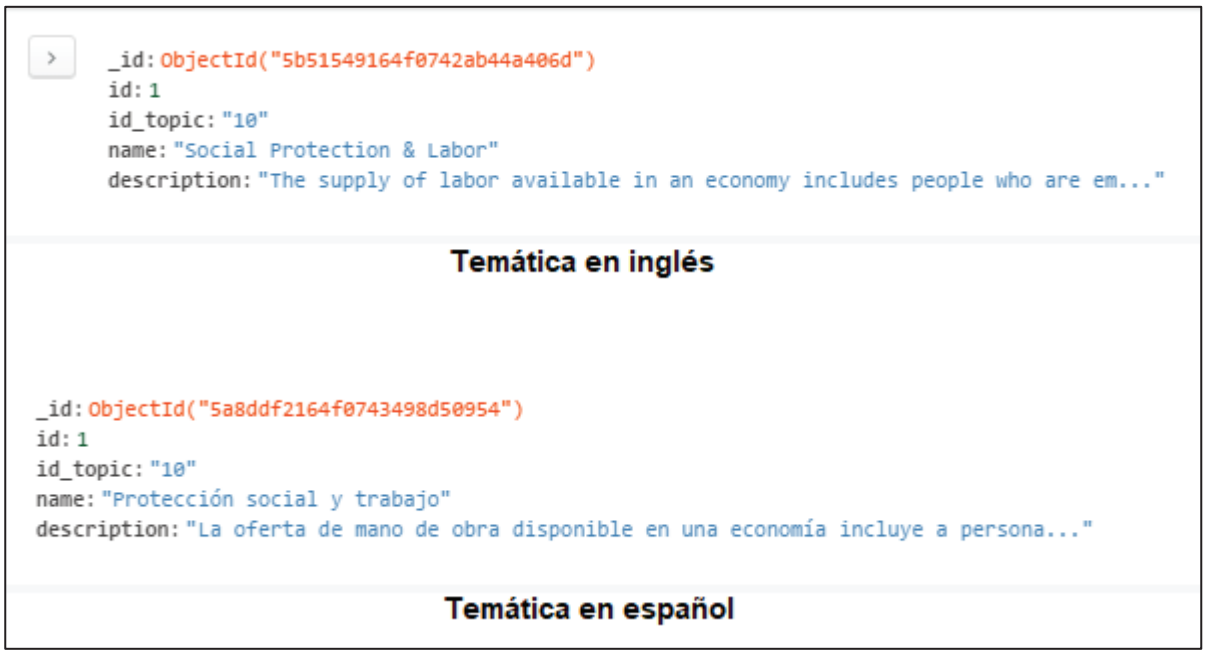

Figura 14. Datos de descripción de Protección Social y Empleo antes y después de traducción

2.1.4 Limpieza de los datos

Los datos procesados y organizados en la **Recolección de los datos** podrían contener datos duplicados, incompletos o erróneos, lo cual podría sesgar los resultados del análisis [11]. Para evitar el sesgo, los datos de los indicadores procesados fueron sometidos a un proceso de limpieza que permitió tener un conjunto de datos válido y de calidad para realizar un análisis más confiable. El proceso de limpieza consistió en 2 partes: **Requisitos de validez y calidad de los datos** y **Eliminación de datos no válidos**.

2.1.4.1 Requisitos de validez y calidad de los datos

Según [15], la calidad se define como el grado en el que un conjunto de características inherentes a un objeto (producto, servicio, proceso, persona, organización, sistema o recurso) cumple con los requisitos. De esta forma se definen los requisitos que los datos, después de haber sido recolectados y procesados, deben cumplir para asegurar calidad en la evaluación de los modelos. Los requisitos se describen a continuación:

- Tener valores numéricos para al menos 3 de los años del periodo establecido entre el año 2003-2017, ya que 3 es el mínimo número de observaciones requeridas para aplicar un modelo de regresión lineal a los datos [16].
- Tener un valor numérico para el año en el que se realizará la predicción, en este caso 2017, ya que este valor representa el valor real con el que se realizará la comparación del valor real vs. valor predicho por los modelos.
- Tener valores numéricos cuya medida de tiempo sea anual, ya que el conjunto de datos en su mayoría tiene indicadores anuales. Si se encuentran datos con otras medidas de tiempo, como trimestral o semestralmente, estos no serán tomados en cuenta.

2.1.4.2 Eliminación de datos no válidos

Para dar cumplimiento a los requisitos establecidos anteriormente, se aplicaron algoritmos de limpieza de datos para cada uno de los requisitos:

- **Eliminar indicadores que no tengan datos numéricos para más de 3 años.**

  Verificando que la cantidad de valores numéricos almacenados en la colección *predictions_indicator*, sea mayor a 3 (ver Figura 15).

```
...
62        if len(item) >= 3:
...
```

Figura 15. Extracto de Anexo 6. Limpieza de valores numéricos de indicadores

- **Eliminar los datos de indicadores que no posean valor numérico para el año 2017.**
  Verificando que el valor numérico para el año 2017 no sea nulo (ver Figura 16).

```
...
73   if (dataIndicator["value"] is None or dataIndicator["date"] !=
'2017'):
...
```

Figura 16. Extracto de Anexo 6. Limpieza de valores numéricos de indicadores

- **Eliminar los datos de indicadores cuya medida de tiempo no sea anual.** Verificando que la longitud del campo *date* sea igual a 4, que corresponde a la longitud de un año (ver Figura 17).

```
…
88          if (len(valueIndicator["date"])==4):
…
```

Figura 17. Extracto de Anexo 6. Limpieza de valores numéricos de indicadores

Con las modificaciones realizadas, el conjunto de datos cumplió con los requisitos de validez y calidad. En la Figura 18 se muestra un extracto de la lista de 60 indicadores que describen a la temática de Protección Social y Empleo. De estos, se toma al indicador con nombre: "Empleados de servicios, varones (% de empleo masculino)", como indicador del cual se toman los datos para la implementación de los modelos de Regresión Lineal y Cadenas de *Márkov* en la siguiente etapa **Modelado y Algoritmos**.

#	ID	Nombre del Indicador
1	SL.SRV.EMPL.MA.ZS	Empleados de servicios, varones (% de empleo masculino)
2	SL.SRV.EMPL.FE.ZS	Empleados de servicios, mujeres (% del empleo femenino)
3	SL.ISV.IFRM.MA.ZS	Empleo informal, masculino (% del total del empleo no agrícola)
4	SL.ISV.IFRM.FE.ZS	Empleo informal, femenino (% del total del empleo no agrícola)
5	SL.IND.EMPL.ZS	Empleos en la industria (% del total de empleos)
…	…	…
…	…	…
…	…	…
…	…	…
60	SL.FAM.WORK.MA.ZS	Trabajadores familiares contribuyentes, hombres (% del empleo masculino)

Figura 18. Extracto de la colección *predictions_indicatorid*

2.1.5 Modelado y Algoritmos

En esta etapa se implementan los algoritmos de los modelos escogidos según el tipo de problema de analítica que se intenta resolver: predicción o descripción [17].

Este proyecto aborda un tipo de problema de predicción, por lo que se realiza un análisis predictivo mediante el cual se intenta determinar el resultado de un evento que pueda ocurrir

en el futuro. Estas predicciones se basan en patrones, tendencias y excepciones que son encontrados en datos históricos y actuales [11].

En el análisis predictivo se pueden utilizar datos de tipo continuo o categórico [17]. En nuestro caso, y por tratarse de datos numéricos provenientes de indicadores de la temática Protección Social y Empleo en Ecuador que corresponden a datos continuos, se seleccionaron los modelos de Regresión Lineal y Cadena de *Márkov* como modelos para realizar la predicción, cuya implementación es descrita en las siguientes subsecciones.

2.1.5.1 Regresión Lineal

Las regresiones son una de las herramientas principales de la estadística referencial cuyo objetivo es describir la relación entre un conjunto de variables, denominadas variables dependientes o de respuesta, y otro conjunto de variables, denominadas variables independientes o predictoras [7].

La regresión lineal simple, denominada así, ya que se tiene solamente una variable independiente, asume que la relación entre las variables dependiente e independiente se describe con la siguiente ecuación (1):

$$Y = aX + b \tag{1}$$

Donde:

- a, b son los coeficientes del modelo lineal, que definen pendiente y la intersección con el eje y, respectivamente.
- Y, variable dependiente o de respuesta
- X, variable independiente o predictora

En la Figura 19, se visualizan los elementos de línea de regresión.

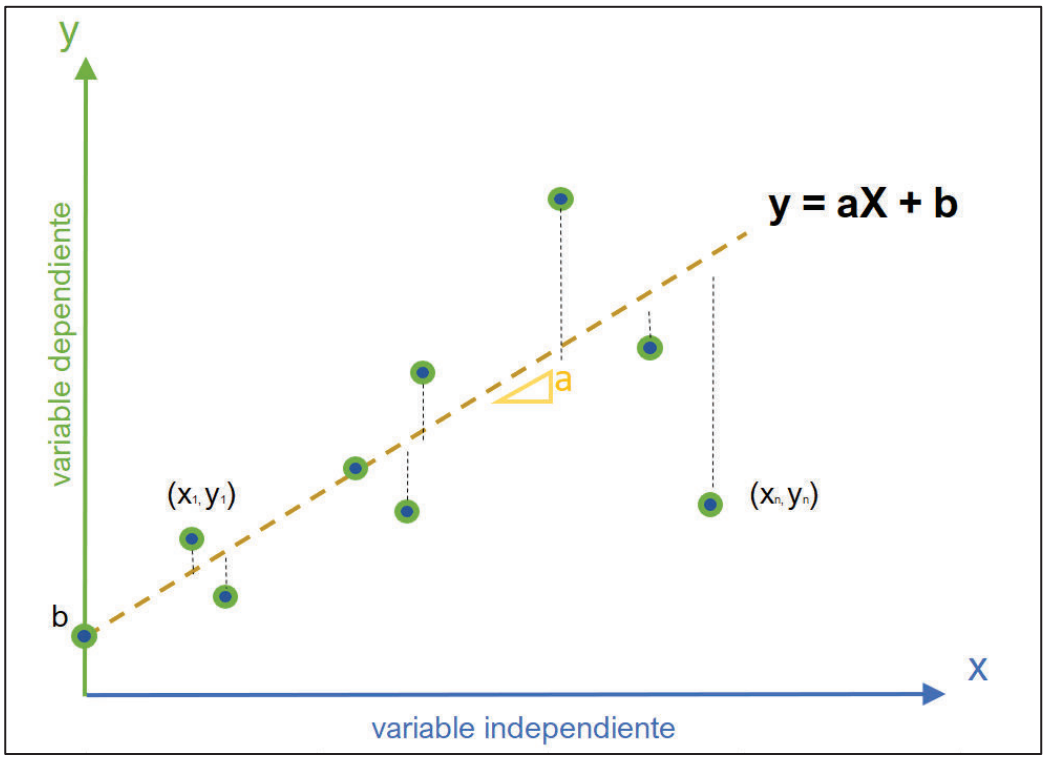

Figura 19. Elementos de línea de regresión

Estimación de parámetros

Los valores estimados de a, b, denominados en adelante como $\hat{a}, \hat{b}$, se obtienen a partir de los datos de entrenamiento, que son representados por un conjunto de datos observados. Una vez se conocen $\hat{a}, \hat{b}$ se puede predecir nuevos valores de Y basándose en valores de X. Dicho esto, la ecuación que describe al modelo se presenta de la siguiente forma (2):

$$\hat{y} = \hat{a}x + \hat{b} \tag{2}$$

El conjunto de datos observado se representa por los pares de datos que se muestran en la expresión:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$$

Estos pares de datos representan medidas observadas de X y Y y son utilizados para obtener los parámetros de ajuste lineal $\hat{a}, \hat{b}$. De modo que se requiere, encontrar una recta con una

pendiente y una intersección con el eje y que minimice la distancia a todos los puntos (x_i, y_i). El método más usado para medir esta proximidad es el ajuste por mínimos cuadrados, del cual se dan detalles a continuación.

Ajuste por mínimos cuadrados

El método por mínimos cuadrados consiste en calcular las estimaciones para los parámetros $\hat{a}, \hat{b}$, tales que hagan mínima la suma de los residuos al cuadrado, entendiendo como residuo a la diferencia entre el valor actual de la variable dependiente y_i y el valor predicho por el modelo $\hat{y_i}$; $(r_i = y_i - \hat{y_i})$ [7]. Lo que se describe con las siguientes ecuaciones ((3), (4)):

$$\frac{\partial \mathrm{RSS}}{\partial \hat{a}} = 0 \longrightarrow \hat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x_i})(y_i - \overline{y_i})}{\sum_{i=1}^{n}(x_i - \overline{x_i})^2} \tag{3}$$

$$\frac{\partial \mathrm{RSS}}{\partial \hat{b}} = 0 \longrightarrow \hat{b} = \bar{y} - \hat{a}\bar{x} \tag{4}$$

Donde:

- RSS, es la suma de los cuadrados de los residuos
- $\hat{a}, \hat{b}$ parámetros pendiente e intersección con eje y, respectivamente
- $\bar{x}, \bar{y}$, son los promedios de las muestras

Precisión en la estimación de los parámetros

Hasta ahora lo que se tiene es una simplificación de una relación que podría no ser exactamente lineal, ya que esta podría contener errores (como errores de medición, o la presencia de otra variable que se desconoce). Por esta razón la ecuación que describe a la relación lineal se modifica, así (5):

$$Y = aX + b + \epsilon \tag{5}$$

Donde:

- ϵ, es error aleatorio independiente de X.

Este error se relaciona con errores en la estimación de los parámetros $\hat{a}, \hat{b}$ calculados como se explico anteriormente. Para determinar el error de la estimación, se utiliza el **error estándar** que muestra la diferencia entre los valores reales, y_i y los valores estimados por la recta de regresión, $\hat{y}_i$ [18]. La fórmula del error estándar de regresión se presenta en la ecuación (6):

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \tag{6}$$

Donde:

n, es el número de observaciones

y_i, valor real de la variable dependiente o de respuesta

$\hat{y}_i$, estimación de la variable dependiente o de respuesta

La Figura 20 y la Figura 21 ejemplifican el cálculo del error estándar, dados los valores de y_i y $\hat{y}_i$.

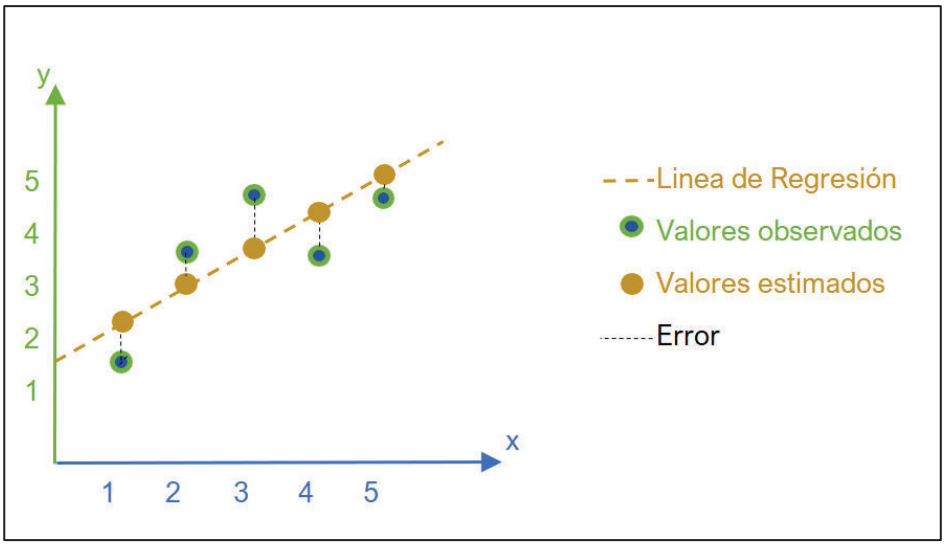

Figura 20. Valores observados y estimados en una línea de regresión

	Valor observado	Valor estimado		
x	y	$\hat{y}$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	2	2,8	-0,8	0,64
2	4	3,4	0,6	0,36
3	5	4	1	1
4	4	4,6	-0,6	0,36
5	5	5,2	-0,2	0,4
				2,4

Error estándar de la estimación $= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{2.4}{5-2}} = \sqrt{0.8} = 0.89$

Figura 21. Ejemplo del cálculo del error estándar de la estimación

Intervalos de confianza

A partir del error estándar pueden construirse intervalos de confianza que son límites entre los cuales se puede esperar que se encuentre el valor de cada parámetro de regresión $\hat{a}, \hat{b}$. Estos límites se obtienen con las siguientes fórmulas $((7.1), (7.2))$:

$$\hat{a} \pm 1.96 * S_e(\hat{a}) \quad (7.1)$$

$$\hat{b} \pm 1.96 * S_e(\hat{b}) \quad (7.2)$$

Donde:

- $\hat{a}, \hat{b}$, parámetros pendiente e intersección con eje y, respectivamente
- S_e , error estándar del parámetro

Los limites se obtienen sumando y restando 1.96 errores estándar al valor del parámetro de regresión. Se establece el valor 1.96 ya que se trabaja con un nivel de confianza del 95%. [19]

Prueba de significación

A partir del error estándar, también se puede encontrar el estadístico *t* que permite contrastar las hipótesis nulas que afirmen que los parámetros de regresión son cero. El estadístico *t*, se distribuye según el modelo de probabilidad *t Student* con *n-2* grados de libertad. Por tanto, pueden ser utilizados para decidir si un determinado parámetro de regresión es significativamente distinto de cero, o, en consecuencia, si la variable independiente está significativamente relacionada con la dependiente. Estos estadísticos *t*, se obtienen dividiendo los parámetros de regresión $\hat{a}, \hat{b}$, entre sus respectivos errores estándar, como se presenta en las ecuaciones $((8.1), (8.2))$

$$t = \frac{\hat{a}}{S_e(\hat{a})} \tag{8.1}$$

$$t = \frac{\hat{b}}{S_e(\hat{b})} \tag{8.2}$$

Implementación del modelo de Regresión Lineal

Como ilustración se presenta la implementación del modelo de Regresión Lineal con los datos del indicador de nombre: “Empleados de servicios, varones (% de empleo masculino)” e id: “SL.SRV.EMPL.MA.ZS”. Este indicador describe a las personas en edad de trabajar que realizan cualquier actividad para producir bienes o prestar servicios remunerados, ya sea en el trabajo durante el período de referencia o no en el trabajo debido a la ausencia temporal de un trabajo. El sector de servicios consiste en comercio mayorista y minorista; restaurantes y hoteles; transporte, almacenamiento y comunicaciones; financiamiento, seguros, bienes raíces y servicios comerciales; y servicios comunitarios, sociales y personales, de acuerdo con las divisiones 6-9 (ISIC2) o categorías G-Q (ISIC3) o categorías G-U (ISIC4). [6]. Estos datos se presentan en la Tabla 9, medidos por año y cuyo valor numérico representa el porcentaje de la población empleada masculina que se dedica a un trabajo relacionado con servicios.

Tabla 9. Datos del indicador con id “SL.SRV.EMPL.MA.ZS”

Año	Valor numérico (%)
2003	42,98
2004	43,79
2005	43,33
2006	43,38
2007	43,90
2008	43,61
2009	43,59
2010	44,71
2011	45,46
2012	45,76
2013	45,73
2014	46,38
2015	46,35
2016	46,83
2017	46,50

A continuación, se grafican los puntos para tener una idea de la distribución de los datos. Como se observa en la Figura 22, en este caso se tiene una aproximación lineal de los datos. Una vez graficados los datos, se procede a la obtención de parámetros de ajuste lineal, por el método de ajuste de mínimos cuadrados.

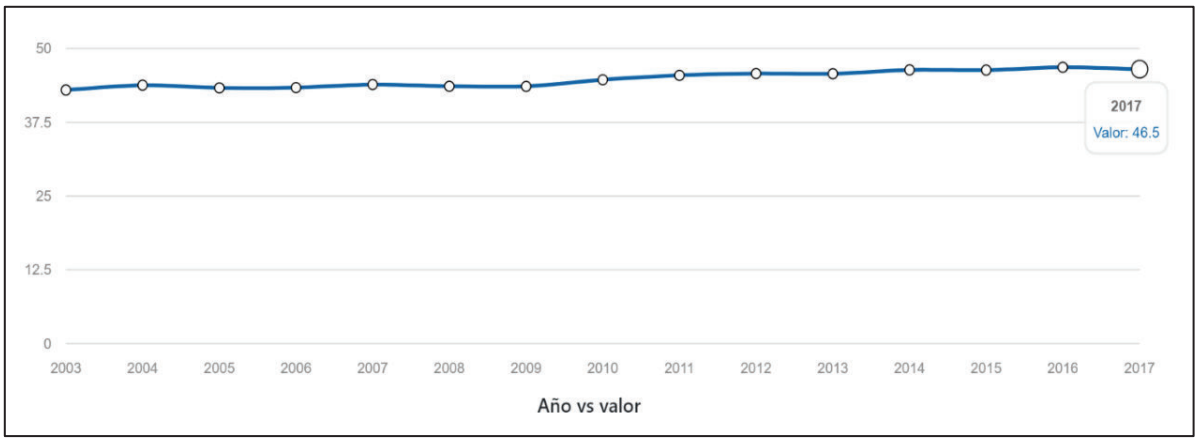

Figura 22. Gráfica de valores por año del indicador con id "SL.SRV.EMPL.MA.ZS", tomada del sistema web

Ajuste por mínimos cuadrados

Se establece la ecuación de la recta de regresión lineal, utilizando la ecuación (5):

$$Y = aX + b + \epsilon$$

Donde:

- Y, es el valor numérico del indicador medido por año
- X, es el año en el que se realiza la medición
- a, b, son las estimaciones para los parámetros de regresión pendiente e intersección con el eje y, respectivamente
- ϵ, error estándar de la estimación de los parámetros

Para la obtención de los parámetros de ajuste lineal se utilizan la ecuación (3) y la ecuación (4).

A continuación, se presenta la Tabla 10 con los datos necesarios para reemplazarlos en las ecuaciones, ecuación (3) y ecuación (4) respectivamente.

Siendo:

- $\overline{x}, \overline{y}$ el valor promedio de x, y respectivamente.
- **n**, corresponde al número de observaciones de y **,** por lo que se conoce son 15 observaciones (2003…2017).

Tabla 10. Valores calculados para la variable dependiente e independiente

n	x	y	xy	x^2	$\overline{x}$	$\overline{y}$
1	2003	42,98	86088,94	4012009	2010	44,82
2	2004	43,79	87755,16	4016016		
3	2005	43,33	86876,65	4020025		
4	2006	43,38	87020,28	4024036		
5	2007	43,90	88107,3	4028049		
6	2008	43,61	87568,88	4032064		
7	2009	43,59	87572,31	4036081		
8	2010	44,71	89867,1	4040100		
9	2011	45,46	91420,06	4044121		
10	2012	45,76	92069,12	4048144		
11	2013	45,73	92054,49	4052169		
12	2014	46,38	93409,32	4056196		
13	2015	46,35	93395,25	4060225		
14	2016	46,83	94409,28	4064256		
15	2017	46,50	93790,5	4068289		
	30150	**672,30**	**1351404,64**	**60601780**		

Los datos reemplazados en la ecuación (3) y ecuación (4) respectivamente se muestran así:

$$\hat{a}= \frac{\sum_{i=1}^{15} x_i y_i - n(\overline{x_i y_i})}{\sum_{i=1}^{15} x_i^2 - n\bar{x}^2} = \frac{1351404{,}64 - 15(2010 * 44{,}82)}{60601780 - 15 * 2010^2} = \frac{81{,}64}{280} = 0{,}29$$

$$\hat{b}= \bar{y} - \hat{a}\bar{x} = 44{,}82 - 0{,}29 * 2010 = -541{,}24$$

Por lo tanto, reemplazando los valores en la ecuación de la recta establecida, ecuación (5), se vería así:

$$Y = 0{,}29X - 541{,}24 + \epsilon$$

Con los datos de los parámetros de la ecuación (5) se tiene la Tabla 11 con las estimaciones de y, $\hat{y}$ (predicciones de y en base a x):

Tabla 11. Valores estimados de y, variable dependiente

x	y	xy	x^2	y^2	$\overline{x}$	$\overline{y}$	$\widehat{y}$	$y-\widehat{y}$	$(y-\widehat{y})^2$
2003	42,98	86088,94	4012009	1847,28			42,78	0,20	0,04
2004	43,79	87755,16	4016016	1917,56			43,07	0,72	0,52
2005	43,33	86876,65	4020025	1877,49			43,36	-0,03	0,00
2006	43,38	87020,28	4024036	1881,82			43,65	-0,27	0,07
2007	43,90	88107,3	4028049	1927,21			43,95	-0,05	0,00
2008	43,61	87568,88	4032064	1901,83			44,24	-0,63	0,39
2009	43,59	87572,31	4036081	1900,09			44,53	-0,94	0,88
2010	44,71	89867,1	4040100	1998,98	2010	44,82	44,82	-0,11	0,01
2011	45,46	91420,06	4044121	2066,61			45,11	0,35	0,12
2012	45,76	92069,12	4048144	2093,98			45,40	0,36	0,13
2013	45,73	92054,49	4052169	2091,23			45,69	0,04	0,00
2014	46,38	93409,32	4056196	2151,10			45,99	0,39	0,16
2015	46,35	93395,25	4060225	2148,32			46,28	0,07	0,01
2016	46,83	94409,28	4064256	2193,05			46,57	0,26	0,07
2017	46,50	93790,5	4068289	2162,25			46,86	-0,36	0,13
30150	**672,30**	**1351404,64**	**60601780**	**30158,82**			**672,3**	**0,00**	**2,53**

Error estándar

El error estándar de la estimación se obtiene con la ecuación (6), reemplazando con los datos obtenidos de la Tabla 11, se tiene:

$$S_e = \sqrt{\frac{2{,}53}{13}} = 0{,}44$$

Lo que se interpreta como que la desviación de los valores estimados $\hat{y}$, con respecto a la recta de regresión, es de 0.44.

Intervalos de confianza

A partir del error estándar de la estimación $\hat{y}$, se calcula el intervalo de confianza para la estimación reemplazando los valores de error estándar en las ecuaciones (7.1) y (7.2).

$$\hat{y} \pm 1{,}96 * S_e(\hat{y}) = \hat{y} \pm 1{,}96 * 0{,}44 = \hat{y} \pm 0{,}86$$

El intervalo se establece entonces como:

$$[\hat{y} + 0.86, \hat{y} - 0.86]$$

Implementación en Python

La implementación en el sistema web de Python del modelo de Regresión Lineal y sus elementos detallados anteriormente se realiza mediante la librería statsmodels, la cual proporciona clases y funciones para la estimación de muchos modelos estadísticos diferentes, así como para realizar pruebas estadísticas y explorar datos estadísticos [20]. Su implementación, se muestra en la Figura 23.

```
…
52  #========2. Entrenamiento
53  result = sm.OLS(y, sm.add_constant(x)).fit()
…
62  #========2. Predicción
63  #Predicción para el año 2017
64  ypred = result.predict(sm.add_constant(x)[0])
…
```

Figura 23. Extracto del Anexo 7. Implementación de Regresión Lineal Simple

El modelo ***result***, se crea mediante la función **OLS** (Ordinary Least Squares) de la librería *statsmodels*, que realiza el ajuste de la recta de regresión por el método de mínimos cuadrados. El modelo se entrena, tomando como conjunto de entrenamiento, a los valores de *y*, variable dependiente que contiene a los valores correspondientes a los datos numéricos del indicador, y los valores de la variable independiente *x*, que corresponde al año en el que fue realizada la medición. Una vez el modelo está entrenado, mediante la instrucción ***predict,*** se predice el valor de la variable dependiente o de respuesta *y* para el año 2017.

Los resultados de la ejecución del código de la Figura 23, se muestran en la Figura 24, donde se resume las posibilidades de la librería statsmodels, y cuyos resultados obtenidos en la regresión lineal, se encierran en el cuadro azul, siguiendo un orden de arriba hacia abajo, el número de observaciones presentes en la regresión lineal, $\hat{a}$ y $\hat{b}$ estimaciones para los parámetros de regresión pendiente e intersección con el eje y, respectivamente. Además, el estadístico t *Student*, el error estándar de la estimación y finalmente, la ecuación de la recta de regresión.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.904
Model:                            OLS   Adj. R-squared:                  0.897
Method:                 Least Squares   F-statistic:                     122.3
Date:                Sun, 18 Nov 2018   Prob (F-statistic):           5.53e-08
Time:                        00:40:26   Log-Likelihood:                -7.9357
No. Observations:                  15   AIC:                             19.87
Df Residuals:                      13   BIC:                             21.29
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -541.2386     52.993    -10.213      0.000    -655.722    -426.755
x1             0.2916      0.026     11.059      0.000       0.235       0.349
==============================================================================
Omnibus:                        1.776   Durbin-Watson:                   1.199
Prob(Omnibus):                  0.412   Jarque-Bera (JB):                0.818
Skew:                          -0.572   Prob(JB):                        0.664
Kurtosis:                       3.004   Cond. No.                     9.35e+05
==============================================================================

Error estándar de la regresión: 0.441161802816
Coeficiente de Determinación, R^2 =  0.903922360013

        -----Ecuación de la recta de regresión ---------

       y =  0.291571428571 x + -541.238571429
```

Figura 24. Resultado de la ejecución de la implementación en Python de Regresión Lineal

2.1.5.2 Cadena de Márkov

La cadena de *Márkov*, también conocida como modelo de *Márkov* o proceso de *Márkov*, es un concepto desarrollado dentro de la teoría de la probabilidad y la estadística que establece una fuerte dependencia entre que tenga lugar un evento y un evento anterior. Su principal utilidad es el análisis del comportamiento de procesos estocásticos (procesos aleatorios).

Según *Márkov*, en sistemas o procesos estocásticos que presentan un estado presente o actual es posible conocer sus antecedentes o desarrollo histórico y, por lo tanto, establecer una descripción de la probabilidad futura de los mismos.

Más formalmente, la definición supone que en procesos estocásticos la probabilidad de que algo suceda solamente depende del pasado histórico de la realidad que es estudiada. Por este motivo, a menudo se dice que estas cadenas cuentan con memoria. La base de las cadenas es la conocida como propiedad de *Márkov*, la cual resume lo dicho anteriormente en la siguiente regla: lo que la cadena experimente en un momento $t + 1$ solamente depende de lo acontecido en el momento t (el inmediatamente anterior) [8].

2.1.5.2.1 Elementos del modelo

Una cadena de *Márkov* implica un sistema que se encuentra en un estado determinado en cada paso, con el estado cambiando aleatoriamente entre pasos. Es una secuencia de variables aleatorias $x1, x2, x3, \ldots$ con la propiedad de *Márkov* explicada previamente. La fórmula matemática probabilista que la describe se presenta a continuación, en la ecuación (9):

$$\Pr(X_{n+1} = x \mid X_1 = x\,1, X_2 = x\,2, \ldots, X_n = x\,n) = \Pr(X_{n+1} = x \mid X_n = x\,n) \qquad (9)$$

Como se puede ver, la probabilidad de X_{n+1}, solo depende de la probabilidad de X_n que la precede.

Los valores posibles de Xi forman un conjunto denominado **espacio de estados**. El espacio de estado puede ser cualquier cosa: letras, números, puntajes de baloncesto o condiciones climáticas.

Una cadena de *Márkov* se representa utilizando un autómata probabilístico. Los cambios de estado del sistema se llaman transiciones. Las probabilidades asociadas con varios cambios de estado se llaman **probabilidades de transición**. Un autómata probabilístico incluye la probabilidad de una transición dada en la función de transición, convirtiéndola en una **matriz de transición**. En ella, cada estado en el espacio de estado se incluye una vez como una fila y otra vez como una columna, y cada celda en la matriz le dice la probabilidad de pasar del estado de su fila al estado de su columna.

Si la cadena de *Márkov* tiene N estados posibles, la matriz será una matriz $N\ x\ N$, de modo que la entrada (i, j) es la probabilidad de transición del estado i al estado j. Además, la matriz de transición debe ser una matriz estocástica, una matriz cuyas entradas en cada fila deben sumar exactamente 1, ya que cada fila representa la distribución de probabilidad de cada estado [21].

Resumiendo, el modelo se caracteriza por un espacio de estados, una matriz de transición que describe las probabilidades de transiciones entre cada estado, y un estado inicial a través del espacio de estados, dado en la distribución inicial.

La Figura 25 muestra un ejemplo de una cadena de *Márkov*, la descripción de un proceso aleatorio del clima durante varios días. Donde las características del clima, soleado y lluvioso, se establecen como **el espacio de estados** y cuyas **probabilidades de transacción** se describen a continuación:

- Si hoy es un día soleado, entonces existe una probabilidad de 0.2 de que mañana sea un día lluvioso y existe una probabilidad de 0.8 de que vuelva a ser soleado
- Si hoy es un día lluvioso, entonces existe una probabilidad de 0.7 de que mañana sea un día soleado y existe una probabilidad de 0.3 de que vuelva a ser lluvioso.

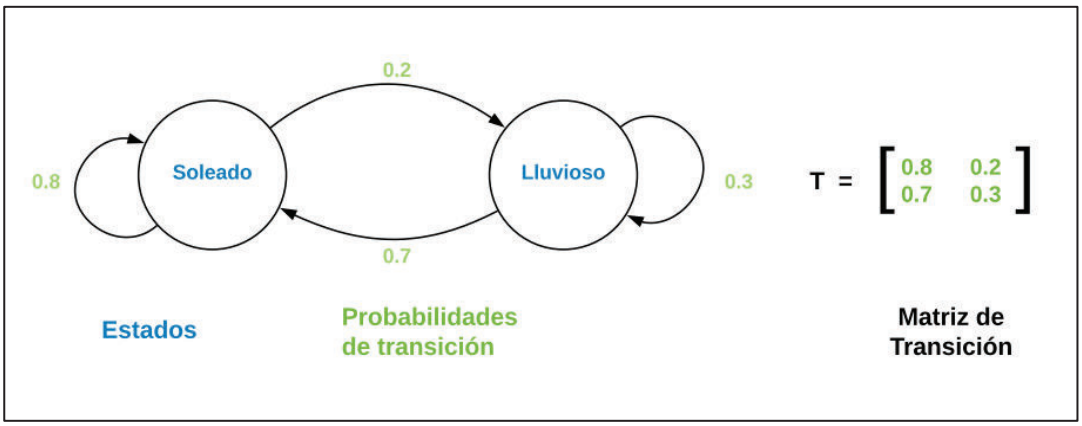

Figura 25. Ejemplo de Cadena de Márkov

2.1.5.2.2 Predicciones con Cadenas de Márkov

Alcanzar un estado estable del sistema, se basa en la suposición de que, en el futuro, el sistema alcance un "equilibrio", en el sentido de que lo que se experimente en el momento t, sea igual a lo experimentado en el momento $t - 1; t = t - 1$. No queriendo decir con esto que las transiciones entre estados no tendrán lugar, al contrario, sucederán, pero estas tenderán al "equilibrio global" tal que la probabilidad de permanecer en un determinado estado no cambie.

Dicho esto, puede observarse que conocer la probabilidad de que un estado ocurra a largo plazo, al llegar a un estado sustentable, ayuda a la predicción y estimación en largos periodos de tiempo [22].

2.1.5.2.3 Implementación de una Cadena de Márkov

Para ilustrar la implementación de una Cadena de *Márkov*, se presentará el caso con los datos del mismo indicador utilizado en la implementación de Regresión Lineal y cuyo es nombre: "Empleados de servicios, varones (% de empleo masculino)". Sus datos se presentan en la Tabla 12, medidos por año y cuyo valor numérico representa el porcentaje de la población empleada masculina que se dedica a un trabajo relacionado con servicios.

Tabla 12. Datos del indicador con id “SL.SRV.EMPL.MA.ZS”

Año	Valor numérico (%)
2017	46,5
2016	46,83
2015	46,35
2014	46,38
2013	45,73
2012	45,76
2011	45,46
2010	44,71
2009	43,59
2008	43,61
2007	43,9
2006	43,38
2005	43,33
2004	43,79
2003	42,98

Con estos datos se describe cada elemento del modelo, como se mencionó en el apartado de **Elementos del modelo***:*

a) Un espacio de estados
b) Una matriz de transición
c) Un estado inicial (distribución inicial)

Una vez construido el modelo con sus respectivos elementos se determina las probabilidades de estado del estado estable del sistema para realizar la predicción del valor numérico para el año 2017. A continuación, se detalla cada elemento del modelo.

a) Espacio de Estados

Para representar a los datos en un espacio finito de estados se calculan los cuartiles del conjunto de datos, que son los tres valores de una variable. En este caso, el valor numérico del indicador que dividen a un conjunto de datos ordenados en cuatro partes iguales.

$Q1, Q2\ y\ Q3$, determinan los valores correspondientes al 25%, al 50% y al 75% de los datos. Y donde además $Q2$ coincide con la mediana [22]. Lo descrito se representa con la ecuación (10).

$$Q_k = k\left(\frac{N}{4}\right) \tag{10}$$

Donde:

- $Q_k =$ Cuartil 1,2,3 o 4
- $N =$ total de datos de la distribución

El cálculo del primer cuartil se ejemplifica con la fórmula presentada en la ecuación (10). Los otros cuartiles son calculados bajo la misma fórmula y sus valores son presentados en la Tabla 14.

Primer cuartil

El cálculo de los cuartiles considera a un conjunto de datos ordenado de menor a mayor, por lo que los valores numéricos del indicador son ordenados para este fin, quedando ordenados como se ve en la Tabla 13.

Tabla 13. Valores numéricos ordenados de menor a mayor del indicador "Empleados de servicios, varones (% de empleo masculino)"

Año	Valor numérico (%)
2003	42,98
2005	43,33
2006	43,38
2009	43,59
2008	43,61
2004	43,79
2007	43,9
2010	44,71
2011	45,46
2013	45,73
2012	45,76
2015	46,35
2014	46,38
2017	46,5
2016	46,83

Una vez son ordenados los datos, se aplica la ecuación (10), donde N corresponde a 14, que son los valores incluidos en el período 2003-2016 y k, que corresponde a 1 por ser el primer cuartil, quedando la ecuación así:

$$Q_k = k\,(\frac{N}{4})$$

$$Q_1 = \left(\frac{14}{4}\right) = 3.5$$

A Q_1 le corresponde la posición 3 en la lista de valores ordenados, que es igual a 43.38. Al existir decimales en la posición se interpola entre la posición 3 y 4, de esta forma:

$$Q_1 = 43.38 + 0.5(43.59 - 43.38)$$

$$Q_1 = 43.38 + 0.5(0.21)$$

$$Q_1 = 43.38 + 0.105$$

$$\boldsymbol{Q_1 = 43.48}$$

La diferencia entre la posición 4 y 3, se multiplica por el valor de los decimales, y esto se suma al valor de la posición encontrada previamente, calculando un valor de 43.48 para el primer cuartil.

A continuación, los valores de los cuartiles se representan en el modelo al espacio de estados. En la Tabla 14 se muestran los valores obtenidos mediante la ecuación (10) para cada cuartil.

Tabla 14. Espacio de estados del modelo de Cadenas de Márkov

Estado 1	E1	42,98 - 43,48
Estado 2	E2	43,48 - 43,90
Estado 3	E3	43,90 - 45,74
Estado 4	E4	45,74 - 46,83

El primer y último valor de la lista de valores numéricos ordenados delimitan el espacio de estados en su inicio y fin (inicio: 42.98, fin:46.83).

b) Matriz de transición

La lista de valores numéricos desordenada del indicador es clasificada de acuerdo con los rangos establecidos de cada estado mostrado en la Tabla 14. Esta clasificación se realiza para obtener una matriz de transición entre estados, cuya construcción se detalla a continuación.

La clasificación se realiza para todos los valores incluidos en el período 2003-2016. El primer valor de la lista es 46.83 que corresponde al año 2016, pertenece al estado 4 que está limitado por los valores 45.74 y 46.83, ya que es mayor al valor 45.74 y es menor o igual al valor 46.83. Los demás valores de la lista son clasificados siguiendo la misma lógica, con lo que se obtiene el siguiente vector de estados: $[E4, E4, E4, E3, E4, E3, E3, E2, E2, E3, E1, E1, E2, E1]$.

Dado un vector de estados se puede construir la matriz de transición. Se posicionan como filas y columnas a los 4 estados determinados por los cuartiles calculados previamente, obteniendo de esta forma una matriz de tamaño 4x4.

	$E1$	$E2$	$E3$	$E4$
$E1$	0	0	0	0
$E2$	0	0	0	0
$E3$	0	0	0	0
$E4$	0	0	0	0

Esta matriz representa las transiciones que se tuvieron entre cada estado, partiendo del vector de estados: $[E4, E4, E4, E3, E4, E3, E3, E2, E2, E3, E1, E1, E2, E1]$. Se tienen :

- El primer elemento del vector de estados E4 va al segundo elemento del vector de estados E4, por lo que la probabilidad de ocurrencia de que, estando inicialmente en el estado E4, se dirija al estado E4, es igual a 1. Formándose la matriz de la siguiente forma:

	$\boldsymbol{E1}$	$\boldsymbol{E2}$	$\boldsymbol{E3}$	$\boldsymbol{E4}$
$\boldsymbol{E1}$	0	0	0	0
$\boldsymbol{E2}$	0	0	0	0
$\boldsymbol{E3}$	0	0	0	0
$\boldsymbol{E4}$	0	0	0	1

- El segundo elemento del vector de estados, E4, va al tercer elemento del vector de estados, E4, por lo que la probabilidad de ocurrencia de que, estando inicialmente en el estado E4, se dirija al estado E4, vuelve a ser igual a 1. Esta probabilidad sumada a la anterior probabilidad ya que los estados involucrados son los mismos, aumenta a 2.

	$E1$	$E2$	$E3$	$E4$
$E1$	0	0	0	0
$E2$	0	0	0	0
$E3$	0	0	0	0
$E4$	0	0	0	2

- El tercer elemento del vector de estados, E4, va al cuarto elemento del vector de estados, E3, por lo que la probabilidad de ocurrencia de que, estando inicialmente en el estado E4, se dirija al estado E3, es igual a 1.

	$E1$	$E2$	$E3$	$E4$
$E1$	0	0	0	0
$E2$	0	0	0	0
$E3$	0	0	0	0
$E4$	0	0	1	2

Así, la matriz de transición es completada con las transiciones de cada estado presente en el vector de estados. Siguiendo el proceso descrito anteriormente para cada uno de ellos se obtiene la siguiente matriz:

	$E1$	$E2$	$E3$	$E4$
$E1$	1	1	0	0
$E2$	1	1	1	0
$E3$	1	1	1	1
$E4$	0	0	2	2

La matriz de transición debe ser una matriz estocástica como se mencionó en el apartado de **Elementos del modelo**, por lo que se calcula la transpuesta (convertir columnas en filas) de la matriz normalizada de transición, para cumplirse que cada fila sume 1, representando cada fila la distribución de probabilidad de cada uno de sus estados. La matriz de probabilidades de transición obtenida se muestra a continuación.

	$E1$	$E2$	$E3$	$E4$
$E1$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
$E2$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0
$E3$	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
$E4$	0	0	$\frac{1}{3}$	$\frac{2}{3}$

c) Estado inicial a través del espacio de estados

La matriz de probabilidades de transición obtenida previamente representa la distribución inicial que tiene los valores numéricos del indicador en el espacio de estados construido a partir de sus cuartiles.

Predicción de valor numérico para el año 2017

La resolución de la matriz de probabilidades de transición permite encontrar las probabilidades de estado estable de cada uno de los estados del espacio de estados del sistema. El estado cuya probabilidad de estado estable proporcione un valor superior al resto, será el estado en donde se encuentre el valor de la predicción para el año 2017.

Para resolver la matriz se plantea un sistema de ecuaciones de la siguiente forma:

$$(w\ \ x\ \ y\ \ z)\begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{pmatrix} = (w\ \ x\ \ y\ \ z) \tag{11}$$

Donde $(w\ x\ y\ z)$ representan la probabilidad de estado estable de cada estado del espacio de estados. De la resolución de las ecuaciones de la matriz, se obtiene:

$$2w - x = 0; \tag{12}$$

$$8x - 4w - 3y = 0; \tag{13}$$

$$9y - 4w - 4x - 4z = 0; \tag{14}$$

$$2z - 3y = 0; \tag{15}$$

$$w + x + y + z = 1; \tag{16}$$

La ecuación (16) del sistema de ecuaciones, se deduce de saber que cada fila suma 1, al ser una matriz estocástica, y que cada fila representa la distribución de probabilidad de cada estado.

Así, los valores obtenidos para $(w\ x\ y\ z\ )$ son:

$$w = \frac{1}{13};$$
$$x = \frac{2}{13};$$
$$y = \frac{4}{13};$$
$$z = \frac{6}{13};$$

A continuación, los valores obtenidos son presentados con valores decimales en un vector de estados, un vector de probabilidades de estado estable:

$$[0.07692308,\quad 0.15384615,\quad 0.30769231,\quad 0.46153846\ ]$$

A partir del vector de probabilidades de estado estable, se establece que el estado con mayor probabilidad de estado estable es el estado 4, con un valor de 0.46, y cuyos valores están delimitados por 45,74 y 46,83. Para tomar el valor más representativo del estado se calcula su promedio, de esta forma: $(45.74 + 46.83)/2 = 46.28$.

Finalmente, el valor obtenido como promedio, se toma como el valor predicción para el año 2017.

Implementación en Python

La implementación en el sistema web de Python de lo anteriormente detallado se realiza mediante la librería *discreteMarkovChain* que aborda el problema de obtener la distribución de estado estable de una Cadena de *Márkov*, también conocida como la distribución estacionaria, la distribución límite, está basado en *numpy* y *scipy* para cálculos eficientes y uso limitado de recursos [23]. La implementación, se muestra en la Figura 26.

```
…
122 #Z = Matriz de probabilidades de transición de estados
123 P = np.array(Z)
124 #Creación del modelo
125 mc = markovChain(P)
126 mc.computePi("linear")
127 #Cálculo del vector predicción
128 vectorPrediction = mc.pi
…
```

Figura 26. Extracto de Anexo 8. Implementación de Cadenas de Márkov

El modelo ***mc*** es creado a partir de la función ***markovChain***, que recibe como parámetro a la matriz de probabilidades de transición de estados **P**. A través de la función ***computePi***, se plantea el sistema de ecuaciones para hallar los valores de probabilidad de estado estable de cada estado del modelo y el cálculo de dichos valores finalmente se realiza con la función ***pi.***

Una vez obtenidas las probabilidades de estado estable en el vector predicción, se obtiene el promedio del estado con mayor probabilidad de estado estable. De esa forma, este valor promedio se considera el valor de predicción para el año 2017.

El resultado de la ejecución del Anexo 8. Implementación de Cadenas de Márkov, es el conjunto de todos los pasos que se describieron a detalle anteriormente. La Figura 27, muestra el resultado de la ejecución.

```
C:\Users\natyd\Desktop>py predictionMarkov.py
        -------Cadena de Márkov---------

Número de datos: 14

Valores desordenados:
 [46.83, 46.35, 46.38, 45.73, 45.76, 45.46, 44.71, 43.59, 43.61, 43.9, 43.38, 43.33, 43.79, 42.98]

Valores ordenados:
 [42.98, 43.33, 43.38, 43.59, 43.61, 43.79, 43.9, 44.71, 45.46, 45.73, 45.76, 46.35, 46.38, 46.83]

Estado 1       ||Estado 2       ||Estado 3       ||Estado 4
42.98 - 43.48  || 43.48 - 43.9 || 43.9 - 45.74 || 45.74 - 46.83

Vector de Estados:
[4, 4, 4, 3, 4, 3, 3, 2, 2, 3, 1, 1, 2, 1]

Matriz de transición sin normalizar:
1 1 0 0
1 1 1 0
1 1 1 1
0 0 2 2


Transpuesta de Matriz Normalizada:
[[ 0.33333333  0.33333333  0.33333333  0.        ]
 [ 0.33333333  0.33333333  0.33333333  0.        ]
 [ 0.          0.25        0.25        0.5       ]
 [ 0.          0.          0.33333333  0.66666667]]

Vector Predicción:
[ 0.07692308  0.15384615  0.30769231  0.46153846]
```

Figura 27. Resultado de la ejecución de la implementación en Python de Cadenas de Márkov

2.1.6 Comunicación

En esta etapa, se interpretan, visualizan, informan o comunican los resultados obtenidos del análisis. Esto se puede realizar a través de gráficas, reportes, resúmenes entre otros [5]. Para esta etapa, un sistema web basado en el framework Django, un framework web Python de alto nivel que fomenta un desarrollo rápido y un diseño limpio y pragmático. Es gratis y de código abierto [23], fue desarrollado siguiendo la metodología de desarrollo de software de prototipado, explicada más adelante en la sección **2.2. Desarrollo de software**. Los scripts que conciernen tanto a la construcción y desarrollo del sistema web, como al análisis mismo de los datos serán entregables incluidos en el proyecto.

El sistema web contempla la representación de los datos de indicadores de forma tabular y en forma gráfica. Los detalles de cada representación se presentan a continuación.

Tabla

Una tabla es la exposición de una serie de datos interrelacionados entre sí. En nuestro caso, se representan los valores numéricos de cada indicador medidos en una unidad tiempo anual. En la Figura 28, se muestra un ejemplo de tabla con valores para los años 2014-2017 para un determinado indicador.

Value	Date
66.41	2017
65.57	2016
66.25	2015
68.51	2014

Figura 28. Extracto de tabla con valores numéricos de indicador

Los valores predichos de cada indicador que describe a la temática Protección Social y Empleo por cada modelo se presentan también en una tabla, como se muestra en la Figura 29.

Predicción 2017

Valor Real	Valor predicho por Cadenas de Márkov	Valor predicho por Regresión Lineal	Coeficiente de Determinación Regresión Lineal
66.41	67.28	67.5	0.15

Figura 29. Tabla con valores predichos por cada modelo

Además, se presenta en el sistema una tabla de resumen, que contiene tanto los valores reales del indicador para el año 2017, como los valores predichos por cada modelo. Cadenas de *Márkov* y Regresión Lineal. Como totales se tiene un recuento de cuantas veces cada modelo fue más exacto para cada indicador. La Figura 30, es un extracto de la tabla de resumen, la tabla original contiene toda la lista de indicadores.

Resumen de Resultados

Notación de los modelos

R : Regresión Lineal
M : Cadenas de Márkov

#	Nombre	Valor Real	Valor Márkov	Valor Regresión	Error Residual Márkov (%)	Error Residual Regresión (%)	Modelo
1	Empleados de servicios, varones (% de empleo masculino)	46.5	46.28	46.86	0.47	0.77	M
2	Empleados de servicios, mujeres (% del empleo femenino)	66.41	67.28	67.5	1.31	1.64	M
3	Empleo informal, masculino (% del total del empleo no agrícola)	49.5	47.16	46.6	4.73	5.86	M
4	Empleo informal, femenino (% del total del empleo no agrícola)	56.12	51.55	52.42	8.14	6.59	R
60	Desempleo, mujeres jóvenes (% de la población activa femenina entre 15 y 24 años de edad) (estimación nacional)	11.44	12.12	9.24	5.94	19.23	M
Total Cadenas de Márkov							**31**
Total Regresión Lineal							**29**

Figura 30. Extracto de tabla de resumen de valores predichos

Gráfica Lineal

Los gráficos de líneas se utilizan para mostrar el valor cuantitativo en un intervalo o intervalo de tiempo continuo. Se usa con mayor frecuencia para mostrar tendencias y relaciones (cuando se agrupan con otras líneas) [24]. Los valores numéricos de indicador en el periodo de tiempo 2003-2017, se reflejan en la gráfica lineal de la Figura 31. En el ejemplo la tendencia que se observa es decreciente.

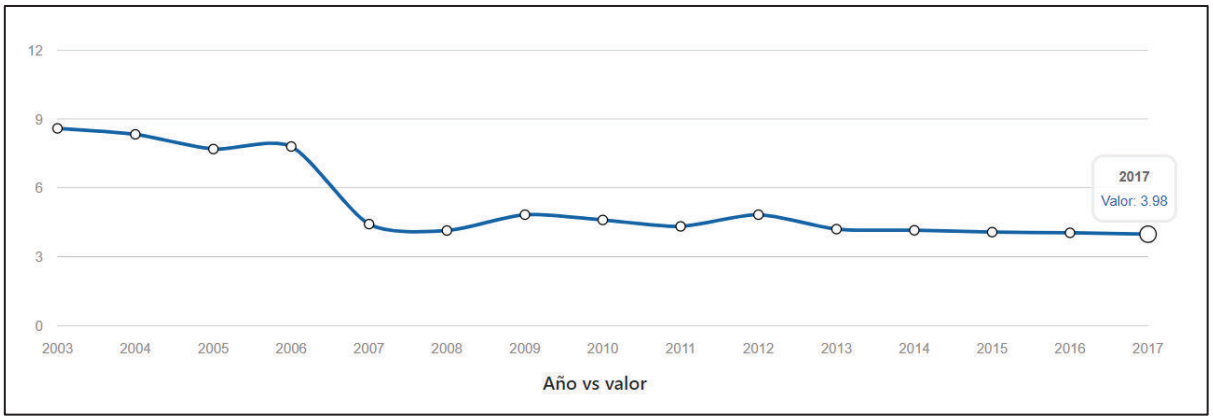

Figura 31. Gráfico de líneas con valores numéricos de indicador

2.2 Desarrollo de software

Para el desarrollo del software se siguió una aproximación basada en prototipado. Esta técnica consiste en modelar el producto final de software para efectuar pruebas sobre determinados atributos de este sin necesidad de que se encuentre completamente disponible [25].

Para probar la funcionalidad del software se utiliza el prototipado vertical, mientras que para probar la fidelidad de la reproducción de la interfaz se utiliza el prototipado de alta fidelidad, en el que se busca que las características de la interfaz sean idénticas a las del producto final, en este caso al sistema web basado en el framework Django.

2.2.1 Prototipo de interfaz

Los prototipos de interfaz fueron construidos con Balsamiq, herramienta para crear maquetas de interfaces gráficas para el usuario [26]. La herramienta cuenta con versiones de escritorio y web. Se utilizó la versión web y se crearon los prototipos de interfaz para:

- Listar temáticas
- Listar Indicadores
- Valores numéricos de cada indicador
- Resumen de los resultados de las predicciones

Las ilustraciones mostradas a continuación no son del todo legibles, porque representan el prototipo del modelo final de la aplicación. Sin embargo, se trata de que sean los más legibles e ilustrativos posibles.

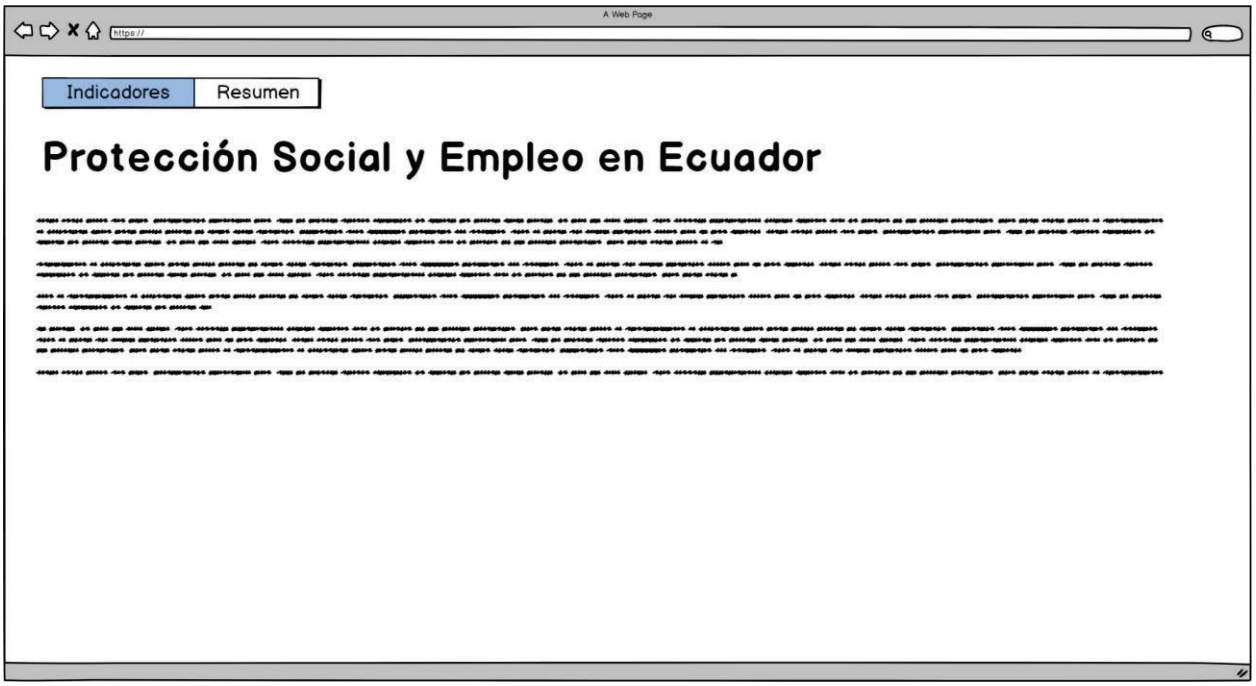

Figura 32. Prototipo de lista de temáticas

La interfaz de la Figura 32 muestra la temática de Protección Social y Empleo con su descripción. Su título navega a la siguiente interfaz (Figura 33) que contiene sus indicadores.

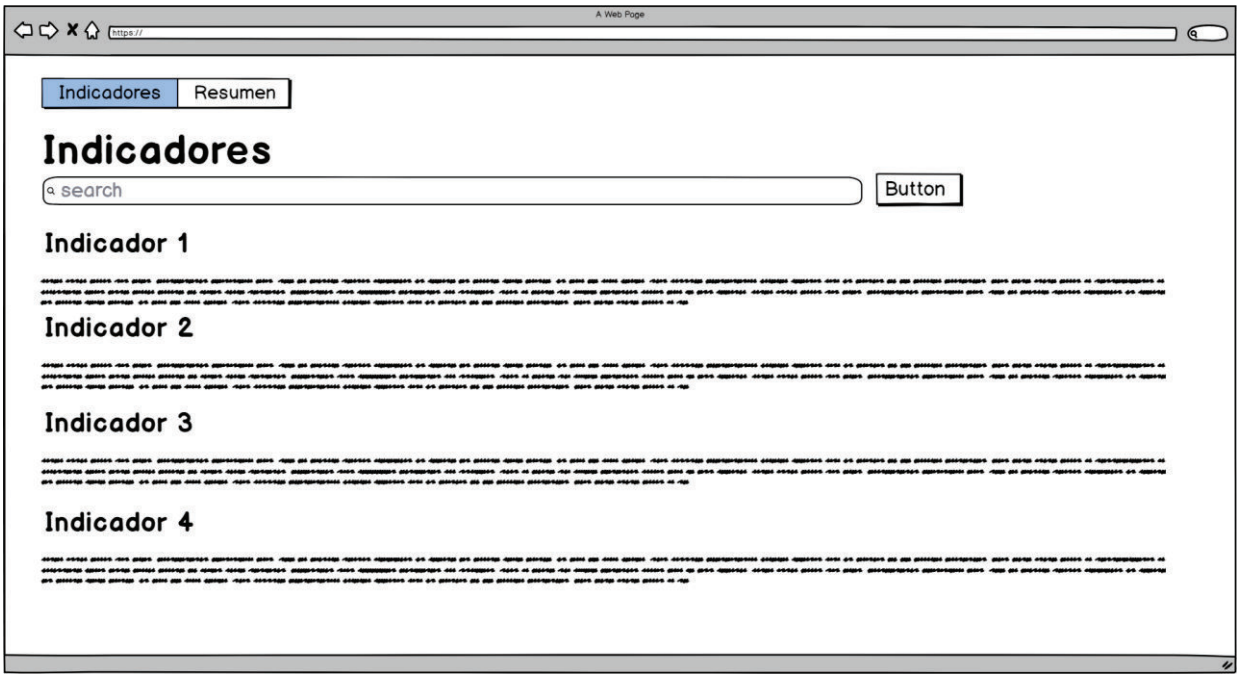

Figura 33. Prototipo de lista de Indicadores de temática

La Figura 33, muestra el conjunto de indicadores para la temática de Protección Social y Empleo con sus respectivas descripciones. Se permite realizar búsquedas de indicadores por frases contenidas en el nombre del mismo. El titulo de cada de uno de los indicadores, navega hacia la siguiente interfaz (Figura 34), que contiene los valores numéricos del indicador.

La Figura 34 muestra la información del indicador en cuanto a su descripción y a sus valores numéricos, así como también una gráfica representando a los valores y el resultado de las predicciones realizadas por los modelos de Cadenas de *Márkov* y Regresión Lineal.

La interfaz de resumen de la Figura 35 muestra una tabla con el conjunto de indicadores de Protección Social y Empleo con sus valores reales y los obtenidos de las predicciones.

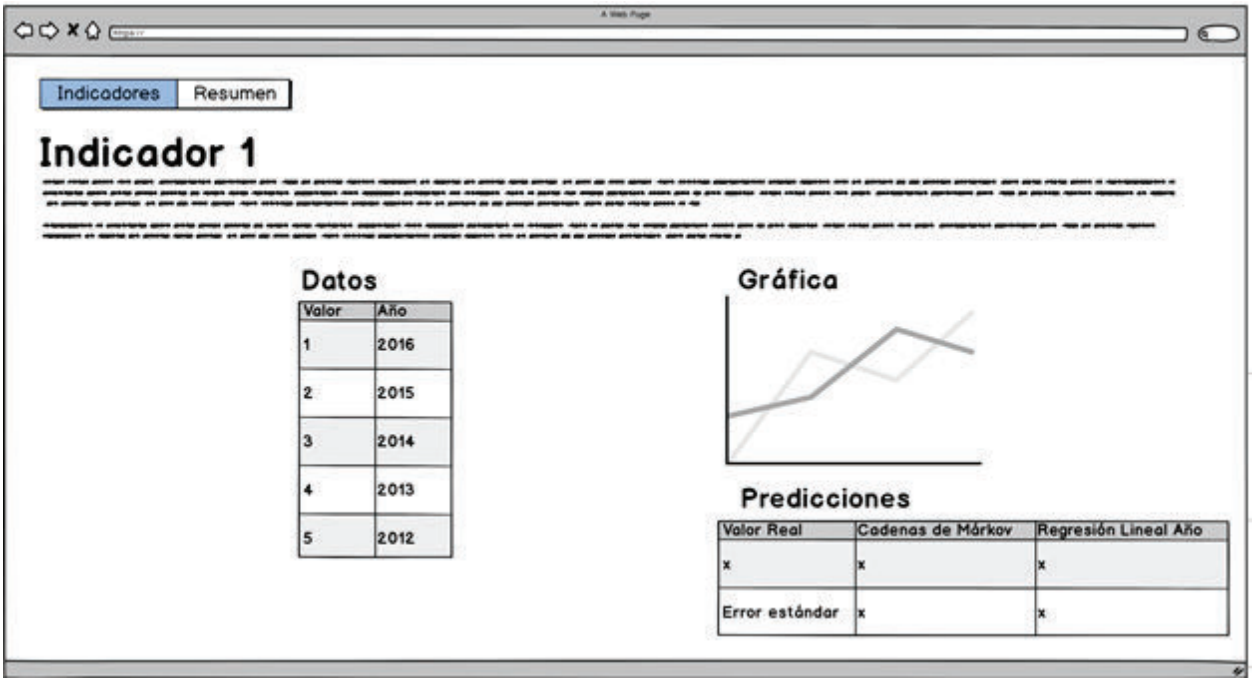

Figura 34. Prototipo de valores numéricos del indicador

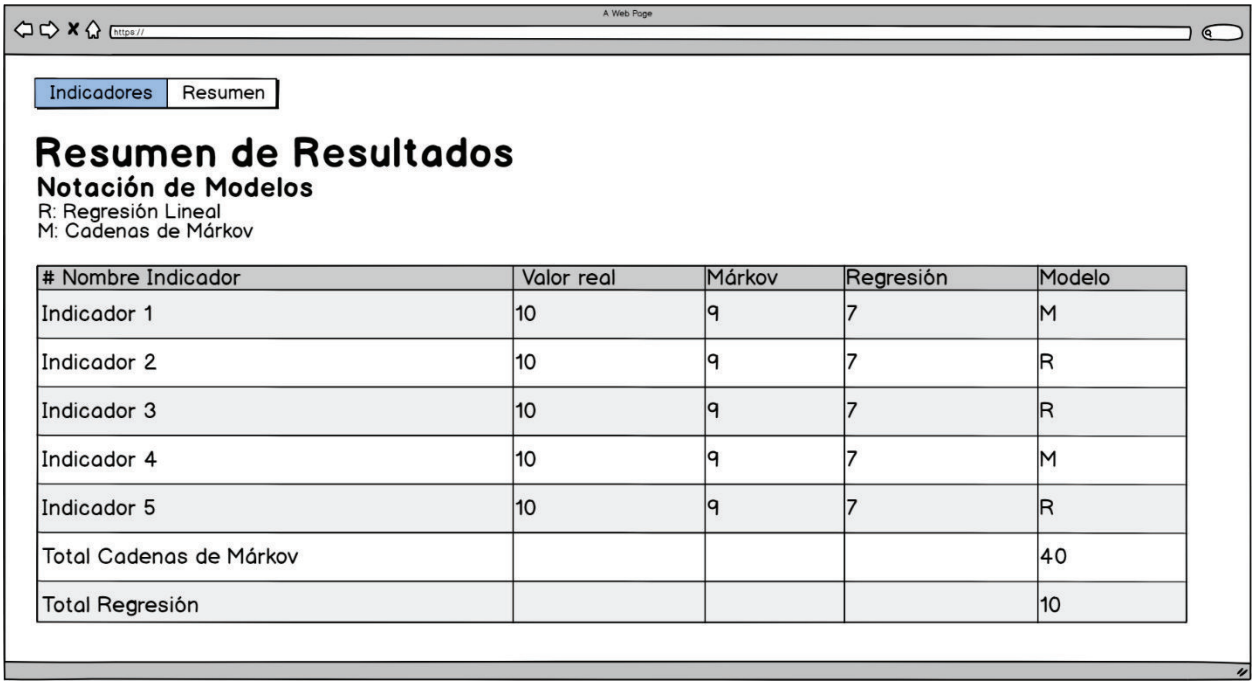

Figura 35. Prototipo de resumen de resultados de predicciones

2.2.2 Prototipo vertical

Un prototipado vertical implementa pocas características del sistema, pero estas características son totalmente funcionales. Es decir, que se puede probar una parte limitada del sistema, pero su funcionalidad puede ser probada a profundidad [25]. En este sentido, los prototipos de interfaz fueron de mucha utilidad, porque en base a ellos y siguiendo el orden en que se presentaron fueron desarrollándose las características descritas en cada prototipo de interfaz. El prototipo de lista de indicadores de la temática de Protección Social y Empleo fue seleccionado para realizar la prueba de las funcionalidades.

Las funcionalidades por probarse son:

- Listar todos los indicadores de la temática.
- Buscar indicadores por su nombre.
- Navegabilidad de indicador hacia sus valores numéricos.

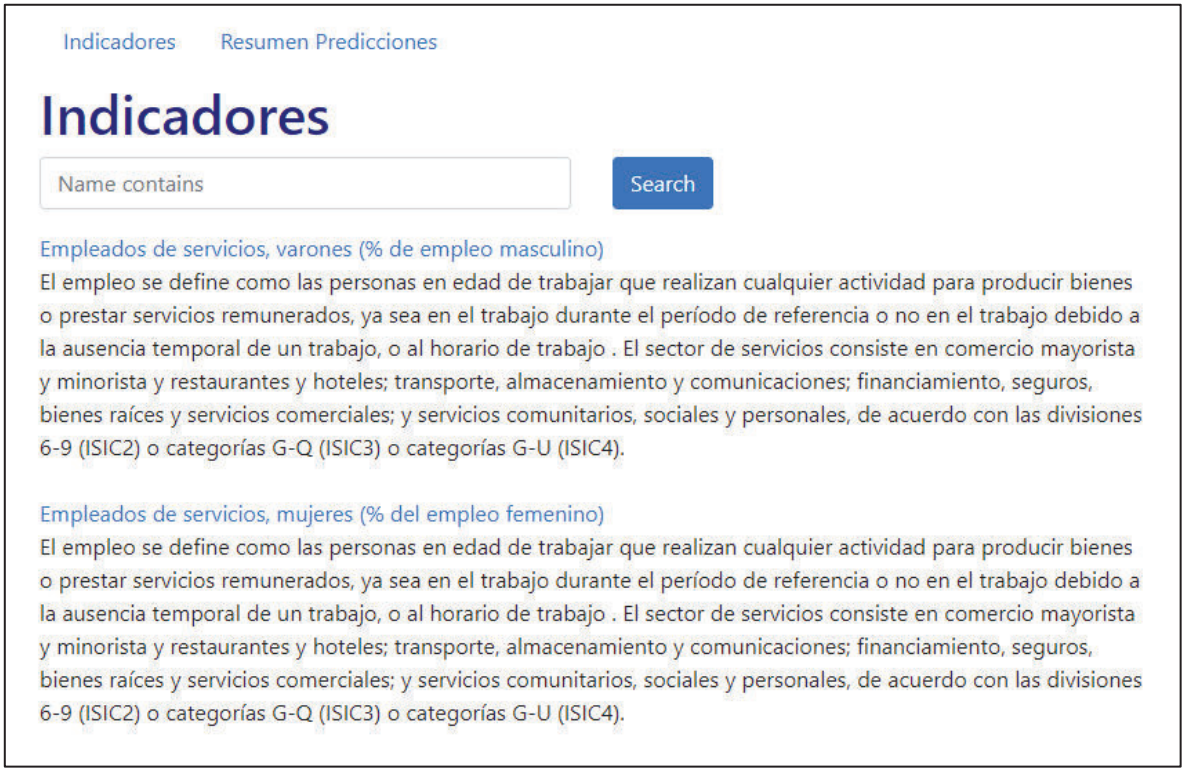

Figura 36. Extracto de la lista de indicadores de temática Protección Social y Empleo

En la Figura 36, se evidencia un extracto de la lista de todos los indicadores de la temática Protección Social y Empleo, con lo que se confirma la implementación exitosa de la primera funcionalidad.

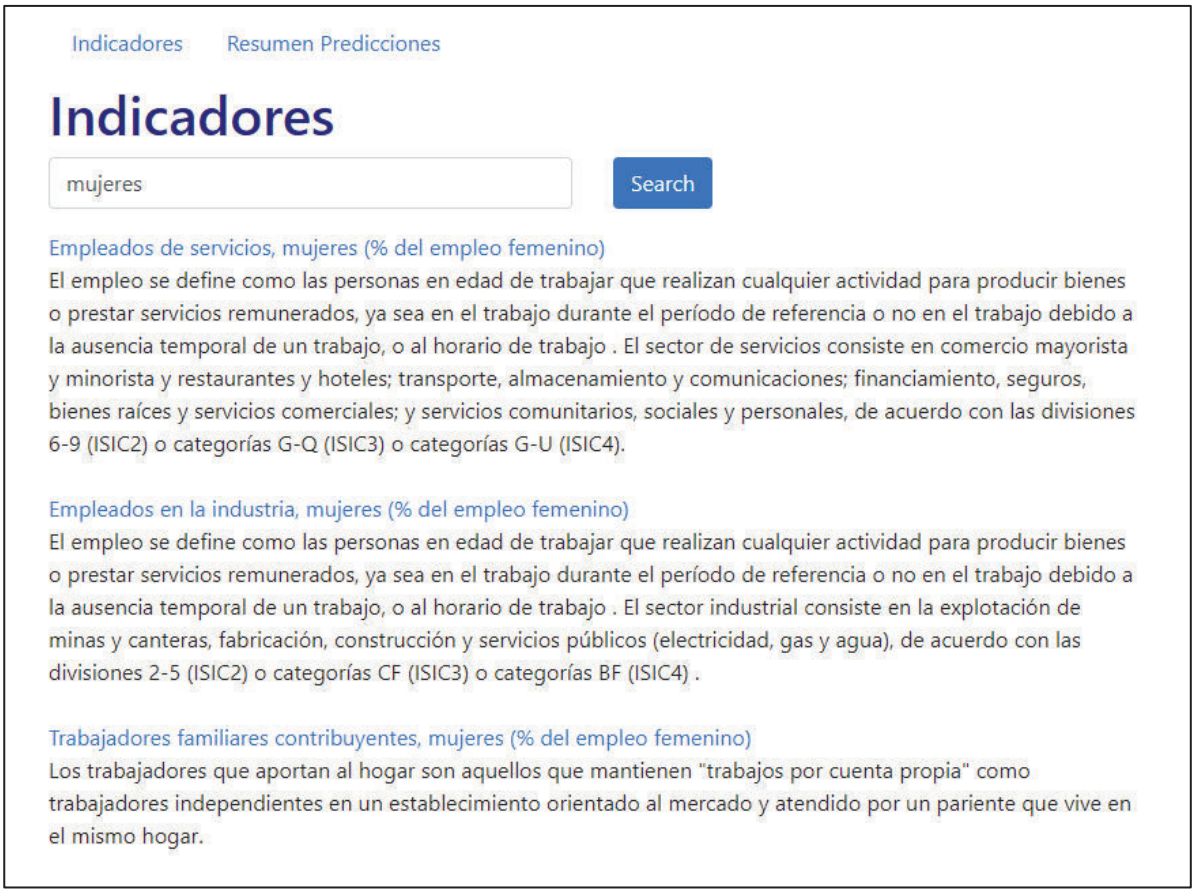

Figura 37. Extracto de la búsqueda de indicadores que contienen la palabra mujeres

La Figura 37, muestra la búsqueda de indicadores de Protección Social y Empleo que contengan la palabra mujeres, comprobando con ello la implementación de la segunda funcionalidad.

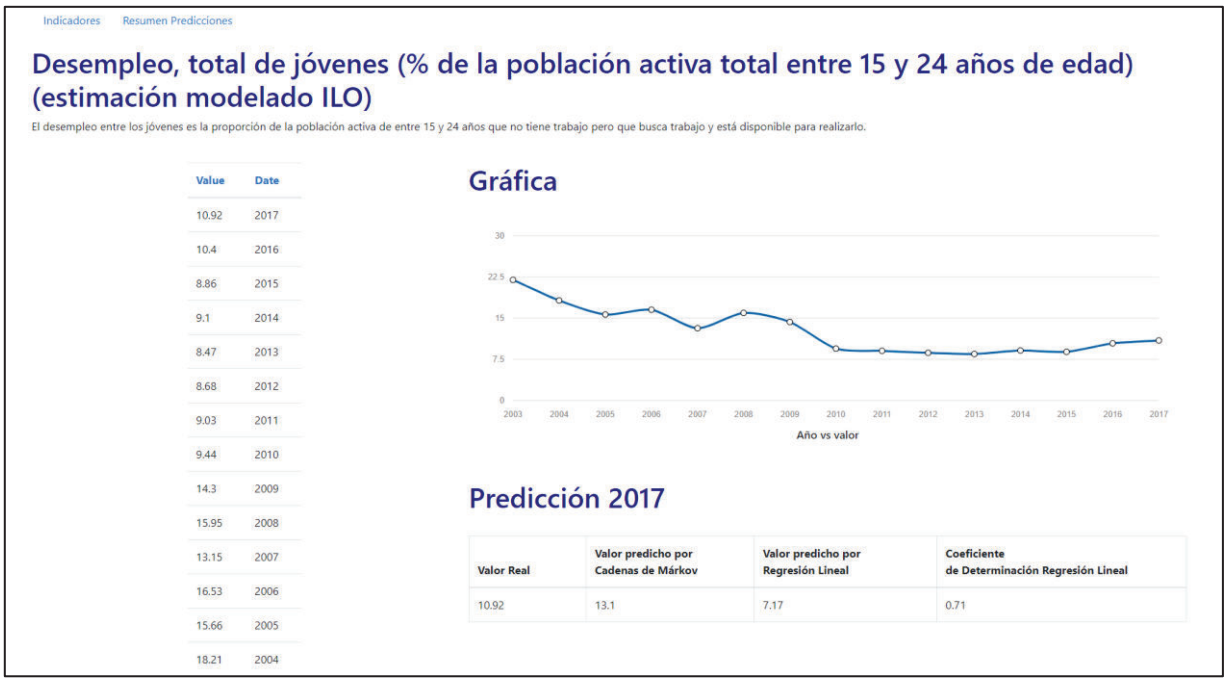

Figura 38. Navegabilidad del indicador hacia sus valores numéricos

Se selecciona al indicador Desempleo, total de jóvenes (% de la población activa total entre 15 y 24 años) (estimación modelado ILO) y se verifica que se navegue hasta la interfaz de valores numéricos, como se muestra en la Figura 38, demostrando así que la funcionalidad de navegación se encuentra bien implementada.

El proceso seguido para probar la funcionalidad de los demás prototipos es el mismo, se establecen las funcionalidades a probar y se realiza la comprobación de cada una de ellas.

3. RESULTADOS Y DISCUSIÓN

3.1 Resultados

Los resultados obtenidos de la implementación de los modelos de Cadenas de *Márkov* y de Regresión Lineal sobre cada uno de los indicadores que describe a la temática de Protección Social y Empleo en Ecuador, son presentados a continuación en la Tabla 15.

Donde:

- **M**, Modelo de Cadenas de *Márkov*
- **R**, Modelo de Regresión Lineal

En la Tabla 15, se especifican los resultados obtenidos al aplicar Cadenas de *Márkov* y Regresión lineal, cuyas columnas describen:

- **Número y nombre del indicador**, de cada uno de los indicadores que representan a la temática de Protección Social y Empleo.
- **Valor real 2017**, al valor obtenido del indicador a través del API de la Banca Mundial para el año 2017.
- **Valor predicción 2017,** al valor obtenido por la predicción del año 2017 realizada por cada modelo predictivo: Cadenas de *Márkov* y Regresión Lineal.
- **Error Residual (%)**, que representa la diferencia entre el valor observado (valor real 2017) y el valor predicho (valor predicción 2017) [27], calculado en forma porcentual.

Tabla 15. Resultados de la Implementación de Cadenas de Márkov y Regresión Lineal

N°	Nombre del Indicador	Valor real 2017	Valor predicción 2017		Error Residual (%)	
			M	R	M	R
1	Empleados de servicios, varones (% de empleo masculino)	46.5	46.28	46.86	0.47	0.77
2	Empleados de servicios, mujeres (% del empleo femenino)	66.41	67.28	67.5	1.31	1.64
3	Empleo informal, masculino (% del total del empleo no agrícola)	49.5	47.16	46.6	4.73	5.86

4	Empleo informal, femenino (% del total del empleo no agrícola)	56.12	51.55	52.42	8.14	6.59
5	Empleos en la industria (% del total de empleos)	18.52	19.37	19.1	4.59	3.13
6	Empleados en la industria, varones (% del empleo masculino)	23.75	24.44	24.46	2.91	2.99
7	Empleados en la industria, mujeres (% del empleo femenino)	10.82	11.28	10.84	4.25	0.18
8	PIB por cada persona empleada (a $ de PPA constantes de 2011)	21912.48	24097.54	24087.95	9.97	9.93
9	Trabajadores familiares contribuyentes, total (% del empleo total)	12.37	10.84	11.47	12.37	7.28
10	Trabajadores familiares contribuyentes, hombres (% del empleo masculino)	6.99	5.92	6.17	15.31	11.73
11	Trabajadores familiares contribuyentes, mujeres (% del empleo femenino)	20.31	18.31	19.63	9.85	3.35
12	Personas que perciben sueldos y asalariados, hombres (% de empleo masculino)	57.34	58.56	58.77	2.13	2.49
13	Personas que perciben sueldos y asalariados, mujeres (% de empleo femenino)	42.35	49.22	43.72	16.22	3.23
14	Empleo vulnerable, total (% del total de empleo)	45.53	43.34	44.82	4.81	1.56
15	Empleo vulnerable, varones (% del empleo masculino)	38.68	37.03	38.16	4.27	1.34
16	Empleo vulnerable, mujeres (% del empleo femenino)	55.64	52.91	55.06	4.91	1.04
17	Relación entre empleo y población, mayores de 15 años, total (%) (estimación modelado ILO)	64.94	64.66	63.9	0.43	1.6
18	Relación entre empleo y población, mayores de 15 años, total (%) (estimación nacional)	65.54	61.77	62.23	5.75	5.05
19	Relación entre empleo y población, mayores de 15 años, hombres (%) (estimación modelado ILO)	78.13	78.16	78.22	0.04	0.12

20	Relación entre empleo y población, mayores de 15 años, hombres (%) (estimación nacional)	78.19	75.08	76.19	3.98	2.56
21	Relación entre empleo y población, mayores de 15 años, mujeres (%) (estimación modelado ILO)	52.0	51.03	49.86	1.87	4.12
22	Relación entre empleo y población, mayores de 15 años, mujeres (%) (estimación nacional)	53.62	50.84	49.14	5.18	8.36
23	Proporción femenina de empleo en la alta y media dirección (%)	35.15	34.59	32.58	1.59	7.31
24	Trabajadores por cuenta propia, total (% del empleo total)	48.72	46.75	47.16	4.04	3.2
25	Trabajadores por cuenta propia, hombre (% de empleo masculino)	42.66	39.73	41.23	6.87	3.35
26	Trabajadores por cuenta propia, mujeres (% del empleo femenino)	57.65	55.05	56.28	4.51	2.38
27	Empleadores, total (% del empleo total)	3.18	3.25	2.34	2.2	26.42
28	Empleadores, hombres (% de empleo masculino)	3.98	4.09	3.07	2.76	22.86
29	Empleadores, mujeres (% del empleo femenino)	2.01	1.88	1.22	6.47	39.3
30	Relación empleo-población, edades 15-24, total (%) (estimación nacional)	41.42	38.31	38.97	7.51	5.92
31	Relación empleo-población, edades entre 15 y 24 años, varones (%) (estimación modelado ILO)	50.69	54.37	50.64	7.26	0.1
32	Relación empleo-población, edades entre 15 y 24 años, varones (%) (estimación nacional)	51.44	48.58	49.15	5.56	4.45
33	Relación empleo-población, edades entre 15 y 24 años, mujeres (%) (estimación modelado ILO)	29.95	31.29	28.01	4.47	6.48
34	Relación empleo-población, edades entre 15 y 24 años, mujeres (%) (estimación nacional)	31.23	27.08	28.62	13.29	8.36

35	Empleos en agricultura (% del total de empleos)	26.93	27.39	25.93	1.71	3.71
36	Empleados en agricultura, hombres (% del empleo masculino)	29.76	29.08	28.69	2.28	3.6
37	Empleados en agricultura, mujeres (% del empleo femenino)	22.77	21.39	21.66	6.06	4.87
38	Desempleo, total (% de la población activa total) (estimación nacional)	3.84	3.28	2.28	14.58	40.63
39	Desempleo, varones (% de la población activa masculina) (estimación modelado ILO)	3.93	4.55	2.3	15.78	41.48
40	Desempleo, varones (% de la población activa masculina) (estimación nacional)	3.05	4.58	2.1	50.16	31.15
41	Desempleo, mujeres (% de la población activa femenina) (estimación modelado OIT)	6.18	4.0	2.88	35.28	53.4
42	Desempleo, mujeres (% de la población activa femenina) (estimación nacional)	4.9	3.92	2.56	20.0	47.76
43	Proporción de jóvenes que no están en educación, empleo o capacitación, total (% de población joven)	17.56	19.34	19.16	10.14	9.11
44	Proporción de jóvenes que no están en educación, empleo o capacitación, hombres (% de la población juvenil masculina)	9.27	10.53	10.37	13.59	11.87
45	Proporción de jóvenes que no están en educación, empleo o capacitación, mujeres (% de población femenina joven)	25.98	28.34	28.11	9.08	8.2
46	Desempleo con educación intermedia (% de la fuerza de trabajo total con educación intermedia)	7.81	7.25	8.31	7.17	6.4
47	Desempleo con educación intermedia, hombres (% de la fuerza laboral masculina con educación intermedia)	5.96	5.78	6.17	3.02	3.52
48	Desempleo con educación intermedia, mujeres (% de la fuerza de trabajo femenina con educación intermedia)	4.71	4.83	4.75	2.55	0.85
49	Desempleo con educación básica (% de la fuerza de	2.96	2.95	2.65	0.34	10.47

	trabajo total con educación básica)					
50	Desempleo con educación básica, hombres (% de la población activa masculina con educación básica)	2.33	2.67	2.19	14.59	6.01
51	Desempleo con educación básica, mujeres (% de la fuerza de trabajo femenina con educación básica)	1.92	2.5	2.13	30.21	10.94
52	Desempleo con educación avanzada (% de la fuerza de trabajo total con educación avanzada)	6.89	5.33	5.16	22.64	25.11
53	Desempleo con educación avanzada, hombres (% de la población activa masculina con educación avanzada)	5.81	4.71	4.25	18.93	26.85
54	Desempleo con educación avanzada, mujeres (% de la fuerza de trabajo femenina con educación avanzada)	4.62	3.97	3.78	14.07	18.18
55	Desempleo, total de jóvenes (% de la población activa total entre 15 y 24 años de edad) (estimación modelado ILO)	10.92	13.1	7.17	19.96	34.34
56	Desempleo, total de jóvenes (% de la población activa total entre 15 y 24 años de edad) (estimación nacional)	8.45	9.12	6.57	7.93	22.25
57	Desempleo, varones jóvenes (% de la población activa masculina entre 15 y 24 años de edad) (estimación modelado ILO)	8.53	10.16	5.57	19.11	34.7
58	Desempleo, varones jóvenes (% de la población activa masculina entre 15 y 24 años de edad) (estimación nacional)	6.57	7.53	5.11	14.61	22.22
59	Desempleo, mujeres jóvenes (% de la población activa femenina entre 15 y 24 años de edad) (estimación modelado ILO)	14.82	17.66	10.05	19.16	32.19
60	Desempleo, mujeres jóvenes (% de la población activa femenina entre 15 y 24 años de edad) (estimación nacional)	11.44	12.12	9.24	5.94	19.23

Se considera más exacta a la predicción cuyo error residual sea el menor, ya que obtener un error residual pequeño, implica que existe una diferencia menor entre el valor observado (real) y el valor predicho (estimado), por lo tanto, el valor predicho es más cercano al valor real, lo que lo hace más exacto. [28].

De los datos presentes en la Tabla 15, se estable que, el modelo de Cadenas de *Márkov* presenta ligeramente una mayor exactitud en cuanto a las predicciones realizadas para el año 2017, ya que cuenta con un total de 31 predicciones más exactas. Por otro lado, el modelo de Regresión Lineal cuenta con un total de 29 predicciones más exactas. Lo que se resumen en que el modelo de Cadenas de *Márkov* tiene un 3 % más de predicciones con mayor exactitud comparado con el modelo de Regresión Lineal. La Tabla 16 y la Figura 39, explican lo anteriormente descrito.

Tabla 16. Predicciones realizadas por los modelos de Cadena de Márkov y Regresión Lineal

	Total, de predicciones	**(%)**
Cadenas de Márkov	31	51,67
Regresión Lineal	29	48,33

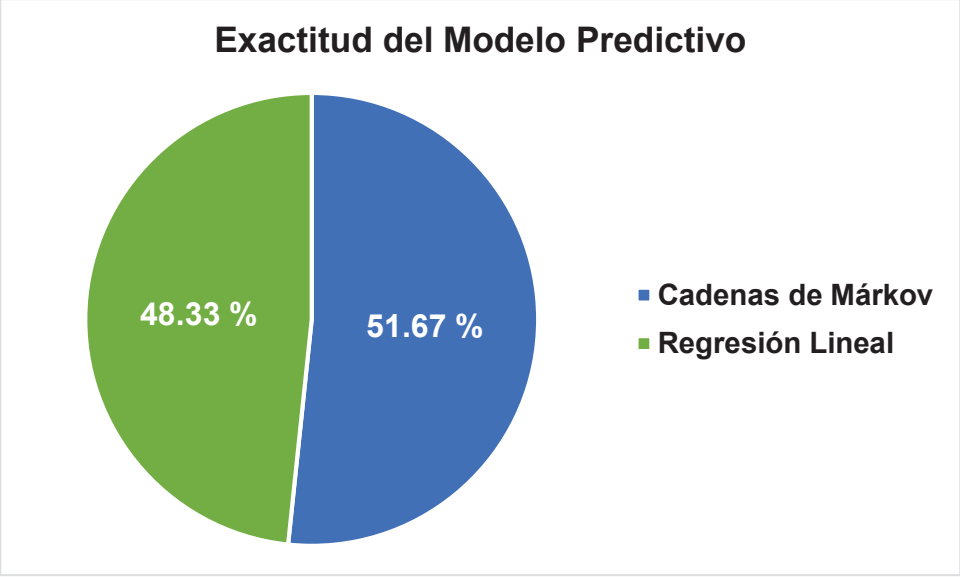

Figura 39. Comparación de exactitud entre modelos de Cadena de Márkov y Regresión Lineal

Fueron escogidos los mejores y los peores casos en cuanto a predicción, del conjunto de datos analizado, para cada modelo: Cadenas de *Márkov* y Regresión Lineal. El comportamiento

encontrado se detalla en las subsecciones: Comportamiento de los datos con el modelo de Cadenas de *Márkov* y Comportamiento de los datos con el modelo de Regresión Lineal.

3.1.1 Comportamiento de los datos con el modelo de Cadenas de Márkov

Mejor Caso

Los datos del indicador Relación entre empleo y población, hombres mayores de 15 años, (%) (estimación modelado ILO), describen a la proporción de la población de hombres de un país que está empleada. Definiendo empleo como las personas en edad de trabajar que durante un período de referencia corto se dedican a cualquier actividad para producir bienes o prestar servicios de pago o ganancia, ya sea en el trabajo durante el período de referencia o a arreglos de tiempo de trabajo [6]. Con estos datos se obtuvo la mejor predicción con el modelo de Cadenas de *Márkov* y están representados en la Figura 40.

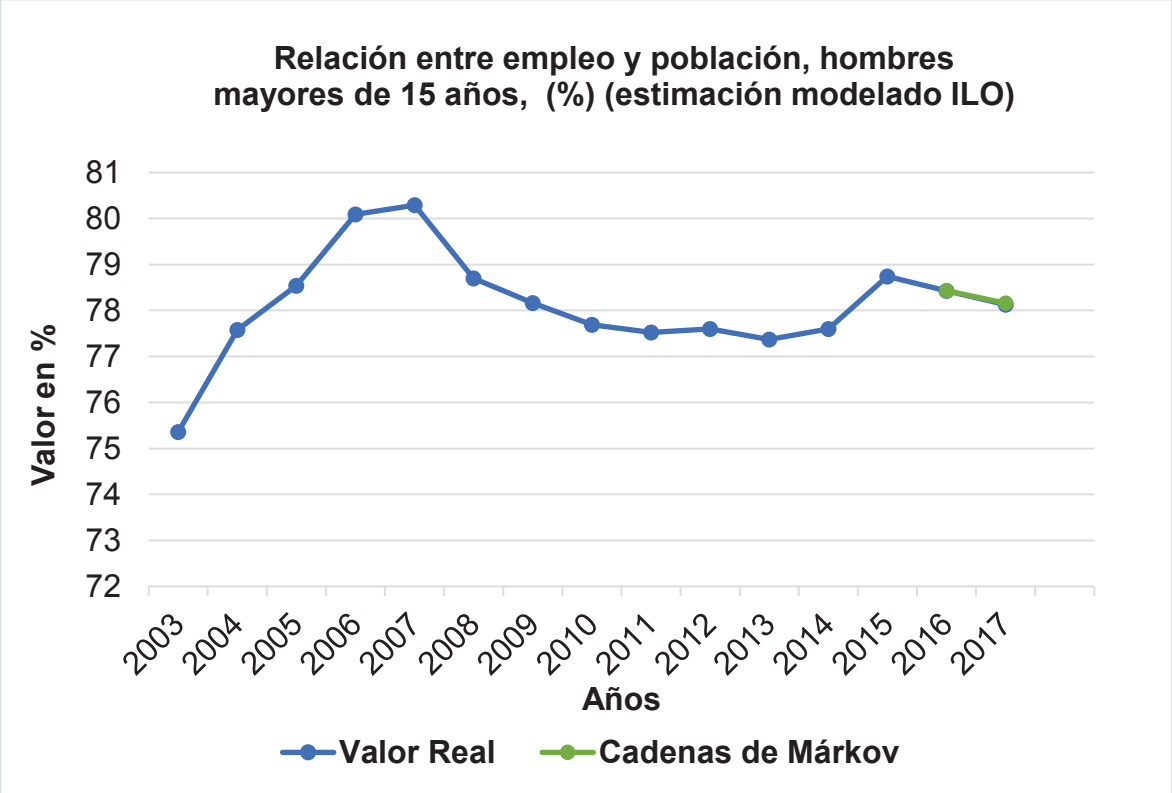

Figura 40. Tendencia de los datos del indicador Relación entre empleo y población, hombres mayores de 15 años, (%) (estimación modelado ILO)

Tabla 17. Datos medidos por año del indicador Relación entre empleo y población, hombres mayores de 15 años, (%) (estimación modelado ILO)

Año	Valor Numérico
2003	75,36
2004	77,58
2005	78,54
2006	80,09
2007	80,29
2008	78,7
2009	78,16
2010	77,69
2011	77,52
2012	77,6
2013	77,37
2014	77,6
2015	78,74
2016	78,43
2017	78,13

El valor real para el año 2017 del indicador Relación entre empleo y población, hombres mayores de 15 años, (%) (estimación modelado ILO), es 78,13 % y el valor de la predicción obtenido por el modelo de Cadenas de *Márkov* es de 78,16 % como se observa en la Figura 40. Además, se evidencia que el comportamiento de los datos observados para el periodo 2003-20017, no tiene una tendencia lineal.

Peor Caso

Los datos del indicador, PIB por cada persona empleada (a $ de PPA constantes de 2011), describen al producto interno bruto (PIB) dividido por el empleo total en la economía del país [6]. Con estos datos se obtuvo la peor predicción con el modelo de Cadenas de *Márkov* y están representados en la Figura 41.

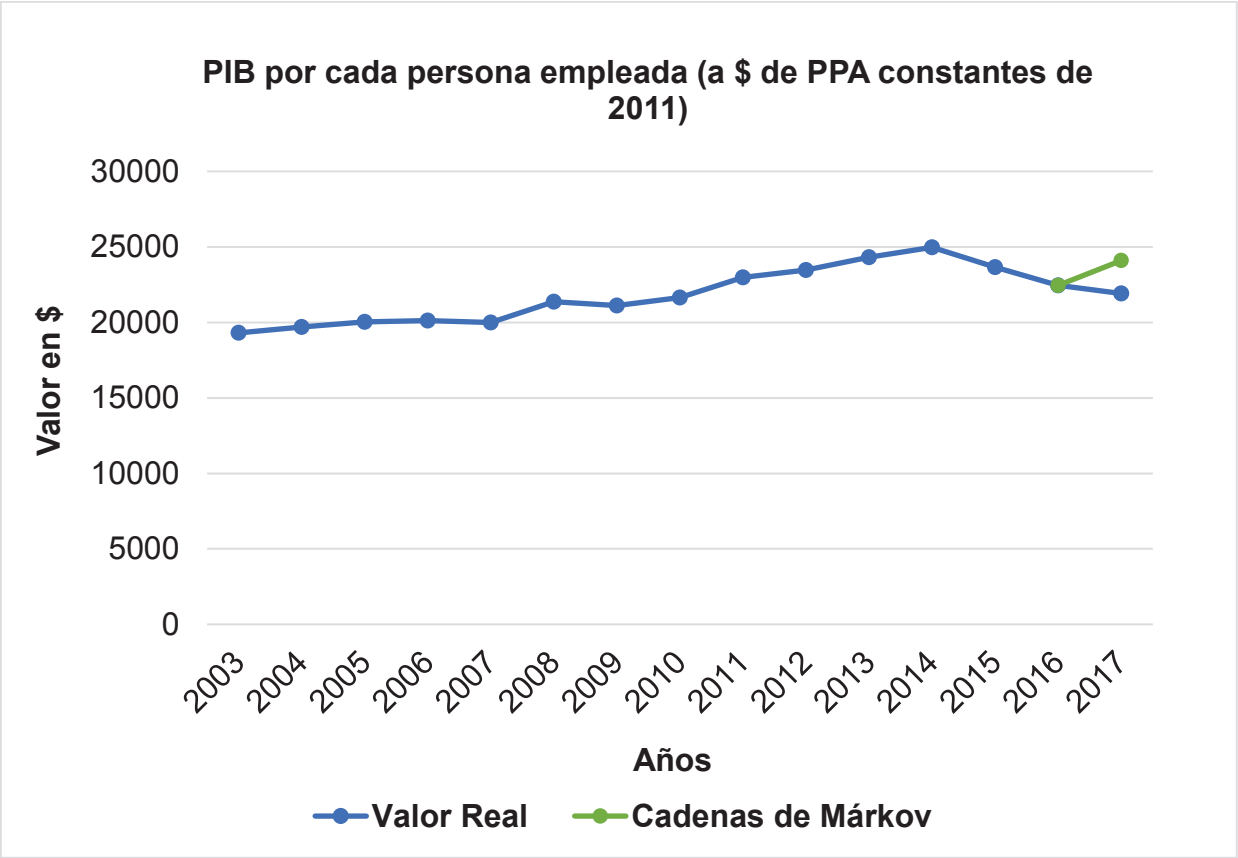

Figura 41. Tendencia de los datos del indicador PIB por cada persona empleada (a $ de PPA constantes de 2011)

Tabla 18. Datos medidos por año del indicador PIB por cada persona empleada (a $ de PPA constantes de 2011)

Año	Valor Numérico
2003	19305,07
2004	19706,34
2005	20045,29
2006	20111,72
2007	19995,66
2008	21362,35
2009	21124,73
2010	21635,75
2011	22970,63
2012	23478,62
2013	24323,97

2014	24970,46
2015	23666,07
2016	22455,1
2017	21912,48

El valor real para el año 2017 del indicador PIB por cada persona empleada (a $ de PPA constantes de 2011), es $ 21912,48 y el valor de la predicción obtenido por el modelo de Cadenas de Márkov es $ 24097,54 como se observa en la Figura 41. Además, se evidencia que el comportamiento de los datos observados para el periodo 2003-20017, tiene una tendencia lineal.

3.1.2 Comportamiento de los datos con el modelo de Regresión Lineal

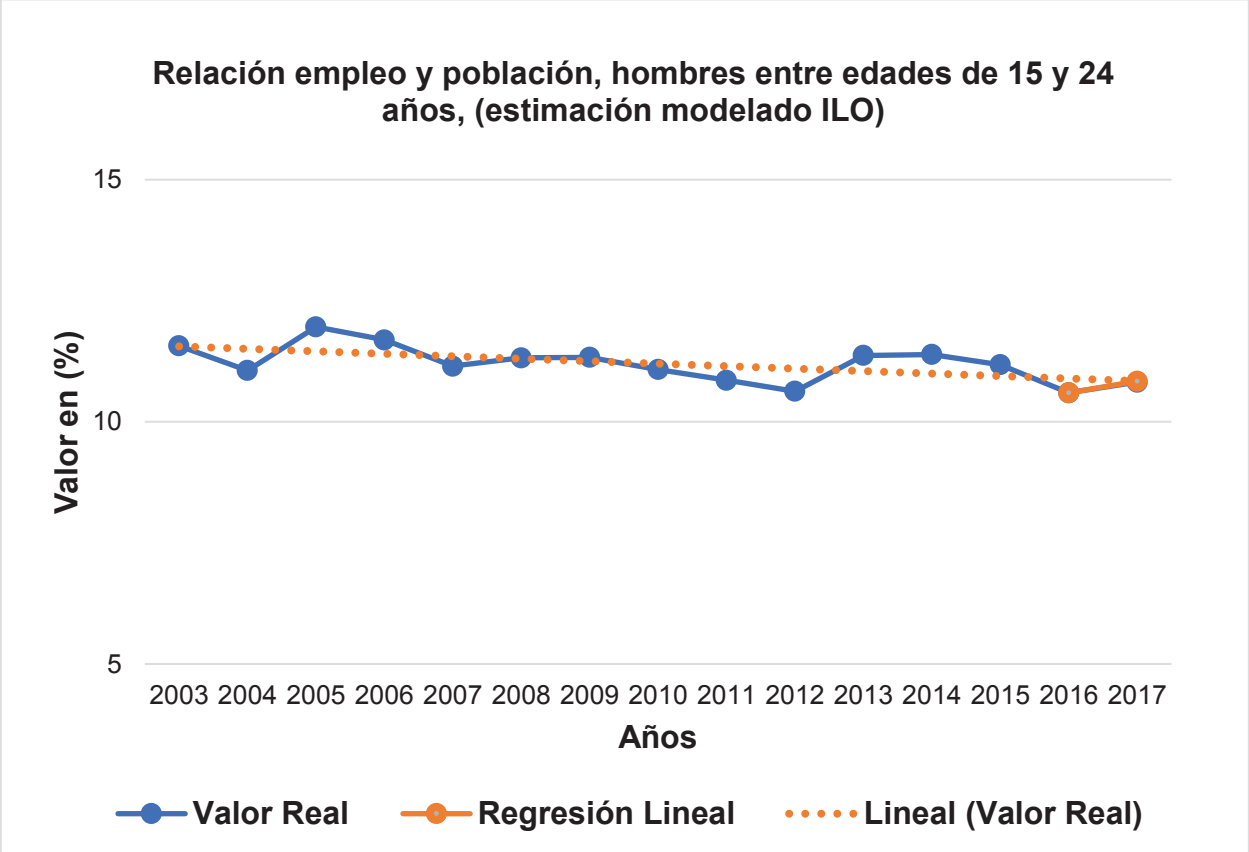

Figura 42. Tendencia de los datos del indicador Relación empleo y población, hombres entre edades de 15 y 24 años, (estimación modelado ILO)

Tabla 19. Datos medidos por año del indicador Relación empleo y población, hombres entre edades de 15 y 24 años, (estimación modelado ILO)

Año	Valor Numérico
2003	11,57
2004	11,06
2005	11,96
2006	11,69
2007	11,15
2008	11,32
2009	11,33
2010	11,08
2011	10,86
2012	10,63
2013	11,37
2014	11,39
2015	11,18
2016	10,6
2017	10,82

Mejor Caso

Los datos del indicador Relación empleo y población, hombres entre edades de 15 y 24 años, (estimación modelado ILO), describen a la proporción de la población de hombres de un país que está empleada. Definiendo empleo como las personas en edad de trabajar que durante un período de referencia corto se dedican a cualquier actividad para producir bienes o prestar servicios de pago o ganancia, ya sea en el trabajo durante el período de referencia o a arreglos de tiempo de trabajo [6]. Con estos datos se obtuvo la mejor predicción con el modelo de Regresión Lineal y están representados en la Figura 42.

El valor real para el año 2017 del indicador Relación empleo y población, hombres entre edades de 15 y 24 años, (estimación modelado ILO), es % 10,82 y el valor de la predicción obtenido por el modelo de Regresión Lineal es % 10,84 como se observa en la Figura 42. Además, se evidencia que el comportamiento de los datos observados para el periodo 2003-20017, tiene una tendencia lineal.

Peor Caso

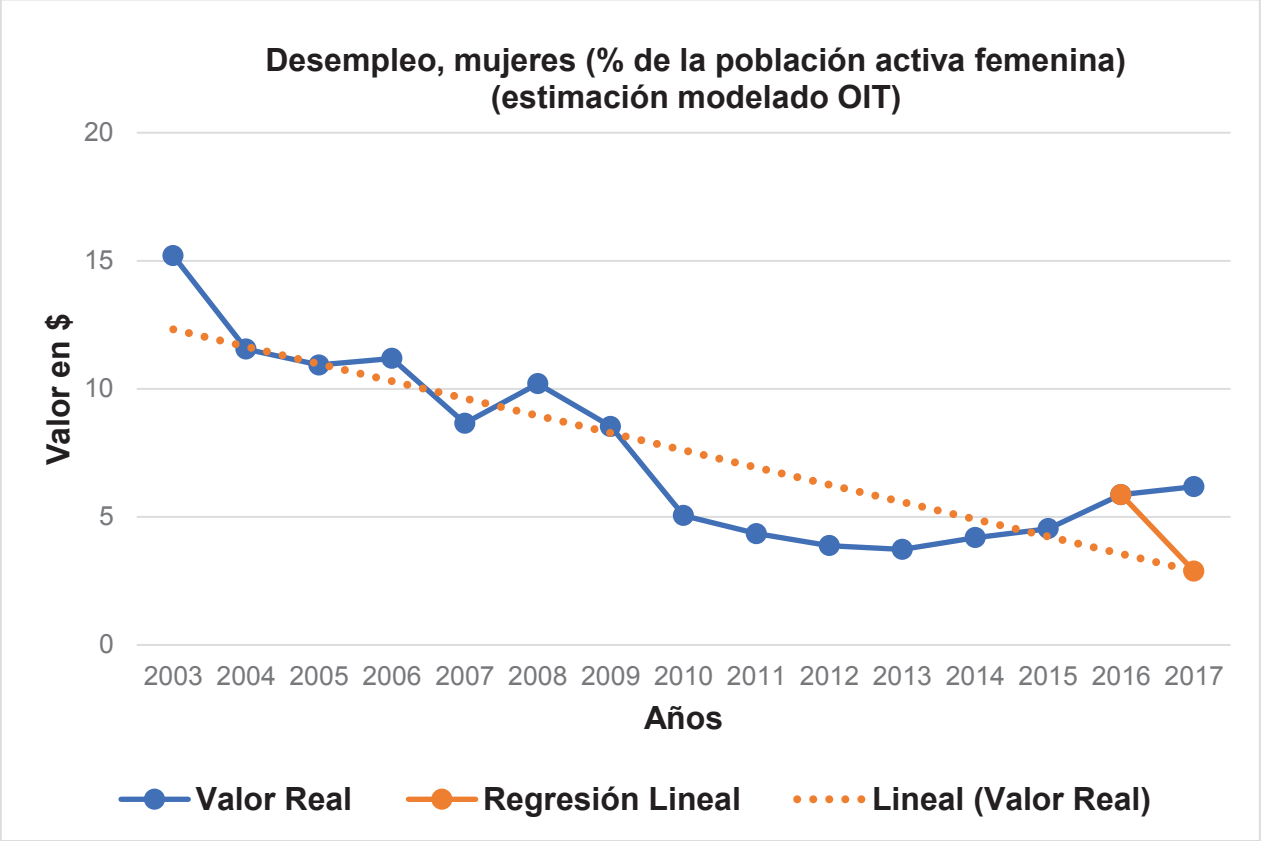

Figura 43. Tendencia de los datos del indicador Desempleo, mujeres (% de la población activa femenina) (estimación modelado OIT)

Tabla 20. Datos medidos por año del indicador Desempleo, mujeres (% de la población activa femenina) (estimación modelado OIT)

Año	Valor Numérico
2003	15,2
2004	11,55
2005	10,93
2006	11,19
2007	8,66
2008	10,2
2009	8,53
2010	5,05
2011	4,35

2012	3,88
2013	3,73
2014	4,19
2015	4,54
2016	5,87
2017	6,18

Los datos del indicador Desempleo, mujeres (% de la población activa femenina) (estimación modelado OIT), describen a la proporción de la población activa que no tiene trabajo pero que busca trabajo y está disponible para realizarlo [6]. Con estos datos se obtuvo la peor predicción con el modelo de Regresión Lineal y están representados en la Figura 43.

El valor real para el año 2017 del indicador PIB por cada persona empleada (a $ de PPA constantes de 2011), es % 6,18 y el valor de la predicción obtenido por el modelo de Regresión Lineal es % 2,88 como se observa en la Figura 43. Además, se evidencia que el comportamiento de los datos observados para el periodo 2003-20017, no tiene una tendencia lineal.

3.2 Discusión

Se compararon los modelos utilizando el error residual, debido a que su cálculo es equivalente en ambos modelos. Sin embargo, al modelo de Regresión Lineal, se le agregó una medida de confiabilidad a la predicción, medida conocida como coeficiente de determinación, que indica que tan confiable es la predicción realizada por el modelo.

La librería de Python utilizada para la aplicación del modelo de Regresión Lineal, statsmodels, ofrece mayor detalle en el análisis de Regresión Lineal, como se ve en la Figura 24, comparada con la librería de Python scikit-learn. Sin embargo, cabe recalcar que ambas librerías cuentan con mucha documentación y son fáciles de usar y de implementar. Por lo que ambas representan una buena opción para realizar el análisis de Regresión Lineal.

A pesar de que el modelo de Cadenas de *Márkov* presentó un 3% más de resultados mejores en cuanto a predicción comparado con el modelo de Regresión Lineal, se considera que con ambos modelos se pueden obtener buenos resultados, partiendo de una elección de modelo

predictivo basada en la observación previa de la distribución de los datos, para nuestro caso y según nuestros resultados, una distribución lineal de los datos analizados, sugieren la elección de un modelo de Regresión Lineal, mientras una distribución no lineal o aleatoria de los datos sugiere la aplicación de un modelo de Cadenas de *Márkov*.

4. CONCLUSIONES

La fuente de datos seleccionada para proveer los datos de la temática Protección Social y Empleo en Ecuador para el período 2003-2017. Esta base de datos de la Banca Mundial extraída del repositorio *Linked Open Data*, proporcionó datos veraces y completos avalados por la Organización Internacional del trabajo, lo que contribuyó en una evaluación de calidad de los métodos predictivos de Cadenas de *Márkov* y Regresión Lineal. Al trabajar con datos veraces se tiene la confiabilidad de realizar la evaluación de la predicción, contrastando la predicción realizada por los modelos con el valor real extraído de la fuente de datos seleccionada.

El prototipado, metodología utilizada para el desarrollo del sistema web en Python en el que se comunican los resultados del análisis, permitió pulir las características del sistema tanto en la parte funcional (implementación de los modelos de Cadenas de *Márkov* y Regresión Lineal para cada indicador de la temática de Protección Social y Empleo), como en el diseño de este. Además, permitió evidenciar errores tempranamente, al añadir y probar funcionalidades periódicamente realizando cambios y mejoras al sistema a tiempo.

La mayor parte de errores funcionales del sistema web se evidenciaron en la etapa de Procesamiento de los datos. Al agregar los datos de los valores de cada indicador para el período 2003-2017 para almacenarlos en la base de datos de MongoDb, la lógica de los scripts diseñados para el almacenamiento fallaba y se omitían ciertos datos, por lo que los modelos arrojaban predicciones deficientes. En trabajos futuros se recomienda tener especial atención en esta etapa del análisis de datos, para entregar a la etapa de Limpieza de los datos, un conjunto coherente de datos sobre los cuales trabajar.

La evaluación de los modelos predictivos de Cadenas de *Márkov* y Regresión Lineal que se desarrolla en el presente proyecto puede ser aplicada en trabajos futuros sobre temáticas de interés como lo son: pobreza, educación, salud, ciencia y tecnología, entre otros y para los

países presentes en la base de datos de la Banca Mundial; siempre que se disponga de datos numéricos continuos para el periodo en el que se requiera trabajar. Dependiendo de la disponibilidad de datos más actuales, además de realizar la evaluación de los modelos, las predicciones otorgadas por los mismos para años futuros como el 2019 o 2020, podrán constituir datos de utilidad para el sector relacionado con la temática de interés escogida.

Las herramientas tecnológicas utilizadas en el desarrollo del sistema web que sirve como medio de comunicación de los resultados de las predicciones y de la evaluación de los modelos, Python como lenguaje de programación, Django como framework de desarrollo web y MongoDb como sistema de base de datos, fueron adecuadas para el proyecto ya que su implementación no significo mayor esfuerzo, al contar con la suficiente documentación para integrar todas las tecnologías, además de ser fáciles de usar. Esto último permitió dedicarle el tiempo y esfuerzo necesarios al análisis de los datos y la evaluación de los modelos predictivos, parte fundamental del proyecto, al no invertir demasiado tiempo en el desarrollo del sistema de comunicación de resultados.

Los resultados obtenidos de las predicciones benefician a personas encargadas de tomar decisiones en base a resultados. Personas como economistas, analistas financieros, estadísticos, entre otros pueden beneficiarse de la aplicación de los modelos y con ello tomar un accionar pertinente. Por otro lado, organizaciones relacionadas a la temática de desempleo en Ecuador, por ejemplo: Instituto Nacional de Estadística y Censos (INEC), Ministerio de trabajo, entre otros; podrán hacer uso de los valores de las predicciones para ejecutar planes de acción en contra del desempleo. Programas para impulsar el empleo en sectores específicos señalados por los resultados de los indicadores predichos, campañas de capacitación a ciertos grupos sociales no beneficiados según las predicciones, son ejemplos de acciones a tomar que beneficiaría a la población desempleada dentro del país.

5. REFERENCIAS

[1] TechTarget, "SearchDataCenter," 25 Julio 2018. [Online]. Available: https://docs.media.bitpipe.com/io_13x/io_139722/item_1600639/Analttica%20para%20la%20transformacion%20empresarial.pdf. [Accessed 21 Marzo 2018].

[2] A. Rayón, "Por qué hablamos del Business Analytics y no solo de Business Intelligence," 6 Diciembre 2015. [Online]. Available: https://blogs.deusto.es/bigdata/por-que-hablamos-del-business-analytics-y-no-solo-de-business-intelligence/. [Accessed 12 Abril 2018].

[3] A. Vehtari and J. Ojanen, " A survey of Bayesian predictive methods for model assessment, selection and comparison.," *Statist. Surv. 6,* no. doi:10.1214/12-SS102. https://projecteuclid.org/euclid.ssu/1356628931, pp. 142--228, 2012.

[4] Biblioteca del Congreso Nacional de Chile, "Biblioteca del Congreso Nacional de Chile," 17 Julio 2018. [Online]. Available: https://datos.bcn.cl/es/informacion/que-es.

[5] C. O'Neil and R. Schutt, Doing Data Science, O'Reilly Media, 2013.

[6] World Bank, "The World Bank Data," 17 Julio 2018. [Online]. Available: https://data.worldbank.org/. [Accessed 17 Julio 2018].

[7] G. Canavos, Probabilidad y Estadísitca, México: McGraw-Hill, 1988.

[8] J. Sánchez, 31 Julio 2018. [Online]. Available: http://economipedia.com/definiciones/cadena-de-markov.html.

[9] J. Behrens, "Principles and Procedures of Exploratory Data Analysis," pp. 131-160, 2018.

[10] EMC Education Services, Data Science & Big Data Analytics:Discovering, Analyzing, Visualizing and Presenting Data, Indianapolis: John Wiley & Sons, Inc., 2015.

[11] T. Erl, W. Khattak and P. Buhler, Big Data Fundamentals, Concepts, Drivers & Techniques, New Jersey: Prentice Hall, 2016.

[12] QuinStreet Inc., "webopedia," [Online]. Available: https://www.webopedia.com/TERM/A/API.html. [Accessed 09 Diciembre 2018].

[13] wbdata, "wbdata," 25 Julio 2018. [Online]. Available: http://wbdata.readthedocs.io/en/latest/index.html.

[14] MongoDB, "mongoDB for Giant Ideas," 21 Julio 2018. [Online]. Available: https://www.mongodb.com/what-is-mongodb.

[15] Organización Internacional de Normalización, "ISO," 28 Julio 2018. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:es.

[16] D. Eastwood, "Quora," 28 Julio 2018. [Online]. Available: https://www.quora.com/What-is-the-minimum-number-of-observations-required-for-doing-regression-analysis-of-a-variable.

[17] C. O'Neil and R. Schutt, Doing Data Science, O'Reilly: ISBN 978-1-449-35865-5., 2013.

[18] J. L. Devore, Probabilidad y Estadística para Ingeniería y Ciencias, Cengage Learning Editores, 2008.

[19] Rice University, University of Houston Clear Lake, and Tufts University, "Online Statistics Education: An Interactive Multimedia Course of Study," Octubre 2017. [Online]. Available: http://onlinestatbook.com/2/estimation/mean.html. [Accessed 25 Noviembre 2018].

[20] scikit learn, "scikit learn," 29 Julio 2018. [Online]. Available: http://scikit-learn.org/stable/index.html.

[21] DataCamp, "DataCamp," 31 Julio 2018. [Online]. Available: https://www.datacamp.com/community/tutorials/markov-chains-python-tutorial.

[22] G. Sánchez, "Vitutor.net," 7 Agosto 2018. [Online]. Available: https://www.vitutor.net/2/11/cuartiles_percentiles.html.

[23] Python , "Python," 7 Agosto 2018. [Online]. Available: https://pypi.org/project/discreteMarkovChain/.

[24] Severino Ribecca, "Catálogo de Visualización de Datos," 23 Julio 2018. [Online]. Available: https://datavizcatalogue.com/ES/metodos/grafica_de_linea.html.

[25] A. Floría, "SIDAR," [Online]. Available: https://www.sidar.org/recur/desdi/traduc/es/visitable/prototype.htm. [Accessed 6 Julio 2018].

[26] balsamiq, "balsamiq," 22 Julio 2018. [Online]. Available: https://balsamiq.com/.

[27] StatTrek.com, "Stat Trek," 2018. [Online]. Available: https://stattrek.com/regression/residual-analysis.aspx. [Accessed 26 Noviembre 2018].

[28] Dataquest Labs, Inc., "DATAQUEST," 2018. [Online]. Available: https://www.dataquest.io/blog/understanding-regression-error-metrics/. [Accessed 26 Noviembre 2018].

[29] I. Kloo, "Textmining: Clustering, Topic Modeling, and Classification," Agosto 2015. [Online]. Available: http://data-

analytics.net/cep/Schedule_files/Textmining%20%20Clustering,%20Topic%20Modeling,%20and%20Classification.htm.

[30] Biblioteca del Congreso Nacional de Chile, "Biblioteca del Congreso Nacional de Chile," [Online]. Available: https://datos.bcn.cl/es/informacion/que-es. [Accessed 11 Abril 2018].

6. ANEXOS

6.1 Tabla de figuras

Figura 1. Arquitectura del sistema web para predicción de valores 4
Figura 2. Temáticas de Linked Open Data 6
Figura 3. Elaboración propia basada en la referencia (World Bank, 2018) 7
Figura 4. Script en Python de Recolección de datos de temática Protección Social y Empleo 10
Figura 5. Contenido de la lista topics de la temática protección social y empleo 11
Figura 6. Script en Python de recolección de datos de indicadores 13
Figura 7. Descripción del indicador con código id: “SL.UEM.NEET.FE.ZS” 14
Figura 8. Extracto del Anexo 1, código para recolectar datos numéricos de indicadores 16
Figura 9. Extracto de los valores numéricos del indicador con id: “SL.SRV.EMPL.FE.ZS” ... 17
Figura 10. Colecciones de la base de datos indicators 18
Figura 11. Extracto de Anexo 3. Organización y procesamiento de datos de temática. 20
Figura 12. Colección predictions_topic 20
Figura 13. Extracto de Anexo 4. Organización y procesamiento de valores numéricos de indicadores. 21
Figura 14. Datos de descripción de Protección Social y Empleo antes y después de traducción 22
Figura 15. Extracto de Anexo 6. Limpieza de valores numéricos de indicadores 23
Figura 16. Extracto de Anexo 6. Limpieza de valores numéricos de indicadores 23
Figura 17. Extracto de Anexo 6. Limpieza de valores numéricos de indicadores 24
Figura 18. Extracto de la colección predictions_indicatorid 24
Figura 19. Elementos de línea de regresión 26
Figura 20. Valores observados y estimados en una línea de regresión 28
Figura 21. Ejemplo del cálculo del error estándar de la estimación 29
Figura 22. Gráfica de valores por año del indicador con id “SL.SRV.EMPL.MA.ZS”, tomada del sistema web 32
Figura 23. Extracto del Anexo 7. Implementación de Regresión Lineal Simple 34
Figura 24. Resultado de la ejecución de la implementación en Python de Regresión Lineal 35

Figura 25. Ejemplo de Cadena de Márkov 38
Figura 26. Extracto de Anexo 8. Implementación de Cadenas de Márkov 45
Figura 27. Resultado de la ejecución de la implementación en Python de Cadenas de Márkov 46
Figura 28. Extracto de tabla con valores numéricos de indicador 47
Figura 29. Tabla con valores predichos por cada modelo 48
Figura 30. Extracto de tabla de resumen de valores predichos 48
Figura 31. Gráfico de líneas con valores numéricos de indicador 49
Figura 32. Prototipo de lista de temáticas 50
Figura 33. Prototipo de lista de Indicadores de temática 51
Figura 34. Prototipo de valores numéricos del indicador 52
Figura 35. Prototipo de resumen de resultados de predicciones 52
Figura 36. Extracto de la lista de indicadores de temática Protección Social y Empleo 53
Figura 37. Extracto de la búsqueda de indicadores que contienen la palabra mujeres 54
Figura 38. Navegabilidad del indicador hacia sus valores numéricos 55
Figura 39. Comparación de exactitud entre modelos de Cadena de Márkov y Regresión Lineal 61
Figura 40. Tendencia de los datos del indicador Relación entre empleo y población, hombres mayores de 15 años, (%) (estimación modelado ILO) 62
Figura 41. Tendencia de los datos del indicador PIB por cada persona empleada (a $ de PPA constantes de 2011) 64
Figura 42. Tendencia de los datos del indicador Relación empleo y población, hombres entre edades de 15 y 24 años, (estimación modelado ILO) 65
Figura 43. Tendencia de los datos del indicador Desempleo, mujeres (% de la población activa femenina) (estimación modelado OIT) 67

6.2 Tabla de tablas

Tabla 1. Temáticas de la Base de datos del Banco Mundial 8
Tabla 2. Descripción de parámetros de la función de wbdata get_topic 9

Tabla 3. Descripción de parámetros de la función de wbdata search_indicators 11
Tabla 4. Descripción de parámetros de la función de wbdata get_indicator 12
Tabla 5. Descripción de parámetros de la función de wbdata get_data 14
Tabla 6. Descripción de campos de la colección predictions_topic 19
Tabla 7. Descripción de campos de la colección predictions_indicatorid 19
Tabla 8. Descripción de campos de colección predictions_indicator 19
Tabla 9. Datos del indicador con id “SL.SRV.EMPL.MA.ZS” 31
Tabla 10. Valores calculados para la variable dependiente e independiente 33
Tabla 11. Valores estimados de y, variable dependiente 33
Tabla 12. Datos del indicador con id “SL.SRV.EMPL.MA.ZS” 39
Tabla 13. Valores numéricos ordenados de menor a mayor del indicador "Empleados de servicios, varones (% de empleo masculino)” 40
Tabla 14. Espacio de estados del modelo de Cadenas de Márkov 41
Tabla 15. Resultados de la Implementación de Cadenas de Márkov y Regresión Lineal 56
Tabla 16. Predicciones realizadas por los modelos de Cadena de Márkov y Regresión Lineal 61
Tabla 17. Datos medidos por año del indicador Relación entre empleo y población, hombres mayores de 15 años, (%) (estimación modelado ILO) 63
Tabla 18. Datos medidos por año del indicador PIB por cada persona empleada (a $ de PPA constantes de 2011) 64
Tabla 19. Datos medidos por año del indicador Relación empleo y población, hombres entre edades de 15 y 24 años, (estimación modelado ILO) 66
Tabla 20. Datos medidos por año del indicador Desempleo, mujeres (% de la población activa femenina) (estimación modelado OIT) 67

6.3 Anexo 1. Código para extracción de datos numéricos de indicadores

```
1   #Autor: Nathaly Cumbicos
2   #Fecha Creación: 15-abril-2018
3   #Archivo: recolectIndicatorsValue.py
4
5   #Librerías
6   import wbdata
7   import datetime
8
```

```
9   #Obteniendo datos de concepto de indicadores
10  #Temática: Protección Social y Empleo
11  #Campos: id, name, source, sourceNote,
12  #sourceOrganization, topic
13  #Temática-Código: Social Protection & Labor-10
14
15  #Buscando indicadores relacionados con empleo en la API
16
17  #función
18
#search_indicators(criteriobusqueda="employ",codigoTematica="10",mostrarso
loId=False)
19
20  employment =
wbdata.search_indicators("employ",topic="10",display=False)
21
22  #Lista de Ids de indicadores obtenidos de la búsqueda
23  employmentList = []
24
25  #Recorriendo lista employment
26  for indicator_employment in employment:
27      employmentList.append(indicator_employment["id"])#Almacenando
únicamente id
28
29  #Definiendo criterios de búsqueda
30  #Período de tiempo: 2003-2017
31  #País: ECU
32
33  data_date = (datetime.datetime(2003, 1, 1), datetime.datetime(2017, 1,
1))
34  country = 'ECU'
35
36  #Lista que contiene los datos de indicadores obtenidos de la búsqueda
37  indicatorList = []
38
39  #Lista con indicadores que no tienen datos o falla su extracción de
datos
40  indicatorFailing = []
41
42  #Lista que ejecuta la función de extracción de datos numéricos del
indicador (get_data)
43  data = []
44
45  #Recorriendo lista employmentList
46  for indicatorID in employmentList:
47      try:
48          #Ejecución función get_data
49
#get_data(codigoIndicador=idsObtenidosBusqueda,codigoPais=ECU,data_date=20
03-2016)
50          data =
wbdata.get_data(indicatorID,country=country,data_date=data_date)
51          if data is not None: #Datos no vacíos
52              indicatorList.append(data) #Guardando los datos numéricos
en indicatorList
53      except Exception:
```

```
54          indicatorFailing.append(indicatorID) #Guardando id de
indicador en indicatorFailing
```

6.4 Anexo 2. Lista de países contemplados en la base de datos del Banco Mundial

Código Id	Nombre
ABW	Aruba
AFG	Afghanistan
AFR	Africa
AGO	Angola
ALB	Albania
AND	Andorra
ANR	Andean Region
ARB	Arab World
ARE	United Arab Emirates
ARG	Argentina
ARM	Armenia
ASM	American Samoa
ATG	Antigua and Barbuda
AUS	Australia
AUT	Austria
AZE	Azerbaijan
BDI	Burundi
BEA	East Asia & Pacific (IBRD-only countries)
BEC	Europe & Central Asia (IBRD-only countries)
BEL	Belgium
BEN	Benin
BFA	Burkina Faso
BGD	Bangladesh
BGR	Bulgaria
BHI	IBRD countries classified as high income
BHR	Bahrain
BHS	Bahamas, The
BIH	Bosnia and Herzegovina
BLA	Latin America & the Caribbean (IBRD-only countries)
BLR	Belarus
BLZ	Belize
BMN	Middle East & North Africa (IBRD-only countries)

BMU	Bermuda
BOL	Bolivia
BRA	Brazil
BRB	Barbados
BRN	Brunei Darussalam
BSS	Sub-Saharan Africa (IBRD-only countries)
BTN	Bhutan
BWA	Botswana
CAA	Sub-Saharan Africa (IFC classification)
CAF	Central African Republic
CAN	Canada
CEA	East Asia and the Pacific (IFC classification)
CEB	Central Europe and the Baltics
CEU	Europe and Central Asia (IFC classification)
CHE	Switzerland
CHI	Channel Islands
CHL	Chile
CHN	China
CIV	Cote d'Ivoire
CLA	Latin America and the Caribbean (IFC classification)
CME	Middle East and North Africa (IFC classification)
CMR	Cameroon
COD	Congo, Dem. Rep.
COG	Congo, Rep.
COL	Colombia
COM	Comoros
CPV	Cabo Verde
CRI	Costa Rica
CSA	South Asia (IFC classification)
CSS	Caribbean small states
CUB	Cuba
CUW	Curacao
CYM	Cayman Islands
CYP	Cyprus
CZE	Czech Republic
DEA	East Asia & Pacific (IDA-eligible countries)
DEC	Europe & Central Asia (IDA-eligible countries)
DEU	Germany
DFS	IDA countries classified as Fragile Situations
DJI	Djibouti
DLA	Latin America & the Caribbean (IDA-eligible countries)

DMA	Dominica
DMN	Middle East & North Africa (IDA-eligible countries)
DNF	IDA countries not classified as Fragile Situations
DNK	Denmark
DNS	IDA countries in Sub-Saharan Africa not classified as fragile situations
DOM	Dominican Republic
DSA	South Asia (IDA-eligible countries)
DSF	IDA countries in Sub-Saharan Africa classified as fragile situations
DSS	Sub-Saharan Africa (IDA-eligible countries)
DXS	IDA total, excluding Sub-Saharan Africa
DZA	Algeria
EAP	East Asia & Pacific (excluding high income)
EAR	Early-demographic dividend
EAS	East Asia & Pacific
ECA	Europe & Central Asia (excluding high income)
ECS	Europe & Central Asia
ECU	Ecuador
EGY	Egypt, Arab Rep.
EMU	Euro area
ERI	Eritrea
ESP	Spain
EST	Estonia
ETH	Ethiopia
EUU	European Union
FCS	Fragile and conflict affected situations
FIN	Finland
FJI	Fiji
FRA	France
FRO	Faroe Islands
FSM	Micronesia, Fed. Sts.
FXS	IDA countries classified as fragile situations, excluding Sub-Saharan Africa
GAB	Gabon
GBR	United Kingdom
GEO	Georgia
GHA	Ghana
GIB	Gibraltar
GIN	Guinea
GMB	Gambia, The
GNB	Guinea-Bissau
GNQ	Equatorial Guinea

GRC	Greece
GRD	Grenada
GRL	Greenland
GTM	Guatemala
GUM	Guam
GUY	Guyana
HIC	High income
HKG	Hong Kong SAR, China
HND	Honduras
HPC	Heavily indebted poor countries (HIPC)
HRV	Croatia
HTI	Haiti
HUN	Hungary
IBB	IBRD, including blend
IBD	IBRD only
IBT	IDA & IBRD total
IDA	IDA total
IDB	IDA blend
IDN	Indonesia
IDX	IDA only
IMN	Isle of Man
IND	India
INX	Not classified
IRL	Ireland
IRN	Iran, Islamic Rep.
IRQ	Iraq
ISL	Iceland
ISR	Israel
ITA	Italy
JAM	Jamaica
JOR	Jordan
JPN	Japan
KAZ	Kazakhstan
KEN	Kenya
KGZ	Kyrgyz Republic
KHM	Cambodia
KIR	Kiribati
KNA	St. Kitts and Nevis
KOR	Korea, Rep.
KWT	Kuwait
LAC	Latin America & Caribbean (excluding high income)

LAO	Lao PDR
LBN	Lebanon
LBR	Liberia
LBY	Libya
LCA	St. Lucia
LCN	Latin America & Caribbean
LCR	Latin America and the Caribbean
LDC	Least developed countries: UN classification
LIC	Low income
LIE	Liechtenstein
LKA	Sri Lanka
LMC	Lower middle income
LMY	Low & middle income
LSO	Lesotho
LTE	Late-demographic dividend
LTU	Lithuania
LUX	Luxembourg
LVA	Latvia
MAC	Macao SAR, China
MAF	St. Martin (French part)
MAR	Morocco
MCA	Central America
MCO	Monaco
MDA	Moldova
MDE	Middle East (developing only)
MDG	Madagascar
MDV	Maldives
MEA	Middle East & North Africa
MEX	Mexico
MHL	Marshall Islands
MIC	Middle income
MKD	Macedonia, FYR
MLI	Mali
MLT	Malta
MMR	Myanmar
MNA	Middle East & North Africa (excluding high income)
MNE	Montenegro
MNG	Mongolia
MNP	Northern Mariana Islands
MOZ	Mozambique
MRT	Mauritania

MUS	Mauritius
MWI	Malawi
MYS	Malaysia
NAC	North America
NAF	North Africa
NAM	Namibia
NCL	New Caledonia
NER	Niger
NGA	Nigeria
NIC	Nicaragua
NLD	Netherlands
NLS	Non-resource rich Sub-Saharan Africa countries, of which landlocked
NOR	Norway
NPL	Nepal
NRS	Non-resource rich Sub-Saharan Africa countries
NRU	Nauru
NXS	IDA countries not classified as fragile situations, excluding Sub-Saharan Africa
NZL	New Zealand
OED	OECD members
OMN	Oman
OSS	Other small states
PAK	Pakistan
PAN	Panama
PER	Peru
PHL	Philippines
PLW	Palau
PNG	Papua New Guinea
POL	Poland
PRE	Pre-demographic dividend
PRI	Puerto Rico
PRK	Korea
PRT	Portugal
PRY	Paraguay
PSE	West Bank and Gaza
PSS	Pacific island small states
PST	Post-demographic dividend
PYF	French Polynesia
QAT	Qatar
ROU	Romania
RRS	Resource rich Sub-Saharan Africa countries

RSO	Resource rich Sub-Saharan Africa countries, of which oil exporters
RUS	Russian Federation
RWA	Rwanda
SAS	South Asia
SAU	Saudi Arabia
SCE	Southern Cone
SDN	Sudan
SEN	Senegal
SGP	Singapore
SLB	Solomon Islands
SLE	Sierra Leone
SLV	El Salvador
SMR	San Marino
SOM	Somalia
SRB	Serbia
SSA	Sub-Saharan Africa (excluding high income)
SSD	South Sudan
SSF	Sub-Saharan Africa
SST	Small states
STP	Sao Tome and Principe
SUR	Suriname
SVK	Slovak Republic
SVN	Slovenia
SWE	Sweden
SWZ	Swaziland
SXM	Sint Maarten (Dutch part)
SXZ	Sub-Saharan Africa excluding South Africa
SYC	Seychelles
SYR	Syrian Arab Republic
TCA	Turks and Caicos Islands
TCD	Chad
TEA	East Asia & Pacific (IDA & IBRD countries)
TEC	Europe & Central Asia (IDA & IBRD countries)
TGO	Togo
THA	Thailand
TJK	Tajikistan
TKM	Turkmenistan
TLA	Latin America & the Caribbean (IDA & IBRD countries)
TLS	Timor-Leste
TMN	Middle East & North Africa (IDA & IBRD countries)
TON	Tonga

TSA	South Asia (IDA & IBRD)
TSS	Sub-Saharan Africa (IDA & IBRD countries)
TTO	Trinidad and Tobago
TUN	Tunisia
TUR	Turkey
TUV	Tuvalu
TWN	Taiwan, China
TZA	Tanzania
UGA	Uganda
UKR	Ukraine
UMC	Upper middle income
URY	Uruguay
USA	United States
UZB	Uzbekistan
VCT	St. Vincent and the Grenadines
VEN	Venezuela, RB
VGB	British Virgin Islands
VIR	Virgin Islands (U.S.)
VNM	Vietnam
VUT	Vanuatu
WLD	World
WSM	Samoa
XKX	Kosovo
XZN	Sub-Saharan Africa excluding South Africa and Nigeria
YEM	Yemen, Rep.
ZAF	South Africa
ZMB	Zambia
ZWE	Zimbabwe

6.5 Anexo 3. Organización y procesamiento de valores de temática

```
1   #Autor: Nathaly Cumbicos
2   #Fecha Creación: 10-mayo-2018
3   #Archivo: baseTopics.py
4
5   import wbdata
6   #Librería de Python para establecer conexión con MongoDB
7   from pymongo import MongoClient
8
9   #Instancia de cliente para la conexión a la base
10  Client = MongoClient()
11  #Base de datos
12  db = Client ["Indicators"]
```

```
#Colección
topics = db["predictions_topic"]

#Obteniendo datos de temática: Protección Social y Empleo
#Campos: id, value, sourcenote
#Temática-Código: Social Protection & Labor-10

topicList = [] #Lista que extrae los datos del API
#función get_topic(codigo_tematica=10, mostrarsoloId=False)
topicList = wbdata.get_topic(10,display=False)

#Secuencial para id
cont = 0
#Recorriendo la lista de temáticas
for item in topicList:
    #Creando diccionario topic que guarda la data extraída
    topic = {}
    cont +=1
    topic ["id"] = cont
    topic ["id_topic"] = item["id"]
    topic ["name"] = item["value"]
    topic ["description"] = item["sourceNote"]
    #Guardando el diccionario topic en lista topics
    topics.insert(topic)
```

6.6 Anexo 4. Organización y procesamiento de valores numéricos de indicadores

```
#Autor: Nathaly Cumbicos
#Fecha Creación: 10-mayo-2018
#Archivo: baseValueIndicators.py

#Librerías
import wbdata
import datetime
from pymongo import MongoClient

#Instancia de cliente para la conexión a la base
Client = MongoClient()
#Base de datos
db = Client ["Indicators"]
#Colección
indicators = db["predictions_indicator"]
#Obteniendo datos de concepto de indicadores
#Temática: Protección Social y Empleo
#Campos: id, name, source, sourceNote,
#sourceOrganization, topic
#Temática-Código: Social Protection & Labor-10

#Buscando indicadores relacionados con empleo en la API

#función
```

```
25
#search_indicators(criteriobusqueda="employ",codigoTematica="10",mostrarsoloId=False)
26
27  employment = wbdata.search_indicators("employ",topic="10",display=False)
28
29  #Lista de Ids de indicadores obtenidos de la búsqueda
30  employmentList = []
31
32  #Recorriendo lista employment
33  for indicator_employment in employment:
34      employmentList.append(indicator_employment["id"])#Almacenando únicamente id
35
36  #Definiendo criterios de búsqueda
37  #Período de tiempo: 2003-2017
38  #País: ECU
39
40  data_date = (datetime.datetime(2003, 1, 1), datetime.datetime(2017, 1, 1))
41  country = 'ECU'
42
43  #Lista que contiene los datos de indicadores obtenidos de la búsqueda
44  indicatorList = []
45
46  #Lista con indicadores que no tienen datos o falla su extracción de datos
47  indicatorFailing = []
48
49  #Lista que ejecuta la función de extracción de datos numéricos del indicador (get_data)
50  data = []
51
52  #Recorriendo lista employmentList
53  for indicatorID in employmentList:
54      try:
55          #Ejecución función get_data
56          #get_data(codigoIndicador=idsObtenidosBusqueda,codigoPais=ECU,data_date=2003-2017)
57          data = wbdata.get_data(indicatorID,country=country,data_date=data_date)
58          if data is not None: #Datos no vacíos
59              indicatorList.append(data) #Guardando los datos numéricos en indicatorList
60      except Exception:
61          indicatorFailing.append(indicatorID) #Guardando id de indicador en indicatorFailing
62
63
64  #Procesamiento
65  #Lista de indicadores que poseen datos para un
66  #número de años mayor a 3
67  indicatorValues = []
68
```

```
69  for item in indicatorList:
70      if len(item) != 3:
71          indicatorValues.append(item)
72
73  cont = 0;
74
75  #Almacenamiento
76  #Recorriendo la lista indicatorValues
77  for element in indicatorValues:
78      for valueIndicator in element:
79          #Diccionario indicator
80          indicator = {}
81          cont += 1
82          indicator ["id"] = cont
83          indicator ["id_indicator"] = valueIndicator["indicator"]["id"]
84          if(valueIndicator ["value"] is None):
85              indicator ["value"] = valueIndicator["value"]
86          else:
87              indicator ["value"] =
round(float(valueIndicator["value"]),2)
88              indicator ["date"] = int(valueIndicator["date"])
89              #Guardando en colección indicators
90              indicators.insert(indicator)
91
```

6.7 Anexo 5.Organización y procesamiento de descripción de indicadores

```
1   #Autor: Nathaly Cumbicos
2   #Fecha Creación: 10-mayo-2018
3   #Archivo: baseDataIndicators.py
4
5
6   import wbdata
7   import datetime
8   from pymongo import MongoClient
9
10  #Instancia de cliente para la conexión a la base
11  Client = MongoClient()
12  #Base de datos
13  db = Client ["Indicators"]
14  #Colecciones
15  indicators = db["predictions_indicator"]
16  indicators_id = db["predictions_indicatorid"]
17
18  #Obteniendo datos de concepto de indicadores
19  #Temática: Protección Social y Empleo
20  #Campos: id, name, source, sourceNote,
21  #sourceOrganization, topic
22  #Temática-Código: Social Protection & Labor-10
23
24  #Buscando indicadores relacionados con empleo en la API
```

```
25  employment = indicators.distinct("id_indicator")
26  employmentList = employment
27
28  #Lista que contiene los datos de indicadores obtenidos de la búsqueda
29  indicatorList = []
30
31  #Lista con indicadores que no tienen datos o falla su extracción de
datos
32  indicatorFailing = []
33
34  #Lista que ejecuta la función de extracción de datos del indicador
(get_indicator)
35  data = []
36
37  #Recorriendo lista employmentList
38  for indicatorID in employmentList:
39      try:
40          #Ejecución función get_indicator
41
#get_indicator(codigoIndicador=idsObtenidosBusqueda,codigoTematica="10",mo
strarsoloId=False)
42          data =
wbdata.get_indicator(indicator=indicatorID,topic="10",display=False)
43          if data is not None: #Datos no vacíos
44              indicatorList.append(data) #Guardando los datos de
indicadores en indicatorList
45      except Exception: #Datos con error
46          indicatorFailing.append(indicatorID) #Guardando id de
indicador en indicatorFailing
47
48  cont = 0
49  #Recorriendo indicatorList
50  for item in indicatorList:
51      for objectIndicator in item:
52          indicator = {}#Diccionario indicator
53          cont +=1
54          indicator ["id"] = cont
55          indicator ["id_indicator"] = objectIndicator["id"]
56          indicator ["name"] = objectIndicator["name"]
57          indicator ["description"] = objectIndicator["sourceNote"]
58          indicator ["topicId"] = "10"
59          #Almacenando indicator en colección predictions_indicatorid
60          indicators_id.insert(indicator)
```

6.8 Anexo 6. Limpieza de valores numéricos de indicadores

```
1   #Autor: Nathaly Cumbicos
2   #Fecha Creación: 15-mayo-2018
3   #Archivo: cleanValuesIndicator.py
4
5   import wbdata
6   import datetime
```

```
7
8   #Obteniendo datos de concepto de indicadores
9   #Temática: Protección Social y Empleo
10  #Campos: id, name, source, sourceNote,
11  #sourceOrganization, topic
12  #Temática-Código: Social Protection & Labor-10
13
14  #Buscando indicadores relacionados con empleo en la API
15
16  #función
17
#search_indicators(criteriobusqueda="employ",codigoTematica="10",mostrarso
loId=False)
18
19  employment =
wbdata.search_indicators("employ",topic="10",display=False)
20
21  #Lista de Ids de indicadores obtenidos de la búsqueda
22  employmentList = []
23
24  #Recorriendo lista employment
25  for indicator_employment in employment:
26      employmentList.append(indicator_employment["id"])#Almacenando
únicamente id
27
28  #Definiendo criterios de búsqueda
29  #Período de tiempo: 2003-2016
30  #País: ECU
31
32  data_date = (datetime.datetime(2003, 1, 1), datetime.datetime(2017, 1,
1))
33  country = 'ECU'
34
35  #Lista que contiene los datos de indicadores obtenidos de la búsqueda
36  indicatorList = []
37
38  #Lista con indicadores que no tienen datos o falla su extracción de
datos
39  indicatorFailing = []
40
41  #Lista que ejecuta la función de extracción de datos numéricos del
indicador (get_data)
42  data = []
43
44  #Recorriendo lista employmentList
45  for indicatorID in employmentList:
46      try:
47          #Ejecución función get_data
48
#get_data(codigoIndicador=idsObtenidosBusqueda,codigoPais=ECU,data_date=20
03-2016)
49          data =
wbdata.get_data(indicatorID,country=country,data_date=data_date)
50          if data is not None: #Datos no vacíos
51              indicatorList.append(data) #Guardando los datos numéricos
en indicatorList
```

```
52      except Exception:
53          indicatorFailing.append(indicatorID) #Guardando id de
indicador en indicatorFailing
54
55
56  #Limpieza
57  #Lista de indicadores que poseen datos para un
58  #número de años mayor a 3
59  indicatorValues = []
60
61  for item in indicatorList:
62      if len(item) >= 3:
63          indicatorValues.append(item)
64
65  #Limpieza
66  lista = []
67  #Recorriendo lista indicatorsValue, lista con valores numéricos de
indicadores
68  for indicador in indicatorValues:
69      #Si el primer registro (año=2016) no tiene valor, se descarta
70      for index, dataIndicator in enumerate(indicador):
71          if(index == 0) :
72              #Eliminando registros que no tienen valor para el año
2016.
73              if (dataIndicator["value"] is None or
dataIndicator["date"] != '2016'):
74                  indicatorValues.remove(indicador)
75              else:
76                  lista.append(indicador)
77
78  cont = 0;
79
80  #Almacenamiento en base
81  #Recorriendo lista
82  for element in lista:
83      #Recorriendo los campos de element
84      for valueIndicator in element:
85          indicator = {}#Diccionario indicator
86          #Descartando registros de periodos de tiempo no anuales:
Trimestrales.
87          #Anuales: tamaño de período de tiempo es 4.
88          if (len(valueIndicator["date"])==4):
89              cont += 1
90              indicator ["id"] = cont
91              indicator ["id_indicator"] =
valueIndicator["indicator"]["id"]
92              if(valueIndicator ["value"] is None):
93                  indicator ["value"] = valueIndicator["value"]
94              else:
95                  indicator ["value"] =
round(float(valueIndicator["value"]),2)
96              indicator ["date"] = int(valueIndicator["date"])
97              indicators.insert(indicator)
98
```

6.9 Anexo 7. Implementación de Regresión Lineal

```
1   #Autor: Nathaly Cumbicos
2   #Fecha : 25-junio-2018
3   #archivo: predictionRegresion.py
4
5   #Predicción Valor 2017: Modelo Regresión Lineal
6   #Indicador ejemplo:
7   #id: "SL.SRV.EMPL.MA.ZS"
8   #nombre: "Empleados de servicios, varones (% de empleo masculino)"
9
10  #Librerías
11  import pandas as pd
12  import numpy as np
13  from pymongo import MongoClient
14  from sklearn.model_selection import train_test_split
15  from sklearn.linear_model import LinearRegression
16  import statsmodels.api as sm
17
18  #Conexión a BD
19  Client = MongoClient()
20  db = Client ["Indicators"]
21  indicators = db["predictions_indicator"]
22
23  years = []
24  values = []
25
26  #Recolección de todos los datos numéricos
27  #Periodo: 2003-2017
28
29  result = indicators.find({"id_indicator":" SL.SRV.EMPL.MA.ZS
",
"date":{ '$gt': 2002, '$lt':2018 }},{"_id":0,"value":1,"date":1})
30
31  for item in result:
32      if (item["value"]):
33          values.append(item["value"])
34          years.append(item["date"])
35
36  #=====1. Preparación de datos========
37  #print('\n\tIndicador:\n\tEmpleados en agricultura, hombres (% del
empleo masculino)')
38  d = { 'años':years, 'valores':values}
39  dataSet = pd.DataFrame(d)
40
41  #Datos
42  print("Datos")
43  print("Y :",values)
44  print("X :",years)
45
46  #Variable predictora: años
47  x = dataSet.iloc[:, :-1].values
48  #Variable respuesta: valor
49  y = dataSet.iloc[:, 1].values
```

```
#====================================

#========2. Entrenamiento

result = sm.OLS(y, sm.add_constant(x)).fit()
print(result.summary())
print('R^2 = ', result.rsquared)

#Ecuación de la recta de regresión
print('\n\t-----Ecuación de la recta de regresión ---------')
print ('\n\ty = ',result.params[1],'x +',result.params[0])
#==================================

#========2. Predicción
#Predicción para el año 2017
ypred = result.predict(sm.add_constant(x)[0])

#Resultado de la predicción
print ('\n\t-----Valor real vs Valor predicho 2016-------')
print ('\n\t',values[0] ,"||",round(ypred.tolist()[0],2))
#=======================================
```

6.10 Anexo 8. Implementación de Cadenas de Márkov

```
#Autor: Nathaly Cumbicos
#Fecha : 25-junio-2018
#archivo: predictionMarkov.py

#Predicción Valor 2017
#Modelo de Cadenas de Márkov
#Indicador ejemplo:
#id: "SL.EMP.OWAC.FE.ZS"
#nombre: "Trabajadores por cuenta propia, mujeres (% de la población empleada)"

#Librerías
from pymongo import MongoClient
from sklearn.preprocessing import normalize
import numpy as np
from discreteMarkovChain import markovChain

#Matriz de Estados
def transition_matrix(transitions):
    n =  max(transitions) #number of states
    M = [[0]*n for _ in range(n)]

    for (i,j) in zip(transitions,transitions[1:]):
        M[i-1][j-1] += 1
    return M

#Valor real del indicador
def realValue(id):
```

```
28      Client = MongoClient()
29      db = Client ["Indicators"]
30      indicators = db["predictions_indicator"]
31
32      values = []
33
34      result =
indicators.find({"id_indicator":id,"date":2017},{"_id":0,"value":1,"date":
1})
35      for item in result:
36          values.append(item["value"])
37
38      value = values[0]
39      return (value)
40
41
42  #Prediccion con Cadenas de Márkov
43  Client = MongoClient()
44  db = Client ["Indicators"]
45  indicators = db["predictions_indicator"]
46
47  values = []
48
49  result = indicators.find({"id_indicator":"SL.SRV.EMPL.MA.ZS","date":{
'$gt': 2002, '$lt':2017 }},{"_id":0,"value":1,"date":1})
50
51  for item in result:
52      if (item["value"]):
53          values.append(item["value"])
54
55  print("\t-------Cadena de Márkov---------\n")
56  #===== 1. Cálculo de Cuartiles ========
57  #Ordenando datos de menor a mayor
58  sortedValues = sorted(values)
59  length = len(sortedValues)
60  print('Número de datos:',length,'\n')
61  print('Valores desordenados:\n',values,'\n')
62  print('Valores ordenados:\n',sortedValues,'\n')
63
64  #Cálculo de los cuartiles
65  #Cuartil 1
66  ran1 = length*(0.25)
67  state1 = sortedValues[int(ran1)-1]+(abs(ran1)-
abs(int(ran1)))*(sortedValues[int(ran1)]-sortedValues[int(ran1)-1])
68  st1 = round(state1,2)
69
70  #Cuartil 2
71  ran2 = length*(0.50)
72  state2 = sortedValues[int(ran2)-1]+(abs(ran2)-
abs(int(ran2)))*(sortedValues[int(ran2)]-sortedValues[int(ran2)-1])
73  st2 = round(state2,2)
74
75  #Cuartil 3
76  ran3 = length*(0.75)
77  state3 = sortedValues[int(ran3)-1]+(abs(ran3)-
abs(int(ran3)))*(sortedValues[int(ran3)]-sortedValues[int(ran3)-1])
```

```
78  st3 = round(state3,2)
79
80  vectorStates = []
81  state = 0
82  #====================================
83
84  #===== 2. Clasificación en estados====
85  print("Estado 1       ||Estado 2       ||Estado 3       ||Estado 4
")
86  print(sortedValues[0],"-",st1," ||",st1,"-",st2,"||",st2,"-
",st3,"||",st3,"-",sortedValues[-1])
87
88  #---------Matriz de estados ---------
89  #Clasificación de los valores en estados
90  for valueIndicator in values:
91      if(valueIndicator >= sortedValues[0] and valueIndicator<st1):
92          state = 1
93          vectorStates.append(state)
94      elif (valueIndicator >= st1 and valueIndicator<st2):
95          state = 2
96          vectorStates.append(state)
97      elif (valueIndicator >= st2 and valueIndicator<st3):
98          state = 3
99          vectorStates.append(state)
100     elif (valueIndicator >= st3):
101         state = 4
102         vectorStates.append(state)
103 print("\nVector de Estados:")
104 print(vectorStates,'\n')
105
106     #Matriz de estados
107 m = transition_matrix(vectorStates)
108
109 print("Matriz de transición sin normalizar:")
110 print("e1 e2 e3 e4")
111 for row in m: print('  '.join('{0}'.format(x) for x in row))
112 print('\n')
113
114 #Normalización de matriz de estados
115 Column_Normalized1 = normalize(m, norm='l1', axis=0)
116 #Cálculo de la transpuesta
117 Z = Column_Normalized1.transpose()
118 print("Transpuesta de Matriz Normalizada:")
119 print(Z)
120 #====================================
121
122 #===== 3. Predicción ================
123 P = np.array(Z)
124 #Creación del modelo
125 mc = markovChain(P)
126 mc.computePi("linear")
127 #Cálculo del vector predicción
128 vectorPrediction = mc.pi
129 print("\nVector Predicción:\n")
130 print(vectorPrediction)
131 print("\n")
```

```
132
133 prediction =0
134
135 probability = np.argmax(vectorPrediction)+1
136
137 #Estado con mayor probabilidad
138 print("Estado con mayor probabilidad:")
139 print(probability)
140
141 #Calculo del promedio de cada cuartil
142 if(probability == 1 ):
143     prediction= round((sortedValues[0]+st1)/2,2)
144
145 elif(probability == 2 ):
146     prediction = round((st1+st2)/2,2)
147
148 elif(probability == 3 ):
149     prediction = round((st2+st3)/2,2)
150
151 elif(probability == 4 ):
152     prediction = round((st3+sortedValues[-1])/2,2)
153
154 #Predicción
155 print("-----Valor real vs Valor predicho 2017--------")
156 print(realValue("SL.SRV.EMPL.MA.ZS"),"||",prediction)
157 print("\n")
158 print("\n")
```

ORDEN DE EMPASTADO

ESCUELA POLITECNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS

CARRERA:

INGENIERÍA EN SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

ORDEN DE EMPASTADO

De acuerdo con lo estipulado en el Art. 27 del Instructivo para la Implementación de la Unidad de Titulación en las carreras y programas vigentes de la Escuela Politécnica Nacional, aprobado por el Consejo de Docencia el 22 de abril de 2015, y una vez verificado el cumplimiento del formato de presentación establecido, se autoriza la impresión y encuadernación final del Trabajo de Titulación presentado por:

NATHALY FERNANDA CUMBICOS MIZHQUERO

Fecha de autorización: Quito, 29 de enero 2019

PhD. Myriam Hernández
DECANA DE LA
FACULTAD DE INGENIERIA DE SISTEMAS