UADY
UNIVERSIDAD AUTÓNOMA DE YUCATÁN

Desarrollo de una plataforma Web basada en Aprendizaje de Máquina para la gestión e inferencia de información de una biorrefinería de producción de hidrógeno a partir de aguas residuales.

TESIS

Presentada como requisito para obtener el grado de

Maestro en Ciencias de la Computación

Por:

Ing. Luis Arturo Rodríguez Filigrana

Directores:

- Dr. Jorge Ricardo Gómez Montalvo, Universidad Autónoma de Yucatán.
- Dr. Francisco Moo Mena, Universidad Autónoma de Yucatán.
- Dr. Jesús Ixbalank Torres Zúñiga, Universidad de Guanajuato.

Mérida, Yucatán, México, septiembre del 2020

Dedicatoria

A mis padres quienes me brindaron todo su apoyo para lograr este gran sueño, y por las herramientas necesarias para salir adelante.

A mi hermano que siempre estuvo para apoyarme hasta en los momentos más difíciles.

A mis amigos, que sin ellos esta gran experiencia no hubiese valido la pena.

Índice general

Índice de figuras

Índice de tablas

1. Introducción

En una biorrefinería, una materia prima (de origen vegetal, de origen animal, de origen mineral, etc.) de base biológica se procesa para producir productos como combustible, productos químicos o energía/calor [1]. El desarrollo de una biorrefinería requiere una cantidad sustancial de información: parámetros, variables, modelos de reacciones conocidas, propiedades termodinámicas, eficiencias de proceso o datos experimentales [2]. A los grandes volúmenes de datos se les conoce como *Big Data* [3]. *Big Data* es un término para conjuntos de datos masivos que tienen una estructura grande, variada y compleja, con las dificultades de almacenamiento, análisis y visualización para futuros procesos o resultados [3].

La ciencia computacional de extraer información útil de grandes volúmenes de datos o bases de datos se conoce como minería de datos. Es una disciplina, que se encuentra en la intersección de estadísticas, aprendizaje automático, administración de datos y bases de datos, reconocimiento de patrones, inteligencia artificial y otras áreas [4]. Así pues, el objetivo de la minería de datos es descubrir información nueva y útil en bases de datos, y repositorios, empleando diversos algoritmos de minería de datos, tales como, *Support Vector Machines* (SVM), *Classification And Regression Trees* (CART), *Clustering*, entre otros, que permiten encontrar patrones de información para un análisis más claro en grandes volúmenes de datos [5].

Al igual que la minería de datos, *Machine Learning* (ML) juega un papel importante como un componente fundamental del análisis de datos, y es uno de los principales impulsores de la revolución del *Big Data* [6]. La razón de esto se debe a su capacidad de aprender de los datos y proporcionar información basada en ellos, así como decisiones y predicciones [7].

Hoy día, diversas áreas del conocimiento (biología, inteligencia artificial, bioquímica, etc.) generan y almacenan enormes volúmenes de datos, que describen sus operaciones, productos y procesos. Las áreas de conocimiento antes mencionadas no cuentan con los algoritmos necesarios tanto de minería de datos (SVM, CART, algoritmo A priori, etc.) como de *Machine Learning* (*Supervised Learning* (SL), *Unsupervised Learning* (UL) y *Reinforcement Learning* (RL)) para procesar toda la información [8].

El campo de la minería de datos y aprendizaje automático, abordan la problemática de extraer patrones de interés, asociaciones, reglas, cambios y anomalías de los datos para mejorar el proceso de toma de decisiones en las diversas áreas del conocimiento antes mencionadas [8]. Además, existe una creciente necesidad de procesar y presentar datos de manera transparente y reproducible, y de proporcionar marcos de análisis que sean útiles y rentables.

En esta tesis se propone una plataforma Web que utiliza métodos y algoritmos de aprendizaje automático y minería de datos para recibir, almacenar, procesar y analizar grandes cantidades de datos que genera una biorrefinería que produce hidrógeno a partir de aguas residuales.

1.1. Contexto y problemática

Esta tesis ocurre en el contexto de una biorrefinería propuesta por las universidades Universidad de Guadalajara (UdG) - Universidad de Guanajuato (UGto), la cual utiliza sensores para la medición y/o estimación de variables de estado de la fase anaeróbica acidogénica, el fotobiorreactor y la celda electroquímica microbiana.

1.1.1. La biorrefinería

La figura 1 muestra la biorrefinería propuesta para tratar residuos agroindustriales. Primeramente, se propone un digestor anaerobio para tratar los residuos agroindustriales y generar principalmente ácidos grasos volátiles (AGVs). Adicionalmente, la digestión anaerobia generará dióxido de carbono (CO_2) como un subproducto. Los ácidos grasos volátiles generados por la digestión anaerobia se utilizarán como sustrato en las celdas electroquímicas microbianas, las cuales producirán hidrógeno. Por otro lado, el CO_2 generado como subproducto se inyecta en el fotobiorreactor de producción de microalgas para generar biomasa revalorizable.

La biomasa microalgal obtenida se caracteriza bioquímicamente mediante técnicas de determinación de carbohidratos, lípidos, proteínas y actividad antioxidante total (DPPH, carotenoides y fenoles totales) para su posterior revalorización [9].

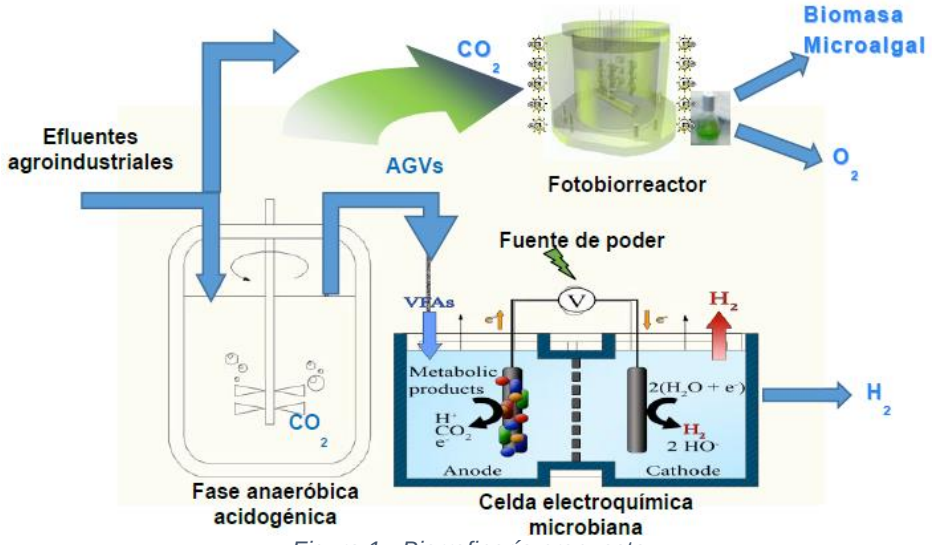


Figura 1.- Biorrefinería propuesta

1.1.2. Toma de decisiones

Como se mencionó en la subsección anterior, la biorrefinería propuesta produce AGVs, CO_2, biomasa de microalgas e hidrógeno, principalmente. Estos productos se generan cada cierto tiempo, de manera que el proceso en su conjunto genera datos mediante sensores (por hardware y por software). Para determinar si la biorrefinería está funcionando correctamente o no, es necesario llevar a cabo un proceso de toma de decisiones utilizando la información generada por la biorrefinería para conocer el estado actual de la misma y aplicar alguna corrección si es necesario.

Sin embargo, el llevar a cabo un proceso de toma de decisiones lleva un cierto tiempo en completarse, desde la captura de los datos hasta tener las acciones correspondientes a aplicar en la biorrefinería. En la figura 2, se muestra un esquema propuesto del proceso de toma de decisiones y sus posibles consecuencias, cada uno de los pasos se describen a continuación:

- **(2a)** La biorrefinería propuesta genera datos cada 10 minutos mediante sensores.
- **(2b)** Estos datos son capturados de forma manual por un biotecnólogo que trabaja con la biorrefinería **(2c)**.
- **(2d)** La información capturada de forma manual puede ocasionar que algún dato contenga valores erróneos y, afecte el proceso de toma de decisiones.
- **(2e)** La toma de decisiones se lleva a cabo cuando se tiene información recopilada, de la biorrefinería, de 3 o 5 meses dependiendo de los criterios establecidos por los biotecnólogos.
- **(2g)** Una vez que se tiene una decisión, un biotecnólogo se encarga de aplicar las correcciones, si las hay, a la biorrefinería.

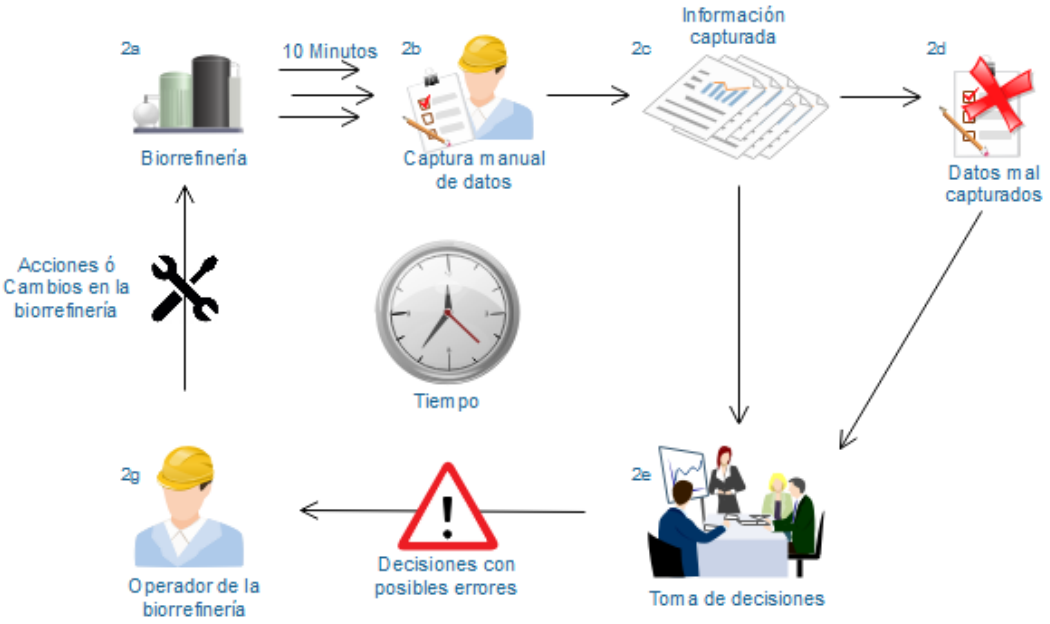


Figura 2.- Propuesta de proceso de toma de decisiones y sus posibles consecuencias

En las siguientes subsecciones se explican las problemáticas de los procesos planteados anteriormente.

1.1.3. Recepción y almacenamiento de datos

La biorrefinería genera, aproximadamente, 550 datos cada 10 minutos mediante sensores. En la Tabla 1 se muestran las entradas de la fase anaeróbica acidogénica: Tasa de dilución (Dil), Demanda Química de Oxígeno (DQO) y AGVs. De igual manera, en la misma tabla 1, la fase anaeróbica acidogénica tiene como variables de estado: biomasa, DQO, AGVs y CO_2.

Fase anaeróbica acidogénica	
Entradas	**Variables de estado**
Tasa de dilución (d^{-1})	Biomasa (gL^{-1})
Demanda química de oxígeno (gL^{-1})	Demanda química de oxígeno (gL^{-1})
Acetato (g/L)	Acetato (mgL^{-1})
	Dióxido de carbono (CO_2)

Tabla 1.- Variables de entrada y de estado de la fase anaeróbica acidogénica

Los AGVs generados por la fase anaeróbica acidogénica, se utilizan como dato de entrada en las celdas electroquímicas microbianas, tal y como se muestra en la tabla 2. De igual manera, tienen como variables de estado: biomasa de bacterias anodofílicas, biomasa de bacterias metanogénicas, biomasa de bacterias hidrogenotrópicas, mediador de oxidación, corriente, flujo de hidrógeno y AGVs.

Celda electroquímica microbiana	
Entradas	**Variables de interés**
Tasa de dilución (d^{-1})	Acetato (mgL^{-1})
Acetato (mgL^{-1})	Bacterias anodofílicas (mgL^{-1})
	Bacterias metanogénicas (mgL^{-1})
	Bacterias hidrogenotrópicas (mgL^{-1})
	Mediador de oxidación (L^{-1})
	Corriente (A)
	Flujo de hidrógeno (Ld^{-1})

Tabla 2.- Variables de entrada y de estado de la celda electroquímica microbiana

Por otro lado, el CO_2 generado por la fase anaeróbica acidogénica se inyecta, a la entrada del fotobiorreactor de producción de microalgas. En la tabla 3 se

muestran las entradas del fotobiorreactor: flujo de entrada (Q_in) y concentración de nutriente. Las variables de estado de interés del fotobiorreactor son la biomasa, el nutriente, la cuota de nutriente y el carbono inorgánico.

Fotobiorreactor	
Entradas	**Variables de interés**
Tasa de dilución (d^{-1})	Biomasa (gL)
Concentración de nutriente (mg/L)	Nutriente (gL)
	Cuota nutriente (gN/gC)
	Carbono inorgánico total (mol TIC/L)

Tabla 3.- Datos de entrada y de interés del fotobiorreactor

Uno de los problemas es que, los biotecnólogos capturan de forma manual los datos generados por la biorrefinería. En algunos casos, es posible que el biotecnólogo haya capturado mal uno o varios datos en sus reportes, esto es un problema al momento de tomar una decisión para realizar un cambio en la biorrefinería, ya que, puede afectar al rendimiento de la biorrefinería y causar errores.

Otro problema de la captura de los datos generados por la biorrefinería es su almacenamiento. Ya que los datos son generados mediante sensores y los biotecnólogos ser encargan de hacer la captura manual de los mismos y no cuentan con un medio computacional necesario para almacenar dicha información.

El tiempo utilizado desde la captura de los datos de forma manual hasta llevar a cabo la toma de decisiones para aplicar alguna acción a la biorrefinería es otro problema. Se sabe que la biorrefinería genera datos cada 10 minutos, esto quiere decir que cada 10 minutos la biorrefinería genera nueva información. Si pasan mas de 10 minutos entre la captura de datos hasta la toma de decisiones y, se va a aplicar alguna acción, es posible que no se detecten las posibles fallas a tiempo y la biorefinería entre en un estado de falla prolongado. Esto es debido a que las

acciones a aplicar consideran la información generada en un estado anterior de la biorrefinería y no del estado actual al momento de realizar una acción.

1.1.4. Visualización de resultados

Cuando se aplica una acción en la biorrefinería, debido a la toma de decisiones, los biotecnólogos tienen que esperar mucho tiempo, horas o días, para que la biorrefinería produzca nueva información y llevar a cabo otra toma de decisiones para determinar si la acción aplicada con anterioridad dio efecto. Esto resulta ser un problema como se planteó en la subsección pasada, ya que, al momento de visualizar los resultados de una acción, la biorrefinería ya cambió a otro estado de tiempo y ya produjo nueva información. Para tener un mejor control del funcionamiento de la biorrefinería, es necesario conocer los estados de esta en casi tiempo real.

1.2. Objetivos

El propósito de esta tesis es diseñar e implementar una plataforma Web que reciba, almacene, infiera y permita visualizar información generada por la biorrefinería propuesta, con el objetivo de agilizar el proceso de toma de decisiones, producir diversos resultados estadísticos utilizando diversos métodos de análisis estadísticos y, generar tablas de recomendaciones de rendimiento para la biorrefinería, aplicando algoritmos de Machine Learning (ML).

1.2.1. Objetivo general

Desarrollar una plataforma Web que permita recibir, almacenar, inferir y visualizar información, para la toma de decisiones, cercanos a tiempo real, utilizando algoritmos de minería de datos y de ML, en el contexto de una biorrefinería que produce hidrógeno a partir de aguas residuales, a partir de los datos que genera.

1.2.2. Objetivos específicos

- Desarrollar un sistema que permita obtener, almacenar y procesar los datos de una biorrefinería en una base de datos para generar información estadística que pueda ser visualizada a través de una interfaz amigable en un ambiente Web y que permita conocer el estado de sus procesos.

- Desarrollar una Interfaz Humano-Máquina (HMI, por sus siglas en inglés), en un ambiente Web, para la visualización de la información almacenada de la biorrefinería, con el propósito de que los biotecnólogos puedan acceder a esta información para la inferencia de datos

- Establecer un estándar de almacenamiento de información extraída de la biorrefinería para su uso genérico para que cualquier otra biorrefinería pueda tener acceso a este sistema y poder brindarle las mismas herramientas para el procesamiento de la información.

- Determinar los algoritmos de análisis estadístico requeridos por los biotecnólogos para implementarlos dentro de la plataforma Web y generar resultados estadísticos, a partir de los datos almacenados de la biorrefinería, mediante el uso de gráficas y/o tablas.

- Realizar una selección de algoritmos de ML para realizar pruebas de clasificación, detección de patrones, etc.; utilizando los datos almacenados de la biorrefinería, para generar tablas de recomendaciones de rendimiento de los procesos de la biorrefinería propuesta.

1.3. Resumen de contribuciones

A continuación, a modo de resumen, se explican cada una de las contribuciones de la tesis.

La biorrefinería envía datos a través de internet hasta llegar al módulo de recepción y procesamiento de datos, el cual se encarga de capturar y almacenar esa información, en una base de datos, para su uso posterior dentro de la plataforma Web. En el módulo de análisis estadístico los datos pasan por diversas fórmulas de cálculo estadístico: promedio, desviación estándar, varianza y covarianza. Los resultados estadísticos son almacenados en la base de datos para su visualización, mediante tablas y/o gráficas, en la plataforma Web.

Los biotecnólogos pueden acceder a la plataforma mediante un navegador Web para realizar: consultas de la información almacenada, cargar nuevos archivos de datos, y visualizar los resultados obtenidos mediante el uso del módulo estadístico, ya sea mediante tablas y/o gráficas, acordes al tipo de consulta realizada.

La plataforma Web cuenta con diversos formularios para realizar cada uno de los cálculos estadísticos, para almacenar nueva información en la base de datos, generar tablas de recomendaciones de rendimiento de la biorrefinería que indican los niveles mínimos y críticos de algún proceso de la misma y, también permiten visualizar diversos tipos de gráficas enseñando el desempeño de los procesos de la biorrefinería.

1.4. Organización de la tesis

En el capítulo 2 se presenta un breve resumen de aprendizaje automático y sus métodos más utilizados para la clasificación de datos y, se estudia el estado del arte de las soluciones de aprendizaje automático actuales para la clasificación de datos con énfasis en el método utilizado en esta tesis.

En el capítulo 3 se muestra con mayor detalle toda la contribución realizada en esta tesis. Desde el desarrollo del Web framework y sus funciones hasta los módulos desarrollados que lo conforman.

En el capítulo 4 se muestran las pruebas y resultados realizados por el producto final de esta tesis. Desde figuras de las diversas gráficas realizadas dentro del entorno Web y de cómo realizarlas, hasta muestras de cómo realizar cada uno de los diversos análisis estadísticos y como generar las recomendaciones de rendimiento correspondientes.

2. Conceptos preliminares

En este capítulo se menciona algunos conceptos preliminares de Machine Learning (ML) que abordan la cuestión de clasificación, esto con el fin de entender la importancia de tener información agrupada y así, lograr generar recomendaciones. De igual manera, se revisa el estado del arte de algunas soluciones de clasificación de datos y se comparan diversos trabajos relacionados con esta tesis en términos de algoritmos de ML.

2.1. Machine Learning

El aprendizaje automático es una rama en evolución de algoritmos computacionales que están diseñados para emular la inteligencia humana al aprender del entorno [10]. Se basa en ideas de diferentes disciplinas, como inteligencia artificial, probabilidad y estadística, informática, teoría de la información, psicología, teoría de control y filosofía [11]. Esta área computacional se ha aplicado en campos tan diversos como el reconocimiento de patrones [12], visión computacional [13], ingeniería de naves espaciales [14], finanzas [15], entretenimiento [16], ecología [17], biología computacional [18] y aplicaciones biomédicas y médicas [19]. La propiedad más importante de estos algoritmos es su capacidad para aprender el entorno a partir de los datos de entrada.

El aprendizaje no es la tarea en sí, sino el medio para lograr la capacidad de realizar una tarea [20]. En el contexto de ML, la experiencia denota un conjunto de datos, donde cada instancia es representada con el mismo conjunto de características [21].

La clasificación en ML ocurre cuando se tiene variables de entrada (X) y una variable de salida (Y) y se utiliza un algoritmo para aprender la función de mapeo de la entrada a la salida $y = f(x)$ [22]. *Machine Learning* está dividido en las siguientes categorías: aprendizaje supervisado, aprendizaje no supervisado, redes neuronales y por refuerzo [23].

Los algoritmos de aprendizaje automático supervisado requieren que los datos utilizados para entrenar el algoritmo ya estén etiquetados. Por ejemplo, un algoritmo de clasificación aprenderá a identificar animales después de ser entrenado con un conjunto de datos de imágenes que están debidamente etiquetadas con la especie del animal y algunas características de identificación [21].

En contraste, los algoritmos de aprendizaje automático no supervisados infieren patrones de un conjunto de datos sin referencia a resultados conocidos o etiquetados. Estos algoritmos, también conocidos como algoritmos de agrupamiento (Clustering), pueden usarse para descubrir nuevas clases de elementos [21]. Los algoritmos de agrupamiento se utilizan comúnmente para formar subconjuntos de datos para encontrar anomalías y similitudes en los datos [24].

2.1.1. K-means Clustering

En esta subsección se hace una revisión del algoritmo de clasificación de aprendizaje no supervisado denominado *k-means. En este trabajo se considera este algoritmno* debido a que no se cuenta con información validada a partir de los datos generados por la biorrefinería propuesta. El algoritmo *k-means*, es un método comúnmente utilizado para particionar automáticamente un conjunto de datos en *k* grupos (cluster) [25].

El algoritmo *K-means* utiliza grupos definidos por centroides del conjunto de datos. Un centroide es un vector que contiene un número de cada variable, donde cada número es la media de una variable para las observaciones en ese grupo [26]. Se considera que un punto está en un grupo particular si está más cerca del centroide de ese grupo que cualquier otro centroide.

En [27] crearon un modelo de clasificación basado en la red neuronal perceptrónica multicapa (MLPNN) utilizando el algoritmo *k-means* como mecanismo de apoyo para la toma de decisiones diagnósticas en el tratamiento de la epilepsia. De igual manera, en [28] proponen una técnica, utilizando el algoritmo de *k-means*, para reconocer movimientos fundamentales del antebrazo humano (extensión, flexión y rotación) y que pueda usarse como herramienta clínica para evaluar el progreso de la rehabilitación en patologías neurodegenerativas.

El algoritmo *k-means* esta denotado de la siguiente manera:

$$j = \sum_{i=1}^{m} \sum_{k=1}^{k} \omega_{ik} \left\| X^i - \mu_k \right\|^2$$

En donde $\omega_{ik} = 1$ para el punto de datos X^i si pertenece al clúster *k*; de lo contrario, $\omega_{ik} = 0$. Además, μ_k es el centroide del grupo de X^i.

La forma en que funciona el algoritmo *k-means* es la siguiente:

- Especifica el número de grupos *k.*
- Inicializa los centroides seleccionando aleatoriamente *k* puntos de datos para los centroides sin reemplazo.
- El algoritmo sigue iterando hasta que no haya cambios en los centroides, es decir, la asignación de puntos de datos a grupos no está cambiando.
- Asigna cada punto de datos al grupo más cercano (centroide).

2.2. Servicio Web

En esta tesis se desarrolló una plataforma Web, desarrollada con el estándar SOAP, la cual puede ser accedida por los biotecnólogos mediante un navegador Web. Los servicios Web son sistemas de software autónomos identificados por un Identificador de Recursos Uniforme (URI, por sus siglas en inglés) que pueden anunciarse, ubicarse y accederse a través de mensajes codificados de acuerdo con estándares basados en XML (por ejemplo, SOAP, WSDL y REST [29]) y transmitidos mediante protocolos de Internet [30]. Los servicios Web encapsulan la funcionalidad de la aplicación y los recursos de información y los hacen disponibles a través de interfaces programáticas, a diferencia de las interfaces proporcionadas por las aplicaciones Web tradicionales que están destinadas a interacciones manuales. Además, dado que están destinados a ser descubiertos y utilizados por otras aplicaciones en la Web, los servicios Web deben describirse y entenderse tanto en términos de capacidades funcionales como de propiedades de calidad de servicio (QoS, por sus siglas en inglés).

2.3. Web framework

La plataforma Web, desarrollada en esta tesis, se construyó utilizando un Web framework denominado *Django*. Un “Framework” para aplicaciones Web se puede considerar como una aplicación genérica incompleta y configurable, con directrices arquitectónicas ofreciendo al desarrollador un conjunto de herramientas para agilizar el proceso de construir una aplicación Web concreta, siempre teniendo en cuenta que es necesario adaptarlo para cada una de las aplicaciones a desarrollarse [31]. Es una estructura que permite la reutilización de sus componentes, los cuales facilitan la creación de estas aplicaciones permitiendo ahorrar tiempo y mantenimiento.

Algunos de los objetivos de usar Web frameworks son:

- Facilita el desarrollo de aplicaciones Web.
- Permite acelerar el proceso de desarrollo de aplicaciones Web.
- Reutiliza código.

Para el desarrollo de software se hace imprescindible el uso de frameworks ya que incluyen bibliotecas, lenguaje, soportes entre otras herramientas la cuales facilitan el desarrollo de aplicaciones Web [32]. La arquitectura que poseen todos los frameworks interactúan bajo el Modelo Vista Controlador (MVC), tal y como se muestra en la figura 3.

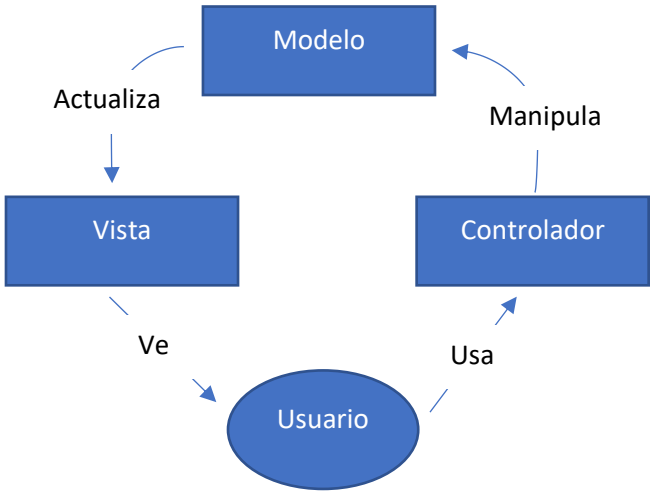


Figura 3.- Modelo Vista Controlador

Modelo

Es aquel que es realizado por el desarrollador y que contiene todos los datos, es decir toda la información y la funcionalidad del programa.

Vista

Es aquella que permite gestionar cómo se presentarán los datos; es decir, cómo interactúa el usuario final con la interfaz, la cual debe ser amigable.

Controlador

Toda la información requerida es enviada al gestor de base de datos para ser guardada; es decir, controla el acceso a los datos y de esta manera el contenido es de forma estática y dinámica.

2.3.1 Django

Es un framework Web de código abierto escrito en Python que permite construir aplicaciones Web más rápido y con menos código [33]. Django es un Web framework, del lenguaje de programación *Python,* que proporciona una solución de alto rendimiento para aplicaciones Web personalizadas y flexibles. Además, la elección de Python como lenguaje de programación en lugar de PHP o Java está motivada por la necesidad de una interfaz simple para interactuar con el sistema del servidor, simplificando la implementación de la comunicación entre el servidor y los lados del clúster.

En [34] crearon una aplicación basada en la Web diseñada para proporcionar una herramienta flexible y fácil de usar para la visualización gráfica de las funciones de distribución de patrones (PDFs). De igual manera, en [35] se habla de un sistema Web que recopila datos cada tres horas de lluvia de la NASA TRMM (Tropical Rainfall Monitoring Mission), los procesa y los compara con series de datos históricos para detectar niveles de alerta.

2.4. Tecnología analítica de procesos

La tecnología analítica de procesos (PAT, por sus siglas en inglés) se ha definido como un sistema para diseñar, analizar y controlar la fabricación a través de mediciones oportunas (es decir, durante el procesamiento) de atributos críticos de calidad y rendimiento de materiales y procesos en bruto y en proceso, con el objetivo de asegurar la calidad del producto final [35]. Un objetivo deseado del marco PAT es diseñar y desarrollar procesos bien entendidos que aseguren constantemente una calidad predefinida al final del proceso de fabricación. Un proceso generalmente se considera bien entendido cuando:

- Se identifican y explican todas las fuentes críticas de variabilidad.
- La variabilidad es gestionada por el proceso.
- Los atributos de calidad del producto pueden predecirse de manera precisa y confiable sobre el espacio de diseño establecido para los materiales utilizados, las condiciones del proceso, la fabricación, el medio ambiente y otras condiciones.

El objetivo para la implementación de PAT puede ser uno o más de los siguientes [35]:

- Mejor comprensión del proceso.
- Rendimiento mejorado debido a la prevención de la chatarra, rechazos y reprocesamiento.
- Reducción del tiempo del ciclo de producción mediante el uso de mediciones y controles en línea.
- Disminución del consumo de energía y eficiencia mejorada de la conversión del proceso por lotes en un proceso continuo.
- Reducción de costos debido a la reducción de desperdicios y consumo de energía.

- Lanzamiento en tiempo real de los lotes.

Desde una perspectiva de implementación, PAT se puede visualizar como un proceso de tres pasos ilustrado en la figura 4 [36].

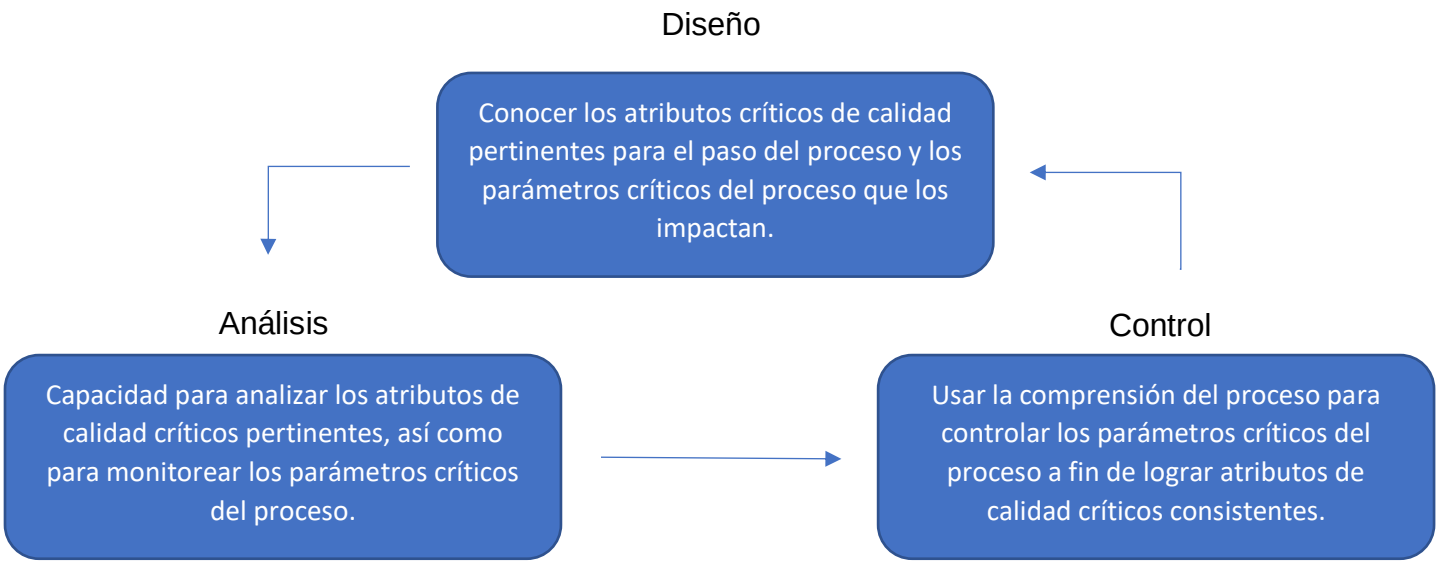


Figura 4.- Los tres pasos que se deben seguir para la implementación de PAT, y los objetivos de cada paso

La fase de diseño comienza en el desarrollo del proceso cuando se diseña la operación de la unidad dada y luego se optimiza y caracteriza [37]. En esta fase, los atributos de calidad críticos (CQA, por sus siglas en inglés) que están siendo afectados por el paso del proceso se identifican junto con los parámetros críticos del proceso (CPP) que se han determinado que afectan el CQA. Esta comprensión del proceso es la esencia de PAT y crítica para las próximas dos fases.

En la fase de análisis, se identifica un analizador adecuado para el monitoreo del CQA y el CPP. La aplicación PAT puede ser “at-line” (muestra extraída, aislada y analizada cerca del flujo del proceso), “on-line” (muestra extraída para análisis del flujo del proceso y devuelta al flujo del proceso nuevamente), “in-line” (muestra no extraída pero analizada en lugar) y “off-line” (muestra extraída y analizada fuera del flujo del proceso) [35, 36, 38]. Para una aplicación PAT, es necesario que los

resultados analíticos estén disponibles en el marco de tiempo necesario para facilitar la toma de decisiones en tiempo real.

Finalmente, la fase de control implica diseñar un esquema de control basado en la comprensión del proceso, de modo que los datos del analizador puedan utilizarse para tomar decisiones de proceso en tiempo real, y se pueda lograr un rendimiento de proceso y calidad de producto consistentes.

En [39] se implementó un sistema PAT en procesos de liofilización para monitorear en tiempo real los puntos finales de los diferentes pasos del proceso de liofilización (congelación, secado primario, secado secundario) y los fenómenos físicos que ocurren durante el proceso.

En [40] se habla sobre un cambio dramático en la producción de alimentos, desde el monitoreo y control inferencial (pH, temperatura, presión, flujo, etc.) hasta la medición de parámetros centrales (concentraciones y perfiles bioquímicos). Debido a esto, implementaron un sistema PAT para controlar los procesos en tiempo real para fabricar productos y materiales dentro de las especificaciones con la ayuda de las tecnologías clave de espectroscopía remota y análisis de datos multivariados.

El sistema desarrollado en esta tesis tiene cierta similitud a un sistema PAT, ya que permite el control de los parámetros y determinar el estado de los procesos de la biorrefinería y así determinar que recomendaciones se deben aplicar y, de igual manera, permite a analizar la información, cercano en tiempo real, para agilizar el proceso de toma de decisiones.

3.- Diseño e Implementación

En este capítulo de se mostrará el diseño e implementación de la interfaz Web y los componentes desarrollados en esta tesis.

3.1.- Diagramas UML: Casos de uso

En las siguientes subsecciones se muestran diversos diagramas UML de casos de uso de la interfaz Web desarrollada en ente trabajo de tesis.

3.1.1.- Carga de datos

La interfaz Web cuenta con una función de carga de datos, en la cual el usuario puede interactuar desde la creación de una base de datos en donde se almacenará su información, cargar archivos de datos de formato CSV y verificar si la carga de datos fue exitosa o no. En la figura 5 se muestra un diagrama UML de casos de uso de la función de carga de datos de la interfaz Web.

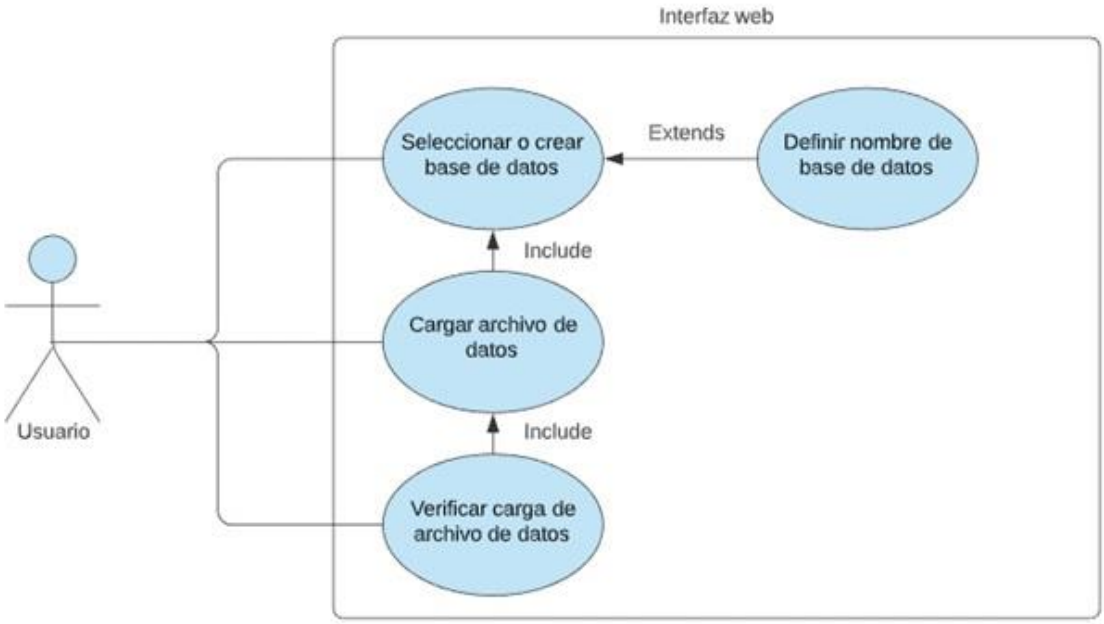


Figura 5.- Diagrama UML de casos de uso: Carga de datos

En las siguientes tablas se muestra la documentación correspondiente del diagrama UML mostrado en la figura 5.

Usuario
Conectores
Uso Desde: Usuario : Actor a: Verificar carga de archivo de datos : Caso de uso
Uso Desde: Usuario : Actor a: Seleccionar o crear base de datos : Caso de uso
Uso Desde: Usuario : Actor a: Cargar archivo de datos : Caso de uso

Tabla 4.- Documentación UML: Usuario

Seleccionar o crear base de datos
Conectores
Extend Desde: Definir nombre de base de datos: Caso de uso a: Seleccionar o crear base de datos : Caso de uso
Uso Desde: Usuario : Actor a: Seleccionar o crear base de datos : Caso de uso
Include Desde: Cargar archivo de datos : Caso de uso a: Seleccionar o crear base de datos : Caso de uso

Tabla 5.- Documentación UML: Seleccionar o crear base de datos

Definir nombre de base de datos

Conectores

Extend
Desde: Definir nombre de base de datos: Caso de uso
a: Seleccionar o crear base de datos : Caso de uso

Tabla 6.- Documentación UML: Definir nombre de base de datos

Cargar archivo de datos

Conectores

Include
Desde: Verificar carga de archivo de datos: Caso de uso
a: Cargar archivo de datos : Caso de uso

Uso
Desde: Usuario : Actor
a: Cargar archivo de datos : Caso de uso

Include
Desde: Cargar archivo de datos : Caso de uso
a: Seleccionar o crear base de datos : Caso de uso

Tabla 7.- Documentación UML: Cargar archivo de datos

Verificar carga de archivo de datos

Conectores

Include
Desde: Verificar carga de archivo de datos: Caso de uso
a: Cargar archivo de datos: Caso de uso

Uso
Desde: Usuario : Actor
a: Verificar carga de archivo de datos: Caso de uso

Tabla 8.- Documentación UML: Verificar carga de archivo de datos

En la subsección 3.5 de este capítulo se detalla el funcionamiento de la carga de datos dentro de la interfaz Web de este trabajo de tesis.

3.1.2.- Cálculo estadístico

La interfaz Web también cuenta con un módulo de cálculo estadístico en donde el usuario puede realizar diversos cálculos (promedio, varianza, covarianza, desviación estándar, prueba de Fisher, análisis de varianza y clustering) utilizando los datos previamente cargados. Para realizar cálculos estadísticos primeramente el usuario debe seleccionar la base de datos en donde se almacenó su información previamente cargada, posteriormente el usuario debe seleccionar los parámetros necesarios que se muestran en la interfaz Web para generar los resultados correspondientes. En la figura 6 se muestra un diagrama UML de casos de uso de la función de cálculo estadístico de la interfaz Web.

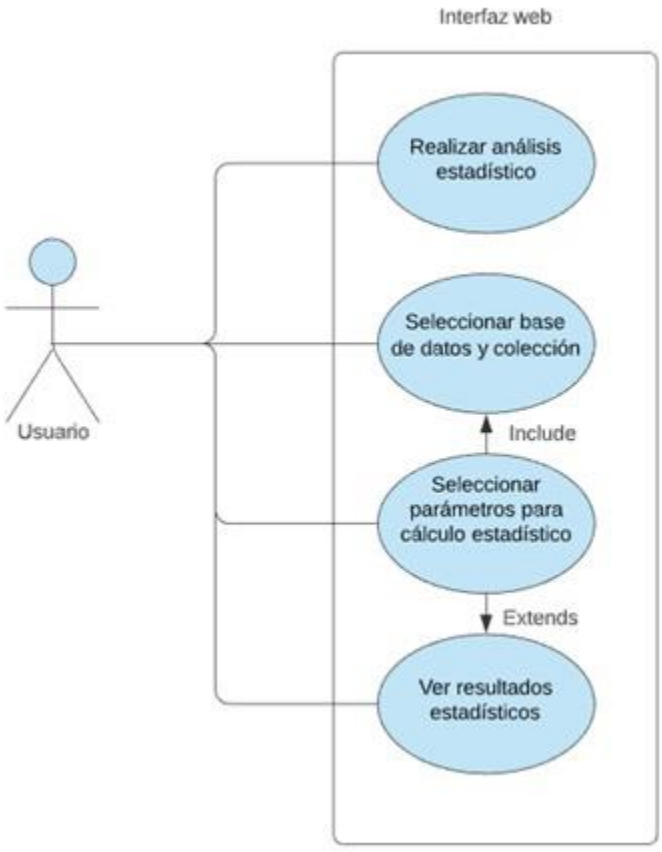


Figura 6.- Diagrama UML de casos de uso: Cálculo estadístico

En las siguientes tablas se muestra la documentación correspondiente del diagrama UML mostrado en la figura 6.

Usuario
Conectores
Uso Desde: Usuario : Actor a: Realizar análisis estadístico : Caso de uso
Uso Desde: Usuario : Actor a: Seleccionar base de datos y colección : Caso de uso
Uso Desde: Usuario : Actor a: Seleccionar parámetros para cálculo estadístico : Caso de uso
Uso Desde: Usuario : Actor a: Ver resultados estadísticos : Caso de uso

Tabla 9.- Documentación UML: Usuario

Realizar análisis estadístico
Conectores
Uso Desde: Usuario : Actor a: Realizar análisis estadístico : Caso de uso

Tabla 10.- Documentación UML: Realizar análisis estadístico

Seleccionar base de datos y colección

Conectores

Include
Desde: Seleccionar parámetros para cálculo estadístico: Caso de uso
a: Seleccionar base de datos y colección: Caso de uso

Uso
Desde: Usuario : Actor
a: Seleccionar base de datos y colección: Caso de uso

Tabla 11.- Documentación UML: Seleccionar base de datos y colección

Seleccionar parámetros para cálculo estadístico

Conectores

Include
Desde: Seleccionar parámetros para cálculo estadístico: Caso de uso
a: Seleccionar base de datos y colección: Caso de uso

Uso
Desde: Usuario : Actor
a: Seleccionar parámetros para cálculo estadístico: Caso de uso

Extend
Desde: Seleccionar parámetros para cálculo estadístico: Caso de uso
a: Ver resultados estadísticos: Caso de uso

Tabla 12.- Documentación UML: Seleccionar parámetros para cálculo estadístico

Ver resultados estadísticos

Conectores

Extend
Desde: Seleccionar parámetros para cálculo estadístico: Caso de uso
a: Ver resultados estadísticos: Caso de uso

Uso
Desde: Usuario : Actor
a: Ver resultados estadísticos: Caso de uso

Tabla 13.- Documentación UML: Ver resultados estadísticos

En la subsección 3.6 de este capítulo se detalla el uso del módulo de cálculo estadístico de la interfaz Web de este trabajo de tesis.

3.1.3.- Gráficas

Otra de las funciones de la interfaz Web es que el usuario puede generar gráficas utilizando los datos almacenados y los resultados estadísticos. El usuario debe llenar los formularios correspondientes en la interfaz Web para poder generar las gráficas correspondientes. En la figura 7 se muestra un diagrama UML de casos de uso de como generar gráficas en la interfaz Web.

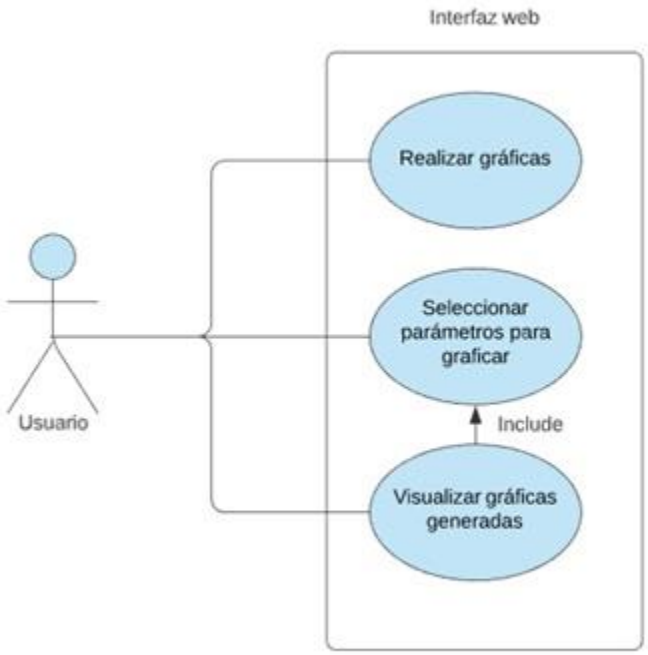


Figura 7.- Diagrama UML de casos de uso: Gráficas

En las siguientes tablas se muestra la documentación correspondiente del diagrama UML mostrado en la figura 6.

Usuario

Conectores

Uso
Desde: Usuario : Actor
a: Realizar gráficas : Caso de uso

Uso
Desde: Usuario : Actor
a: Seleccionar parámetros para graficar : Caso de uso

Uso
Desde: Usuario : Actor
a: Visualizar gráficas generadas : Caso de uso

Tabla 14.- Documentación UML: Usuario

Realizar gráficas

Conectores

Uso
Desde: Usuario : Actor
a: Realizar gráficas: Caso de uso

Tabla 15.- Documentación UML: Realizar gráficas

Seleccionar parámetros para graficar

Conectores

Include
Desde: Visualizar gráficas generadas: Caso de uso
a: Seleccionar parámetros para graficar: Caso de uso

Uso
Desde: Usuario : Actor
a: Seleccionar parámetros para graficar: Caso de uso

Tabla 16.- Documentación UML: Seleccionar parámetros para graficar

Visualizar gráficas generadas
Conectores
Include Desde: Visualizar gráficas generadas: Caso de uso a: Seleccionar parámetros para graficar: Caso de uso
Uso Desde: Usuario : Actor a: Visualizar gráficas generadas: Caso de uso

Tabla 17.- Documentación UML: Visualizar gráficas generadas

En el capítulo 4 de este trabajo de tesis se muestran diversas pruebas y resultados de la interfaz Web, incluyendo diversas gráficas generadas a partir de diversas combinaciones de parámetros.

3.2.- Diagramas UML: Actividad

En las siguientes subsecciones se muestran diversos diagramas UML de actividad de la interfaz Web.

3.2.1.- Carga de datos

En la figura 8 se muestra un diagrama UML que muestra las diversas actividades que se pueden realizar en la interfaz Web en cuanto a carga de datos. Antes de iniciar una carga de datos el usuario debe seleccionar si quiere utilizar una base de datos ya existente o crear una nueva para almacenar sus datos. Cuando el usuario ha hecho su elección, debe llenar los formularios correspondientes para completar la carga de datos. En la subsección 3.5 de este capítulo, se encuentran los pasos a seguir, de manera detallada, de la carga de datos en la interfaz Web.

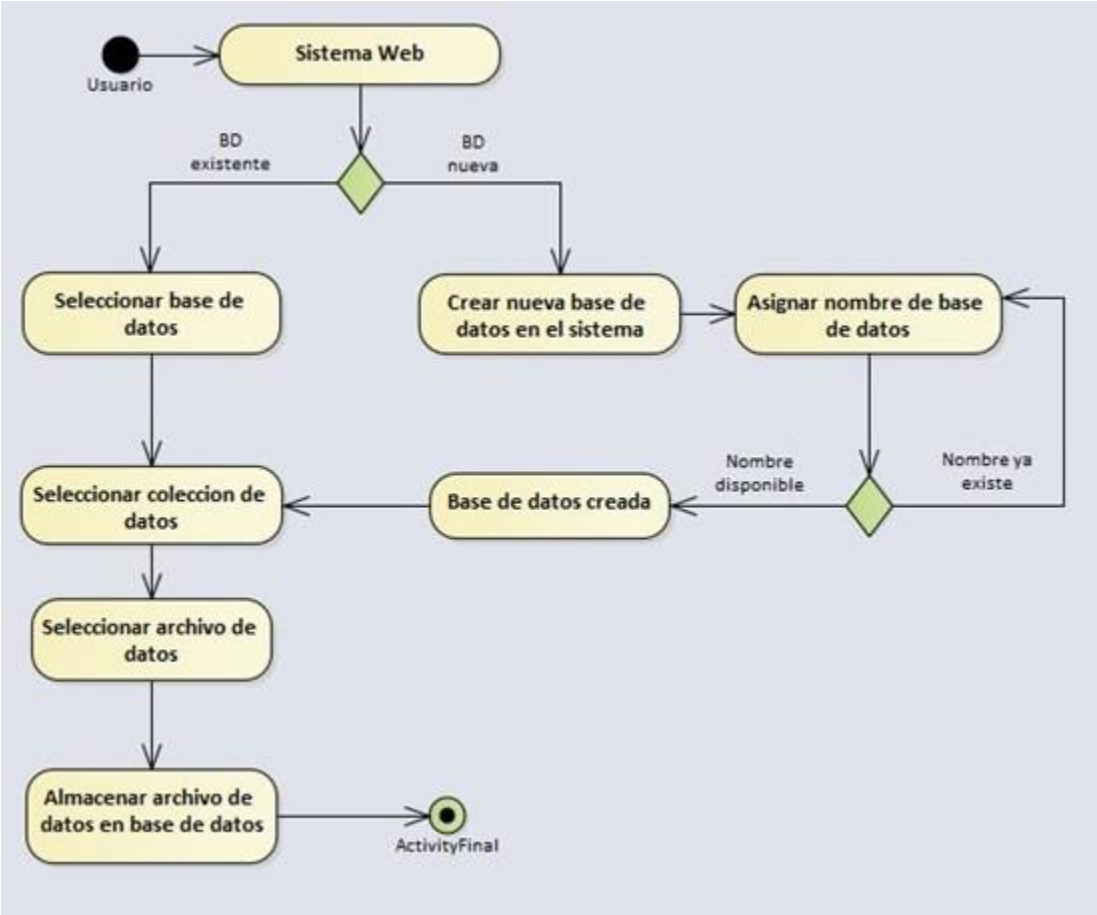


Figura 8.- Diagrama UML de Actividad: Carga de datos

En la tabla 18 se muestra la documentación correspondiente del diagrama UML mostrado en la figura 8.

Usuario
Relaciones de comportamiento salientes
Flujo desde Usuario a Sistema Web
Sistema Web
Relaciones de comportamiento salientes
Flujo desde Sistema Web a Símbolo de decisión 1

Tabla 18.- Parte 1 Documentación UML: Carga de datos

Seleccionar base de datos
Relaciones de comportamiento salientes
Flujo desde Seleccionar base de datos a Seleccionar colección de datos
Relaciones de comportamiento entrantes
Flujo desde Símbolo de decisión 1 (BD existente) a Seleccionar base de datos
Crear nueva base de datos
Relaciones de comportamiento salientes
Flujo desde Crear nueva base de datos a Asignar nombre de base de datos
Relaciones de comportamiento entrantes
Flujo desde Símbolo de decisión 1 (BD nueva) a Crear nueva base de datos
Asignar nombre de base de datos
Relaciones de comportamiento salientes
Flujo desde Asignar nombre de base de datos a Símbolo de decisión 2
Relaciones de comportamiento entrantes
Flujo desde Símbolo de decisión 2 (Nombre ya existente) a Asignar nombre de base de datos
Base de datos creada
Relaciones de comportamiento salientes
Flujo desde Base de datos creada a Seleccionar colección de datos
Relaciones de comportamiento entrantes
Flujo desde Símbolo de decisión 2 (Nombre disponible) a Base de datos creada

Tabla 18.- Parte 2 Documentación UML: Carga de datos

Seleccionar colección de datos
Relaciones de comportamiento salientes
Flujo desde Seleccionar colección de datos a Seleccionar archivo de datos
Relaciones de comportamiento entrantes
Flujo desde Seleccionar base de datos a Seleccionar colección de datos Flujo desde Base de datos creada a Seleccionar colección de datos
Seleccionar archivo de datos
Relaciones de comportamiento salientes
Flujo desde Seleccionar archivo de datos a Almacenar archivo de datos en base de datos
Relaciones de comportamiento entrantes
Flujo desde Seleccionar colección de datos a Seleccionar archivo de datos
Almacenar archivo de datos en base de datos
Relaciones de comportamiento salientes
Flujo desde Almacenar archivo de datos en base de datos a ActivityFinal
Relaciones de comportamiento entrantes
Flujo desde Seleccionar archivo de datos a Almacenar archivo de datos en base de datos
ActivityFinal
Relaciones de comportamiento entrantes
Flujo desde Almacenar archivo de datos en base de datos a ActivityFinal

Tabla 18.- Parte 3 Documentación UML: Carga de datos

3.2.2.- Análisis estadístico y gráficas

En la figura 9 se muestra un diagrama UML de actividad para realizar análisis estadísticos y generar gráficas en la interfaz Web. Primeramente, el usuario selecciona los datos que va a utilizar, para realizar análisis estadísticos, de la base de datos. Posteriormente, el usuario debe elegir si únicamente desea generar datos

estadísticos o generar gráficas con los mismos resultados estadísticos. Una vez seleccionada la opción deseada, la interfaz brinda al usuario los formularios correspondientes para concluir con su análisis y, así, generar los resultados estadísticos. En el capítulo 4 de pruebas y resultados se muestran unos ejemplos de resultados estadísticos y gráficas generadas a partir de los mismos.

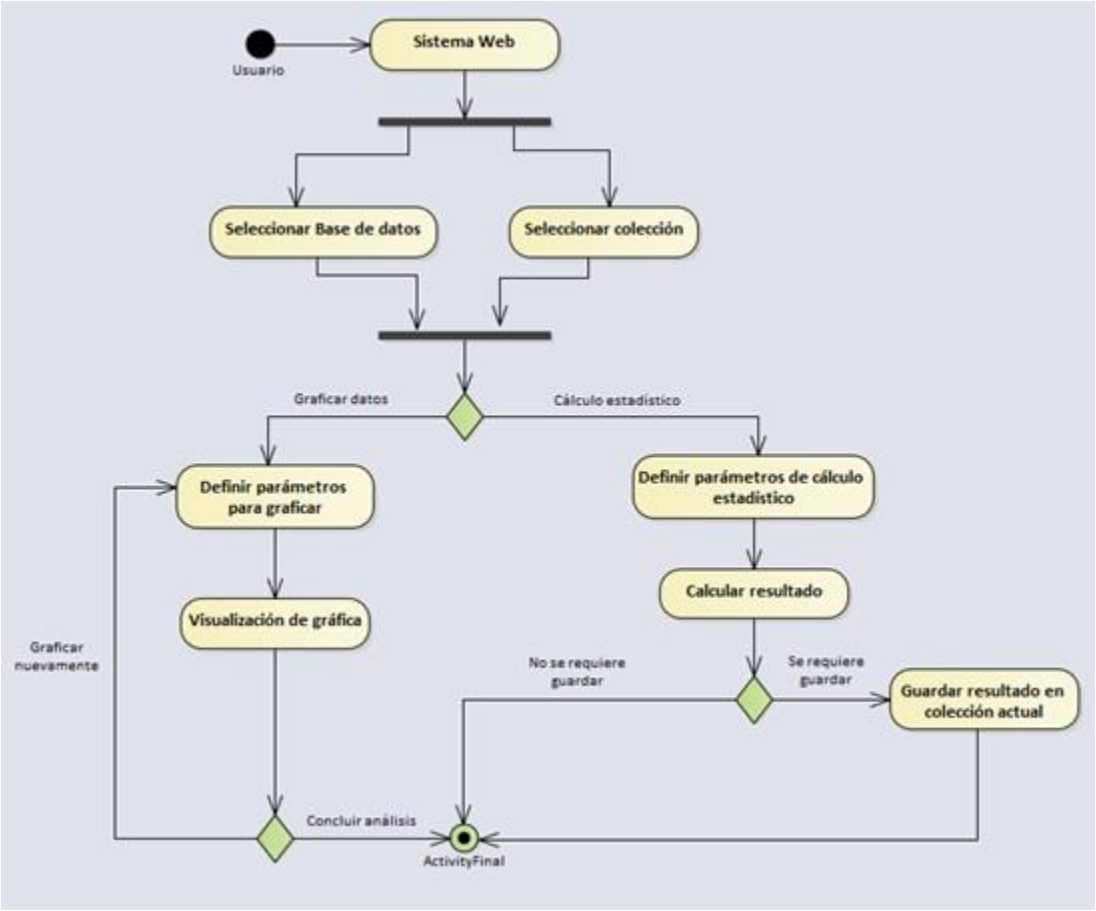


Figura 9.- Diagrama UML de Actividad: Análisis estadístico y gráficas

En la tabla 19 se muestra la documentación correspondiente del diagrama UML mostrado en la figura 9.

Usuario

Relaciones de comportamiento salientes

- Flujo desde Usuario a Sistema Web

Sistema Web

Relaciones de comportamiento salientes

- Flujo desde Sistema Web a Símbolo de bifurcación

Seleccionar base de datos

Relaciones de comportamiento salientes

- Flujo desde Seleccionar base de datos a Símbolo de bifurcación

Relaciones de comportamiento entrantes

- Flujo desde Símbolo de bifurcación a Seleccionar base de datos

Seleccionar colección

Relaciones de comportamiento salientes

- Flujo desde Seleccionar colección a Símbolo de bifurcación

Relaciones de comportamiento entrantes

- Flujo desde Símbolo de bifurcación a Seleccionar colección

Definir parámetros para graficar

Relaciones de comportamiento salientes

- Flujo desde Definir parámetros para graficar a Visualización de gráfica

Relaciones de comportamiento entrantes

- Flujo desde Símbolo de decisión 1 (Graficar datos) a Definir parámetros para graficar
- Flujo desde Símbolo de decisión 2 (Graficar nuevamente) a Definir parámetros para graficar

Tabla 19.- Parte 1 Documentación UML: Análisis estadístico y gráficas

Visualización de gráfica

Relaciones de comportamiento salientes

- Flujo desde Visualización de gráfica a Símbolo de decisión 2

Relaciones de comportamiento entrantes

- Flujo desde Definir parámetros para graficar a Visualización de gráfica

Definir parámetros de cálculo estadístico

Relaciones de comportamiento salientes

- Flujo desde Definir parámetros de cálculo estadístico a Calcular resultado

Relaciones de comportamiento entrantes

- Flujo desde Símbolo de decisión 1 (Cálculo estadístico) a Definir parámetros de cálculo estadístico

Calcular resultado

Relaciones de comportamiento salientes

- Flujo desde Calcular resultado a Símbolo de decisión 3

Relaciones de comportamiento entrantes

- Flujo desde Definir parámetros de cálculo estadístico a Calcular resultado

Guardar resultado en colección actual

Relaciones de comportamiento salientes

- Flujo desde Guardar resultado en colección actual a ActivityFinal

Relaciones de comportamiento entrantes

- Flujo desde Símbolo de decisión 3 (Se requiere guardar) a Guardar resultado en colección actual

Tabla 19.- Parte 2 Documentación UML: Análisis estadístico y gráficas

ActivityFinal
Relaciones de comportamiento entrantes
Flujo desde Símbolo de decisión 2 (Concluir análisis) a ActivityFinal
Flujo desde Símbolo de decisión 3 (No se requiere guardar) a ActivityFinal
Flujo desde Guardar resultado en colección actual a ActivityFinal

Tabla 19.- Parte 3 Documentación UML: Análisis estadístico y gráficas

3.3.- Diagramas UML: Secuencia

En las siguientes subsecciones se muestran diversos diagramas UML de actividad de la interfaz Web.

3.3.1.- Carga de datos

La figura 10 muestra un diagrama UML de secuencia que muestra, de manera detallada, cómo se van comunicando los módulos de la interfaz Web al realizar diversas consultas. Cuando el usuario va a realizar una carga de datos, la interfaz Web realiza diversas consultar para realizar el proceso de carga de datos. La interfaz se comunica con el módulo de recepción y almacenamiento para realizar una carga de datos y devuelve una respuesta. De igual manera, el módulo de recepción y almacenamiento se comunica con la base de datos para realizar la petición de carga de datos y recibir una respuesta, la cual será enviada al usuario para confirmar que su carga de datos ha sido exitosa.

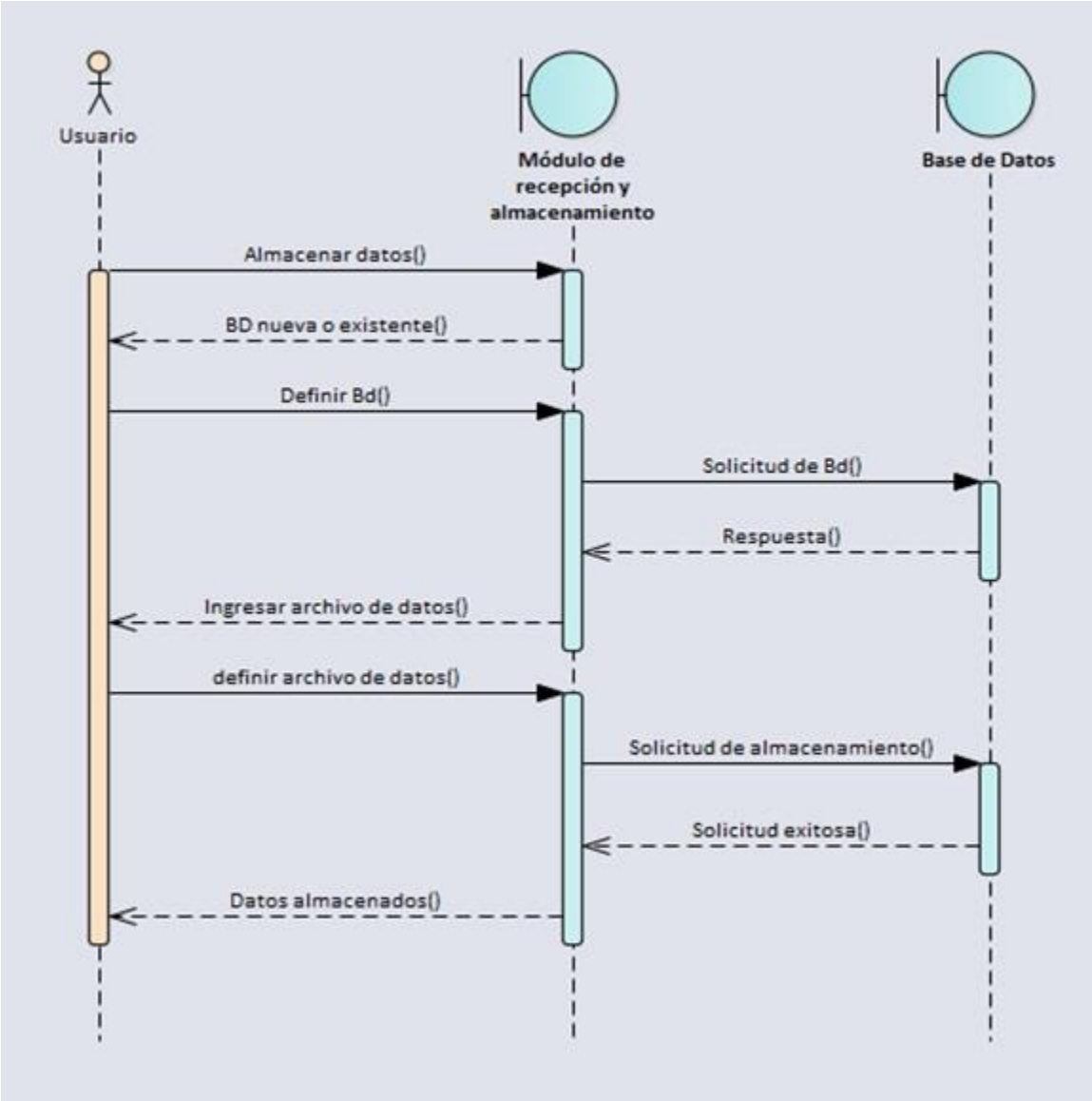


Figura 10.- Diagrama UML de secuencia: Carga de datos

En la tabla 20 se muestra la documentación correspondiente del diagrama UML mostrado en la figura 10.

Usuario

Relaciones de comportamiento salientes

Nombre: Almacenar datos()
Secuencia desde Usuario a Módulo de recepción y almacenamiento

Nombre: Definir Bd()
Secuencia desde Usuario a Módulo de recepción y almacenamiento

Nombre: Definir archivo de datos()
Secuencia desde Usuario a Módulo de recepción y almacenamiento

Relaciones de comportamiento entrantes

Nombre: Bd nueva o existente()
Secuencia desde Módulo de recepción y almacenamiento a Usuario

Nombre: Ingresar archivo de datos()
Secuencia desde Módulo de recepción y almacenamiento a Usuario

Nombre: Datos almacenados()
Secuencia desde Módulo de recepción y almacenamiento a Usuario

Tabla 20.- Parte 1 Documentación UML: Carga de datos

Mensajes de interacción

1.0 ‘Almacenar datos() ' desde ' Usuario ' enviado a ' Módulo de recepción y almacenamiento'.

Synchronous Call. Returns void.

[Return is False. Iteration is False. New group is False.]

1.1 ‘Bd nueva o existente() ' desde ' Módulo de recepción y almacenamiento ' enviado a ' Usuario '.

Synchronous Call. Returns void.

[Return is False. Iteration is False. New group is False.]

✉ 1.2 ‘Definir Bd() ' desde ' Usuario ' enviado a ' Módulo de recepción y almacenamiento'. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]
✉ 1.3 ‘Solicitud de Bd() ' desde ' Módulo de recepción y almacenamiento ' enviado a ' Base de datos'. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]
✉ 1.4 ‘Respuesta() ' desde ' Base de datos ' enviado a ' Módulo de recepción y almacenamiento'. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]
✉ 1.5 ‘Ingresar archivo de datos() ' desde ' Módulo de recepción y almacenamiento ' enviado a ' Usuario '. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]
✉ 1.6 ‘Definir archivo de datos() ' desde ' Usuario ' enviado a ' Módulo de recepción y almacenamiento'. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]
✉ 1.7 ‘Solicitud de almacenamiento() ' desde ' Módulo de recepción y almacenamiento ' enviado a ' Base de datos. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]
✉ 1.8 ‘Solicitud exitosa() ' desde ' Base de datos ' enviado a ' Módulo de recepción y almacenamiento'. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]
✉ 1.9 ‘Datos almacenados() ' desde ' Módulo de recepción y almacenamiento ' enviado a ' Usuario’. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]

Tabla 20.- Parte 2 Documentación UML: Carga de datos

3.3.2.- Cálculo estadístico

En la figura 11 se muestra un diagrama UML de secuencia de las diversas consultas que se ejecutan para realizar cálculos estadísticos en la interfaz Web. El módulo de recepción y almacenamiento se comunica con la base de datos y con el módulo de análisis estadístico para brindarle al usuario los resultados correspondientes. Las consultas que realizan los módulos son desde selección de bases de datos hasta la selección de parámetros para realizar análisis estadísticos, y posteriormente, recibir una respuesta la cual será entregada al usuario mediante la interfaz Web.

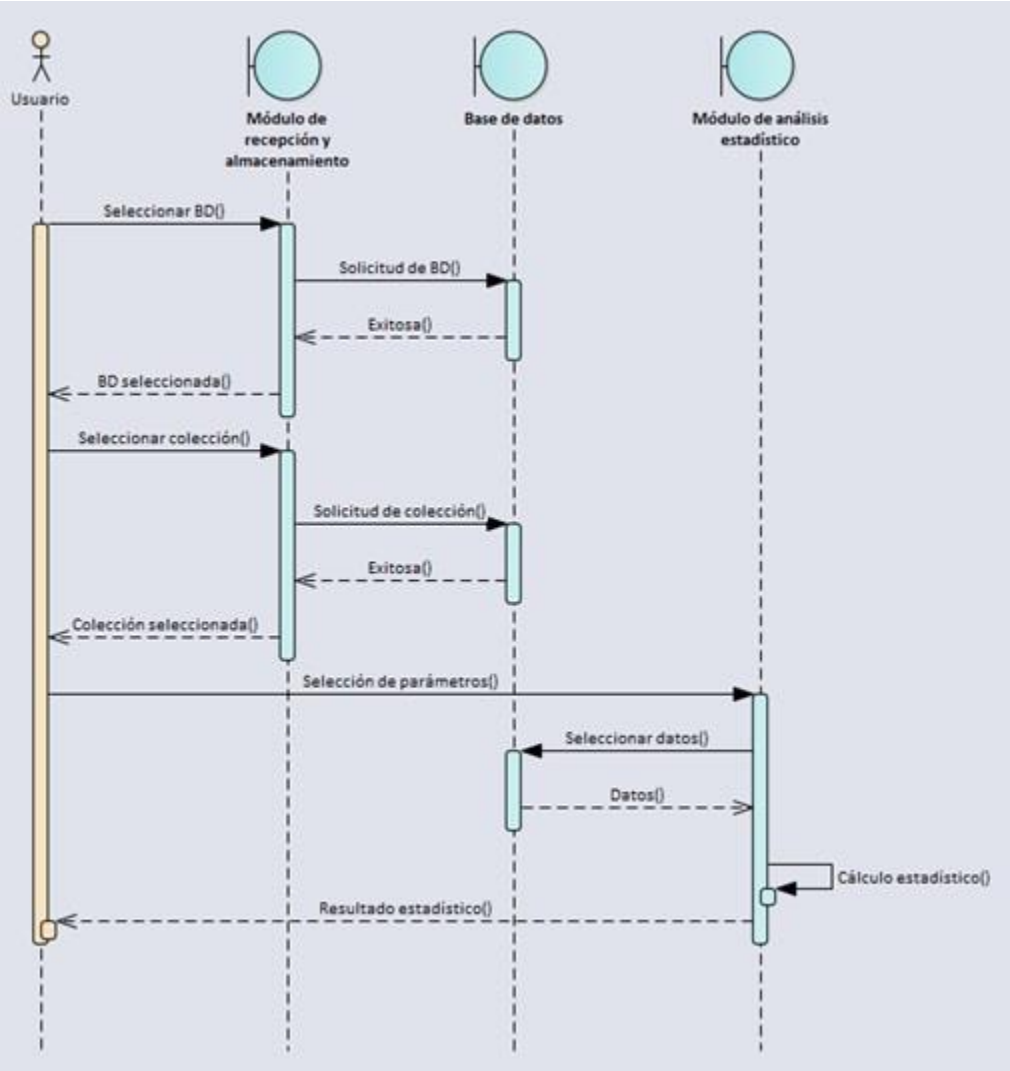


Figura 11.- Diagrama UML de secuencia: Cálculo estadístico

En la tabla 21 se muestra la documentación correspondiente del diagrama UML mostrado en la figura 11.

Usuario
Relaciones de comportamiento salientes
Nombre: Seleccionar Bd() Secuencia desde Usuario a Módulo de recepción y almacenamiento
Nombre: Seleccionar colección() Secuencia desde Usuario a Módulo de recepción y almacenamiento
Nombre: Selección de parámetros() Secuencia desde Usuario a Módulo de análisis estadístico
Relaciones de comportamiento entrantes
Nombre: Bd seleccionada() Secuencia desde Módulo de recepción y almacenamiento a Usuario
Nombre: Colección seleccionada() Secuencia desde Módulo de recepción y almacenamiento a Usuario
Nombre: Resultados estadísticos() Secuencia desde Módulo de análisis estadísticos a Usuario

Tabla 21.- Parte 1 Documentación UML: Cálculo estadístico

Mensajes de interacción
1.0 ‘Seleccionar Bd() ' desde ' Usuario ' enviado a ' Módulo de recepción y almacenamiento'. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]
1.1 ‘Solicitud de Bd() ' desde ' Módulo de recepción y almacenamiento ' enviado a ' Base de datos '. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]

✉ 1.2 ‘Exitosa() ' desde ' Base de datos ' enviado a ' Módulo de recepción y almacenamiento'.

Synchronous Call. Returns void.

[Return is False. Iteration is False. New group is False.]

✉ 1.3 ‘ Bd seleccionada() ' desde ' Módulo de recepción y almacenamiento ' enviado a ' Usuario '.

Synchronous Call. Returns void.

[Return is False. Iteration is False. New group is False.]

✉ 1.4 ‘Seleccionar colección() ' desde ' Usuario ' enviado a ' Módulo de recepción y almacenamiento'.

Synchronous Call. Returns void.

[Return is False. Iteration is False. New group is False.]

✉ 1.5 ‘Solicitud de colección() ' desde ' Módulo de recepción y almacenamiento ' enviado a ' Base de datos '.

Synchronous Call. Returns void.

[Return is False. Iteration is False. New group is False.]

✉ 1.6 ‘Exitosa() ' desde ' Base de datos ' enviado a ' Módulo de recepción y almacenamiento'.

Synchronous Call. Returns void.

[Return is False. Iteration is False. New group is False.]

✉ 1.7 ‘Colección seleccionada() ' desde ' Módulo de recepción y almacenamiento ' enviado a ' Usuario ’.

Synchronous Call. Returns void.

[Return is False. Iteration is False. New group is False.]

✉ 1.8 ‘Selección de parámetros() ' desde ' Usuario ' enviado a ' Módulo de análisis estadístico'.

Synchronous Call. Returns void.

[Return is False. Iteration is False. New group is False.]

✉ 1.9 ‘Seleccionar datos() ' desde ' Módulo de análisis estadístico ' enviado a ' Base de datos ’.

Synchronous Call. Returns void.

[Return is False. Iteration is False. New group is False.]

1.10 ‘Datos() ' desde ' Base de datos ' enviado a ' Módulo de análisis estadístico ’. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]
1.11 ‘Cálculo estadístico() ' desde ' Módulo de análisis estadístico ' enviado a ' Módulo de análisis estadístico’. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]
1.9 ‘Resultado estadístico() ' desde ' Módulo de análisis estadístico ' enviado a ' Usuario ’. Synchronous Call. Returns void. [Return is False. Iteration is False. New group is False.]

Tabla 21.- Parte 2 Documentación UML: Cálculo estadístico

3.4.- Interfaz Web

Antes de comenzar con la recepción de datos de la biorrefinería, primero se implementó una interfaz Web que será la encargada de recibir esa información. Para este trabajo de tesis se decidió trabajar con el Web framework Django. Django es un framework Web de código abierto escrito en Python que permite construir aplicaciones Web más rápido, con menos código y respeta el MVC [34]. En la figura 12 se muestra la interfaz Web que se implementó para este trabajo de tesis.

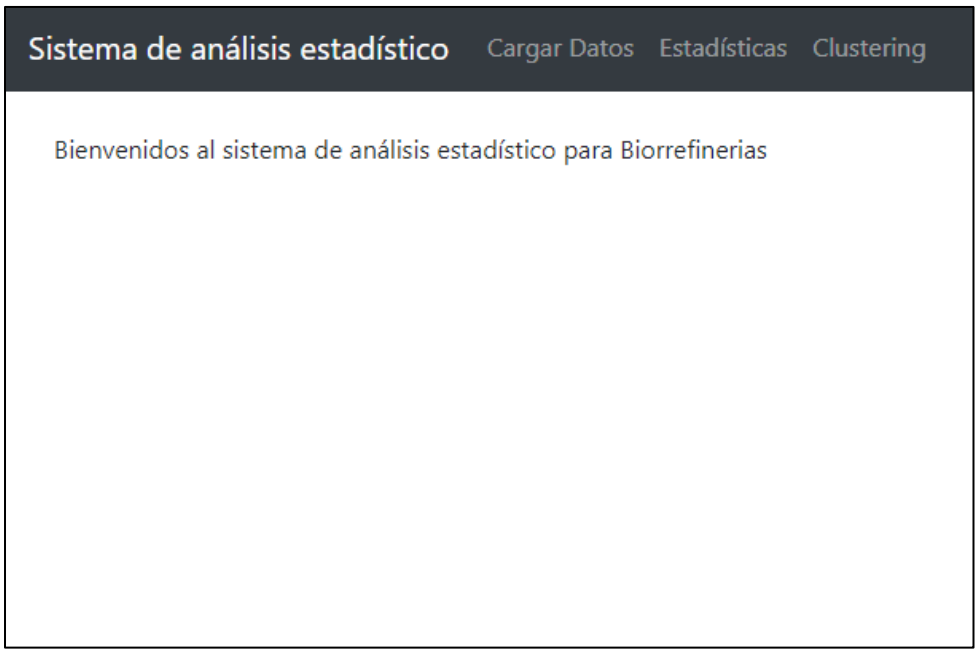


Figura 12.- Interfaz Web Django

La interfaz Web cuenta con diversas funciones las cuales son:

- **Cargar datos:** Este apartado le brinda al usuario la opción de cargar un archivo de datos de valores separados por coma (.CSV) y a su vez, se almacenan en la base de datos. En la subsección 3.5 se explica con más detalle el almacenamiento de la información.

- **Estadísticas:** Con los datos ya almacenados en la base de datos, este apartado le brinda al usuario diversos formularios para realizar análisis estadísticos (promedio, varianza, covarianza, desviación estándar). En la subsección 3.6 se muestran los diversos formularios y sus funciones correspondientes.

- **Clustering:** En este apartado el usuario puede aplicarle un algoritmo de Machine Learning de Clustering a sus datos y, el mismo sistema le arrojará al usuario diversas gráficas de agrupamiento de sus datos y diversas

recomendaciones de rendimiento de la biorrefinería. En la subsección 3.7 se explica con más detalle el funcionamiento de este algoritmo.

3.5.- Módulo de recepción y almacenamiento de datos

La interfaz Web le da la opción al usuario de cargar un archivo de datos en formato CSV y, almacenar la información en la base de datos. En la figura 13 se muestra la interfaz inicial para empezar a cargar datos.

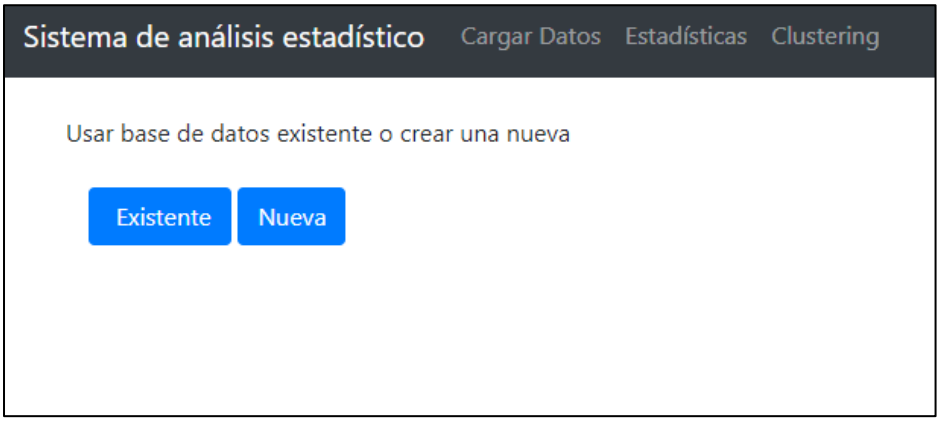


Figura 13.- Interfaz de carga de datos

Antes de cargar el archivo de datos, la interfaz le da la opción al usuario si quiere crear una base de datos nueva o utilizar una ya existente. Para la base de datos se utilizó MongoDB. MongoDB es un sistema de base de datos NoSQL que almacena datos como documentos similares a JSON con esquemas dinámicos (el formato se llama BSON) y tiene un enfoque orientado a cuatro cosas: flexibilidad, potencia, velocidad y facilidad de uso [42]. En la tabla 4 se muestra el formato que se estableció en esta tesis para el almacenamiento de los datos.

Formato de almacenamiento	
_id	Identificador de documentos
Data	Arreglo de datos de valores numéricos
Date	Arreglo de datos de valores de fechas
Name	Nombre del documento

Tabla 22.- Formato de almacenamiento de información

El módulo se encarga de transformar los datos del usuario en el formato establecido que se muestra en la tabla 22. Para cumplir con los objetivos propuestos, se decidió este formato ya que facilita a los algoritmos realizar las consultas más rápido a la base de datos y ofrecer una respuesta casi de inmediata.

Volviendo al almacenamiento de la información, en la figura 14 se muestra la interfaz de selección de una base de datos ya existente.

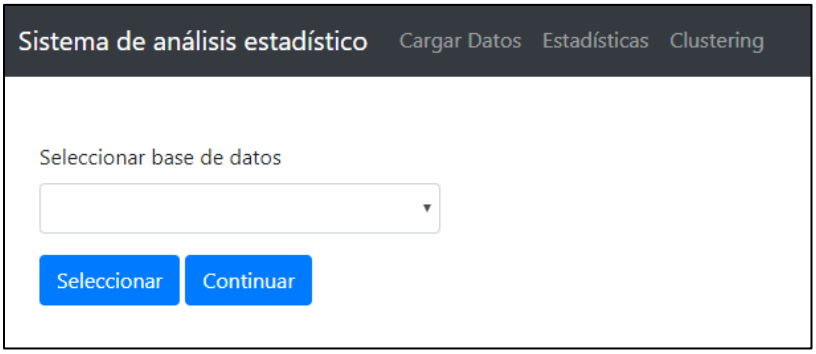


Figura 14.- Interfaz de selección de base de datos existente

Cuando el usuario escoge la opción de almacenar los datos en una base de datos ya existente, el módulo de recepción y almacenamiento hace una consulta a las bases de datos existentes y devuelve los resultados a la interfaz Web para que el usuario seleccione una. En la figura 15 se muestra un ejemplo de las bases de datos creadas para que el usuario seleccione una.

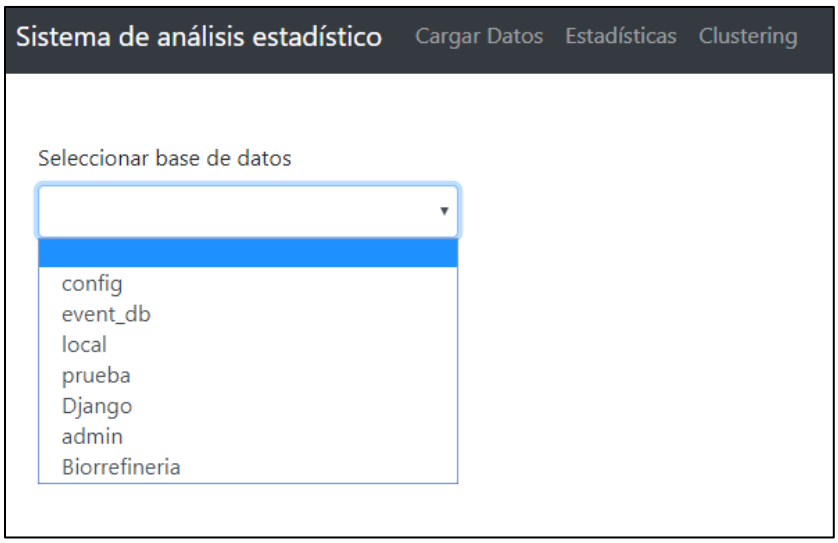


Figura 15.- Ejemplo de bases de datos existentes

En caso de que el usuario requiera crear una base de datos nueva, la interfaz le brinda al usuario el formulario que se muestra en la figura 16.

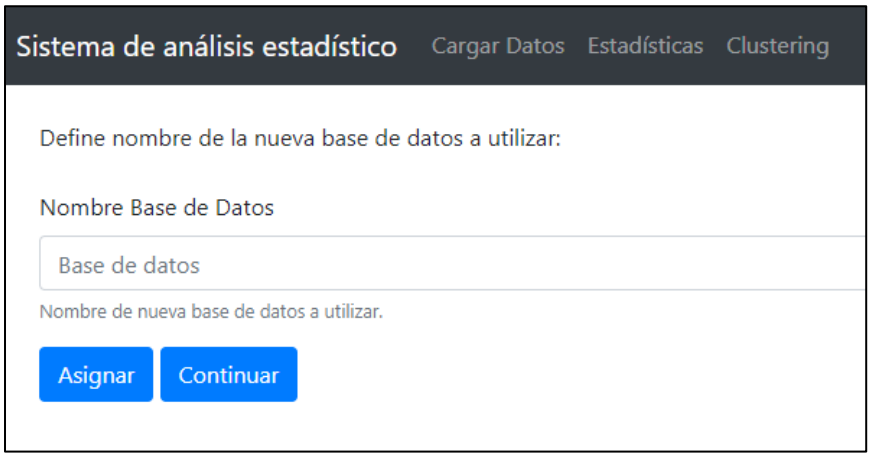


Figura 16.- Interfaz de creación de nueva base de datos

En esta interfaz, el usuario tiene la opción de asignarle un nombre a su base de datos, la cual será utilizada para almacenar la información de su archivo de datos. En caso de que el nombre de la base de datos esté disponible, aparecerá un mensaje diciendo que el nombre está disponible, tal y como se muestra en la figura 17.

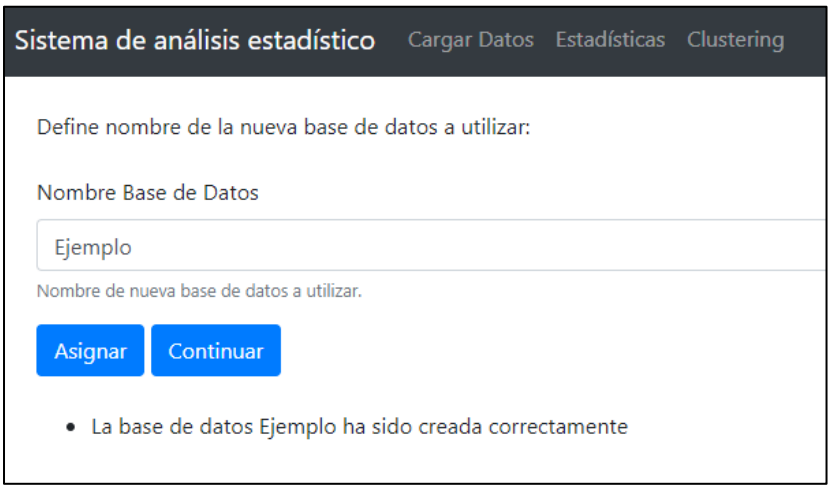


Figura 17.- Mensaje de nombre aceptado de base de datos

En caso de que el nombre ya esté ocupado, apareceré un mensaje advirtiendo que esa base de datos ya está creada. Un ejemplo se muestra en la figura 18.

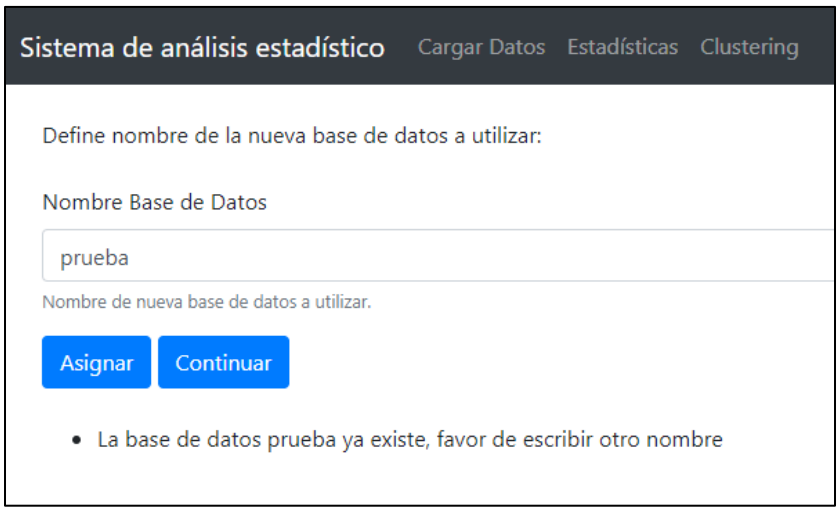


Figura 18.- Mensaje de advertencia de base de datos ya existente

Una vez seleccionado una base de datos existente o creado una nueva, el módulo de recepción y almacenamiento muestra una interfaz para cargar el archivo de datos, tal y como se muestra en la figura 19.

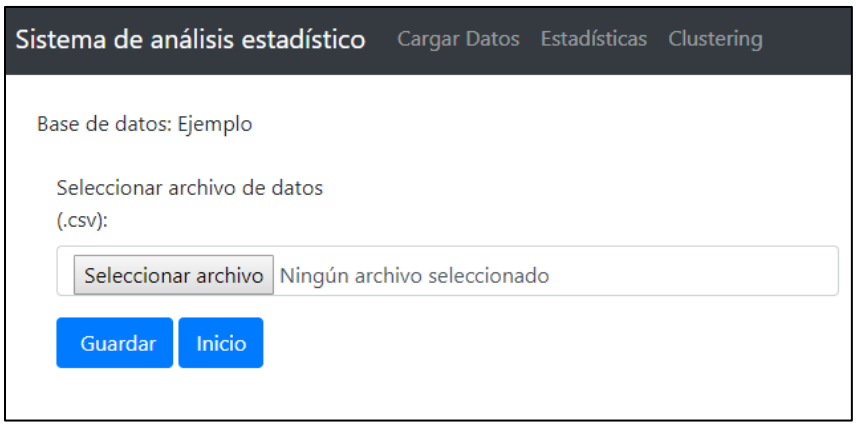


Figura 19.- Interfaz de carga de archivo de datos

Cuando se presiona el botón de “Guardar” para continuar con el almacenado del archivo de datos, aparecerá un mensaje diciendo que el archivo se ha subido correctamente en la base de datos seleccionada. En la figura 20 se muestra un ejemplo del mensaje de almacenamiento exitoso de un archivo de datos llamado “Biomasa”.

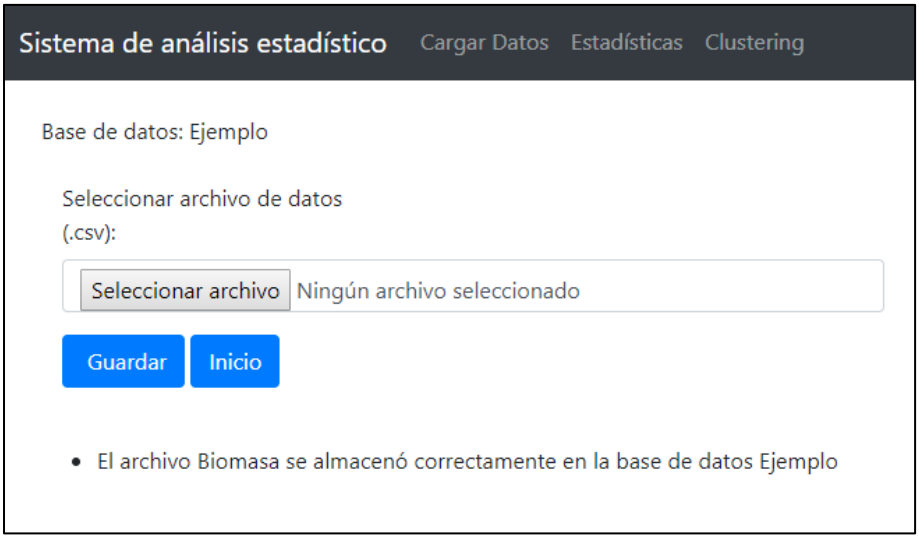


Figura 20.- Mensaje de almacenamiento exitoso de un archivo de datos

3.6.- Módulo de análisis estadístico

Siguiendo con los objetivos planteados, para esta tesis se ha desarrollado un módulo de análisis estadístico básico que se encarga de generar promedios,

desviación estándar, varianza y covarianza de los datos almacenados en la base de datos.

El promedio es un valor central calculado entre un conjunto de números y está definido por la siguiente formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x2 + \cdots + x_n}{n}$$

La desviación estándar es la raíz cuadrada de la varianza. Es una de las medidas de dispersión y es una medida que es indicativa de como los valores individuales pueden diferir de la media y, está definida por la siguiente formula:

$$\sigma = \sqrt{\frac{\sum|x - \mu|^2}{N}}$$

La varianza es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media y, está definida por la siguiente formula:

$$\text{Var(x)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

La covarianza es un valor que indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias y, está definida por la siguiente formula:

$$\text{Cov(X,Y)} = \frac{\sum_{1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

En la figura 21 se muestra la interfaz principal para comenzar a realizar cálculos estadísticos utilizando los datos que están almacenados en la base de datos.

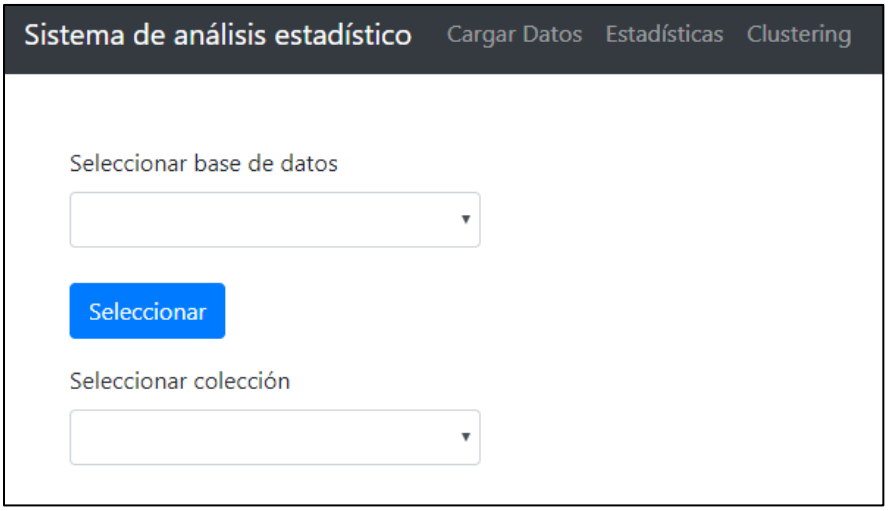


Figura 21.- Interfaz inicial de cálculo estadístico

Como primer paso, en esta interfaz el usuario debe especificar la base de datos y el nombre de la colección en donde están guardados los datos que se van a utilizar para realizar los cálculos estadísticos. Una vez seleccionados, se hace clic en el botón de “Continuar”, tal y como se muestra en la figura 22, para pasar a otra interfaz de cálculo estadístico.

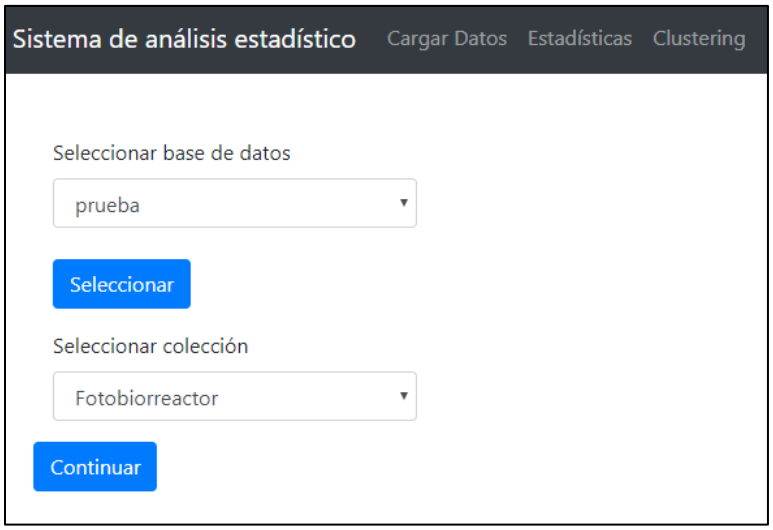


Figura 22.- Interfaz con elementos seleccionados para continuar con el análisis estadístico

En la figura 23 se muestra la nueva interfaz para continuar con el análisis estadístico. En esta ventana, se muestran cuatro botones: Visualización de datos en días, Análisis estadístico en rango de días, prueba de Fisher y ANOVA.

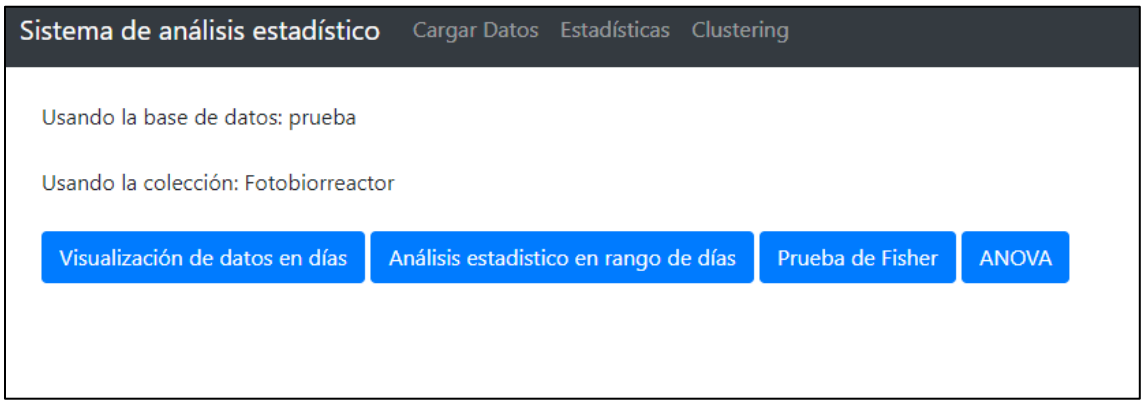


Figura 23.- Interfaz para continuar con análisis estadístico

Las funciones de cada uno de estos botones se mostrarán en las subsecciones siguientes.

3.6.1.- Visualización de datos en días

Al seleccionar la opción de visualización de datos en días se muestra un pequeño formulario, como se muestra en la figura 24, en donde el usuario debe especificar cuantas variables desea utilizar para realizar el cálculo estadístico.

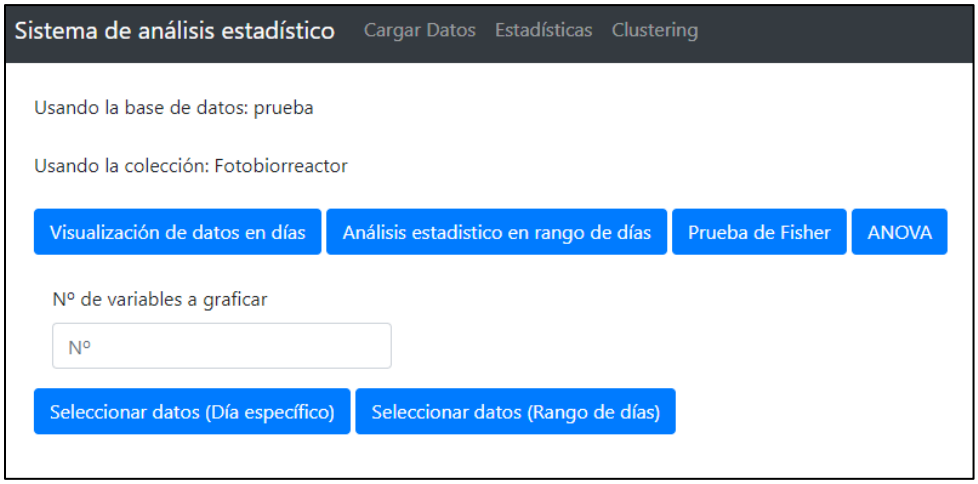


Figura 24.- Interfaz de visualización de datos en días (selección de número de variables)

Una vez que el usuario ha especificado el número de variables a usar durante el análisis estadístico, debe seleccionar entre dos opciones, si quiere utilizar los datos de un día en específico (botón izquierdo) o utilizar los datos de un rango de días (botón derecho). En caso de seleccionar los datos de un día especifico, en la figura 25 se muestra el formulario correspondiente que el usuario debe llenar para que el sistema le genere los resultados correspondientes.

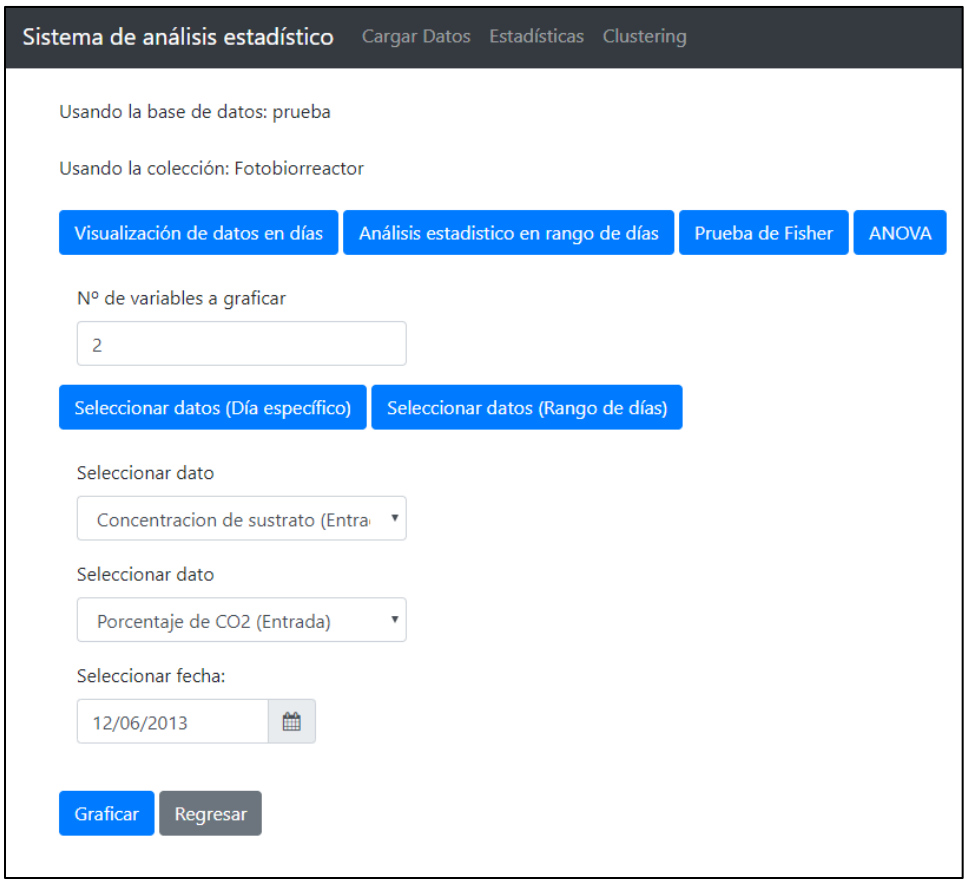


Figura 25.- Formulario de análisis estadístico de datos en un día específico

En otro caso, cuando se selecciona la opción de seleccionar datos en un rango de días se muestra otro formulario, como se muestra en la figura 26, para generar los resultados correspondientes a las selecciones del usuario.

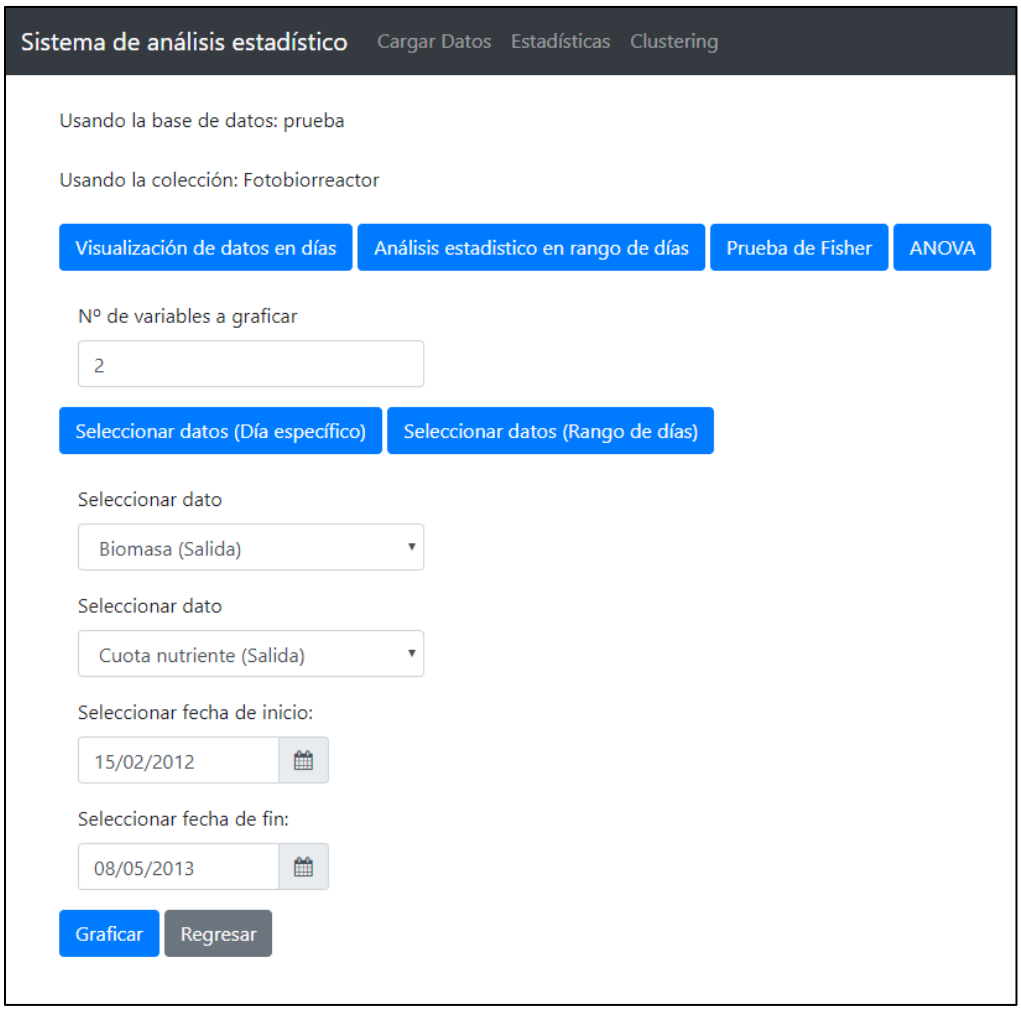


Figura 26.- Formulario de análisis estadístico en rango de días

3.6.2.- Análisis estadístico en rango de días

Al igual que la visualización de datos en días, la opción de análisis estadístico en rango de días inicia con un pequeño formulario que el usuario debe llenar especificando el número de variables a utilizar para el análisis estadístico. Tal y como se muestra en la figura 27.

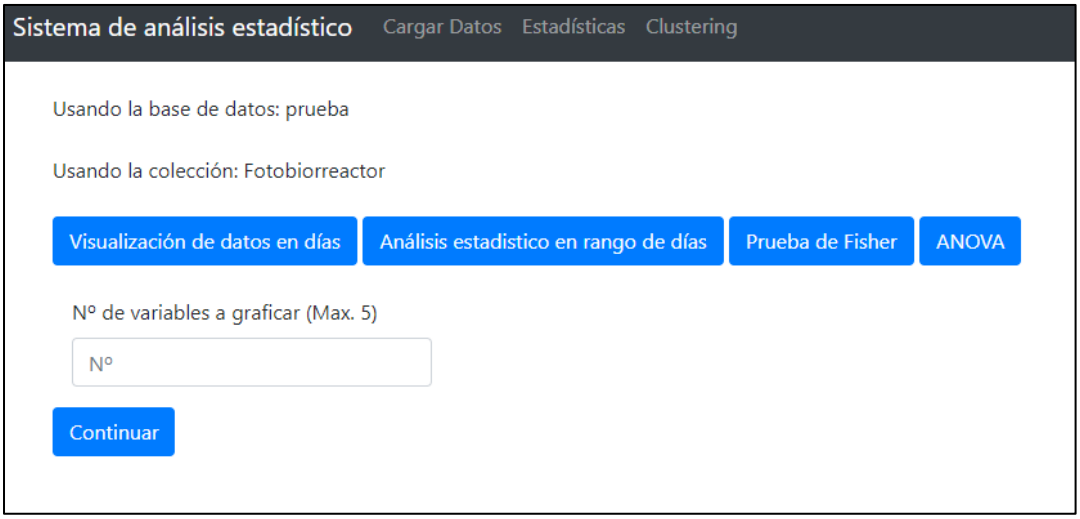


Figura 27.- Interfaz inicial de análisis estadístico en rango de días

Una vez que el usuario ha especificado el número de variables a utilizar y presionado el botón de continuar, el sistema lo redirigirá a un nuevo formulario como se muestra en la figura 28.

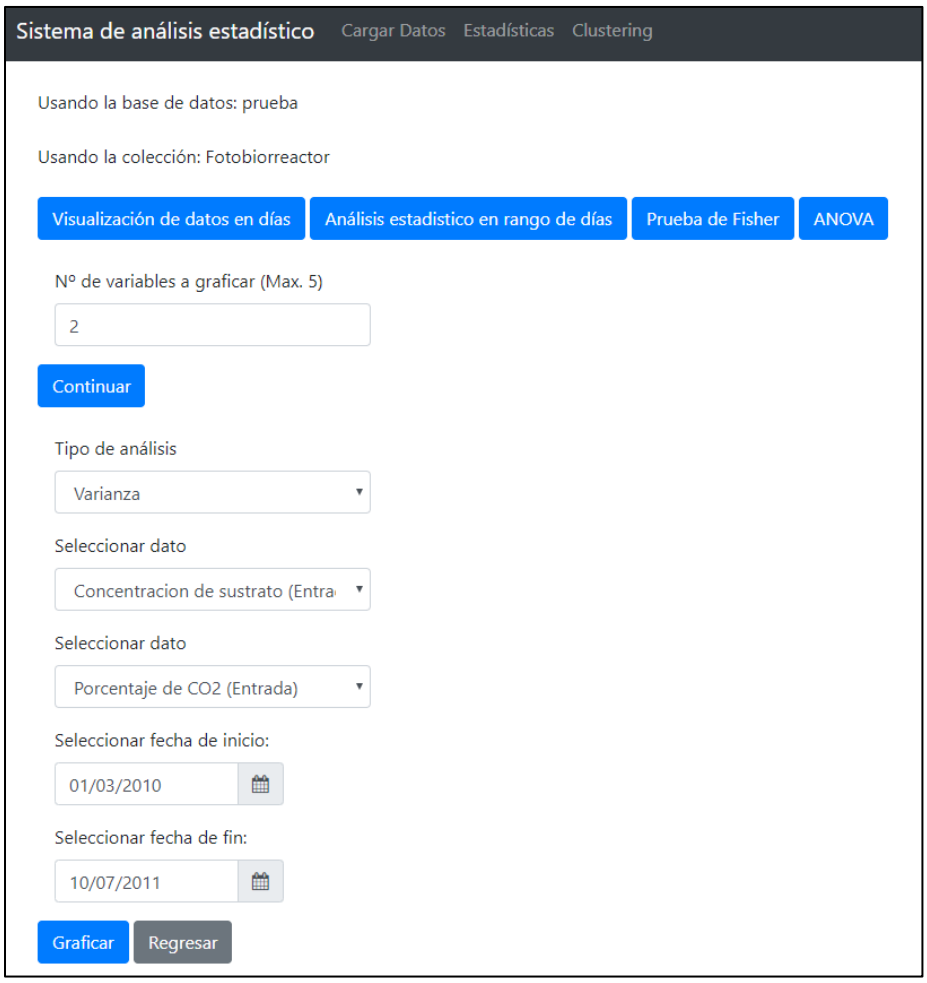


Figura 28.- Formulario de análisis estadístico en rango de días

En este formulario el usuario debe especificar qué tipo de análisis se va a aplicar a los datos (promedio, varianza, covarianza o desviación estándar), que datos se van a utilizar para el análisis y especificar el rango de fechas en donde se tomaran los datos para realizar en análisis correspondiente.

3.6.3.- Prueba de Fisher

La prueba de Fisher fue propuesta por Ronald Aylmer Fisher [41] en la quinta edición de métodos estadísticos para investigadores. Es una prueba de independencia en lugar de asociación en tablas de contingencia de 2 x 2 y está definida por la siguiente formula:

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

Cuando el usuario selecciona la opción de realizar la prueba de Fisher con sus datos, le aparece el siguiente formulario mostrado en la figura 29.

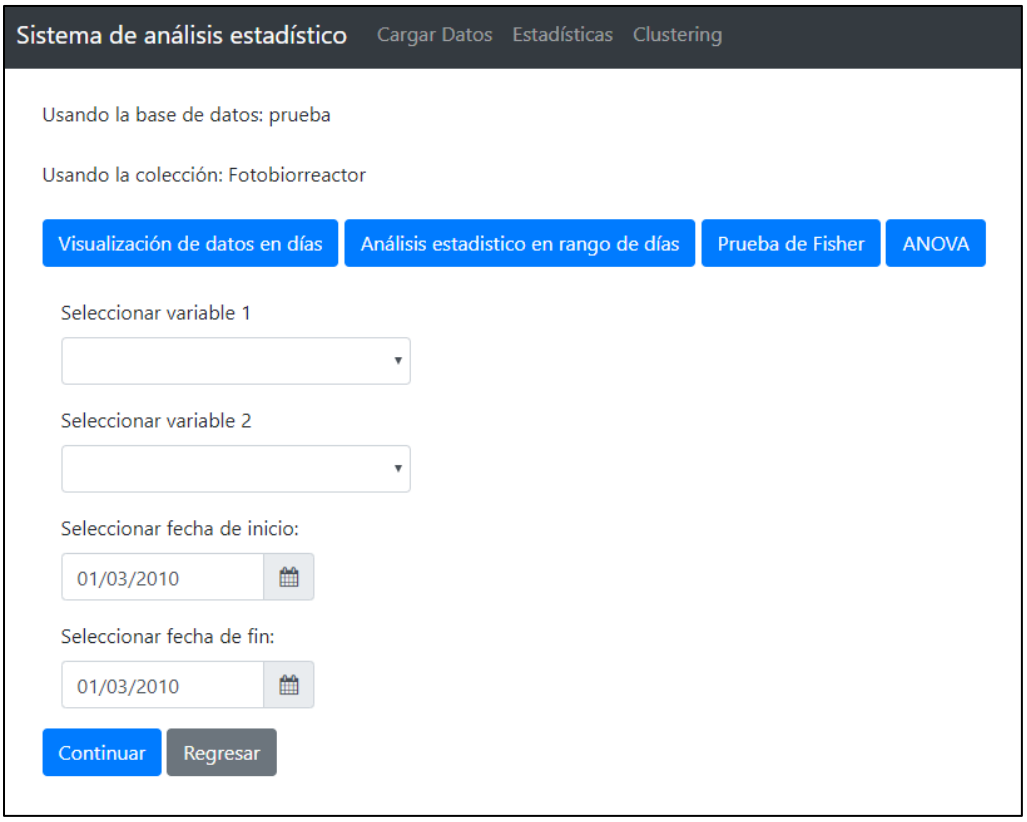


Figura 29.- Formulario de prueba de Fisher

En este formulario el usuario debe especificar 2 variables para crear la tabla de contingencia requerida para realizar la prueba. De igual manera, el usuario debe seleccionar un rango de días en el cual el algoritmo va a tomar los datos para realizar el cálculo y arrojar las respuestas correspondientes.

3.6.4.- Análisis de varianza (ANOVA)

El análisis de varianza se basa en el enfoque en el que el procedimiento utiliza varianzas para determinar si las medias son diferentes. El procedimiento funciona comparando la varianza entre las medias de grupo (entre-grupos) contra la varianza dentro de los grupos (intra-sujetos) como una forma de determinar si los grupos son más distintos entre sí que dentro de sí. El análisis de varianza está definido por la siguiente formula:

$$\sum_i \sum_j (y_{ij} - \overline{y})^2 = n \sum_i (y_i - \overline{y})^2 + \sum_i \sum_j (y_{ij} - y_i)^2$$

Esta ecuación se reescribe frecuentemente como:

$$SS_{total} = SS_{fact} + SS_{int}$$

Donde:

- **SS_{fact}** es un número real relacionado con la varianza, que mide la variación debida al factor o tipo de situación estudiado
- **SS_{int}** es un número real relacionado con la varianza, que mide la variación dentro de cada factor o tipo de situación

Cuando el usuario selecciona la opción de cálculo de ANOVA, el sistema le arroja el siguiente formulario como se muestra en la figura 30.

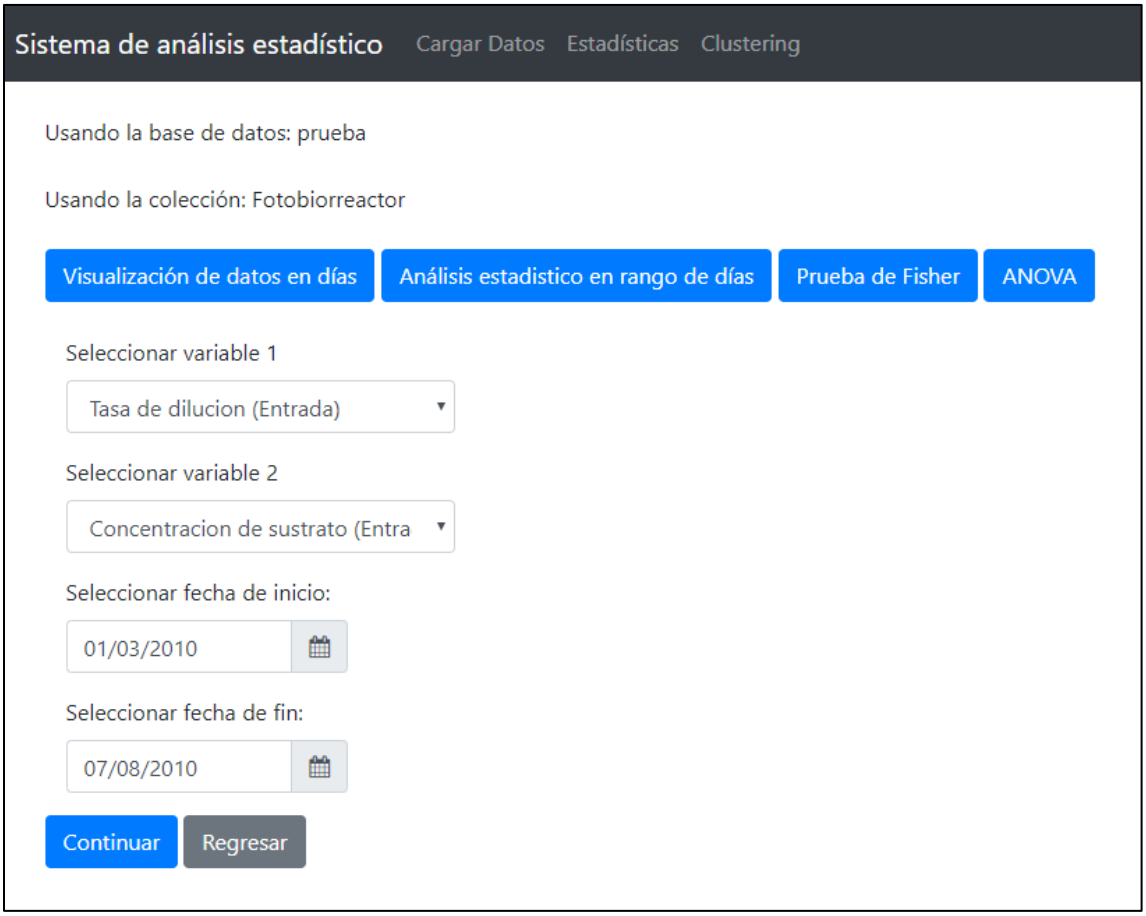


Figura 30.- Formulario de cálculo de ANOVA

Al igual que el formulario de la prueba de Fisher, el formulario de cálculo de ANOVA pide que se seleccione las variables correspondientes para realizar el cálculo y, de igual manera, establecer un rango de fechas en el cual el algoritmo tomará los datos para continuar con el cálculo de ANOVA.

3.6.5.- Clustering

En esta tesis se utilizó el algoritmo de clustering llamado k-means [43] el cual es un método utilizado para particionar automáticamente un conjunto de datos en *k* grupos. K-means implementa el uso de grupos definidos por *k* centroides del conjunto de datos. Se considera que un punto está en un grupo particular si está más cerca del centroide de ese grupo que cualquier otro centroide [44].

Cuando el usuario selecciona la opción de realizar el cálculo de clustering en la interfaz Web, se le muestra una ventana de selección de base de datos, como se muestra en la figura 31, como paso inicial para realizar el cálculo correspondiente.

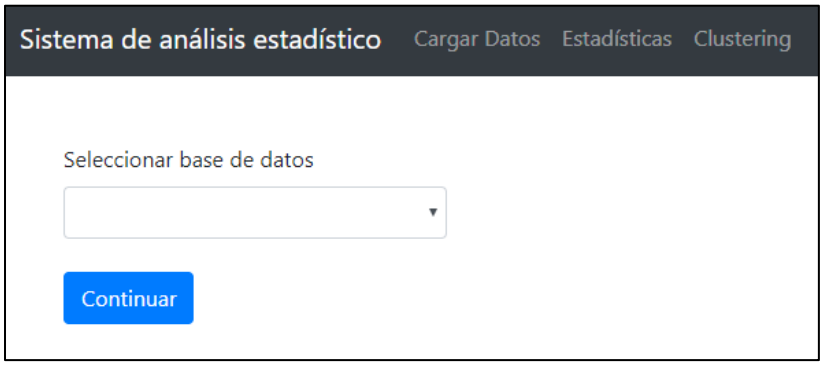


Figura 31.- Ventana inicial de cálculo de clustering

Una vez que el usuario ha seleccionado la base de datos en donde se encuentra la información que va a usar para el algoritmo de clustering, le aparecerá un nuevo formulario como se muestra en la figura 32.

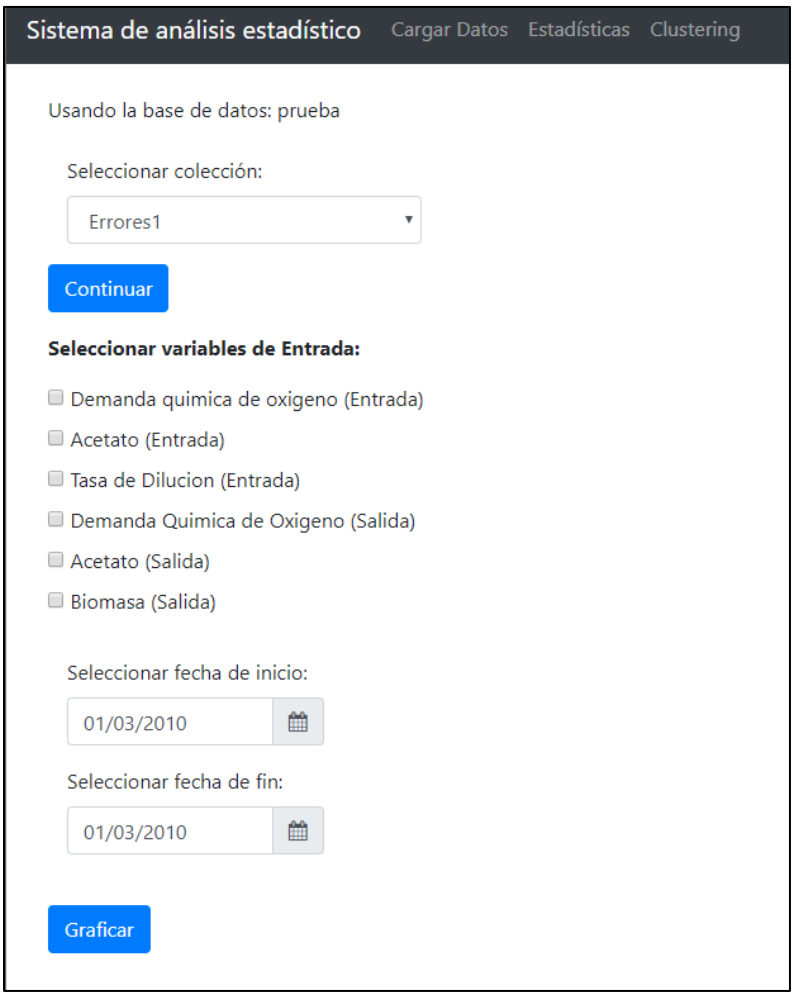


Figura 32.- Formulario para realizar clustering

En este trabajo de tesis, el algoritmo de k-means agrupa los datos de "entrada" con los datos de "Salida". Esto con la finalidad de analizar las diversas similitudes que tienen las entradas con respecto a sus salidas en un periodo de tiempo definido. Para ello, el usuario debe seleccionar la colección en donde están guardados sus datos. Posteriormente, el sistema acomoda todas las variables encontradas dentro de la colección seleccionada y le pide al usuario que seleccione las que son de "Entrada" y que defina un rango de fechas para realizar el cálculo de k-means. Las pruebas y resultados de este algoritmo se encuentran en el capítulo 4 de este trabajo de tesis.

4.- Pruebas y resultados

En este capítulo, se hace una revisión a los resultados obtenidos de utilizar los diversos formularios ilustrados en la sección pasada, desde la carga de datos y su almacenamiento, hasta la obtención de los cálculos estadísticos y clustering.

4.1.- Carga de datos

Para esta prueba se utilizó un archivo de datos CSV que contiene información generada por un simulador de Matlab de la biorrefinería. Primeramente, seleccionaremos la base de datos “prueba” para realizar el ejemplo, posteriormente seleccionamos el o los archivos CSV para almacenar en la base de datos. Para este ejemplo, se utilizó un archivo CSV con nombre “Bacterias anodofilicas” y se almacenó en la base de datos prueba, el resultado de esto se muestra en la figura 33.

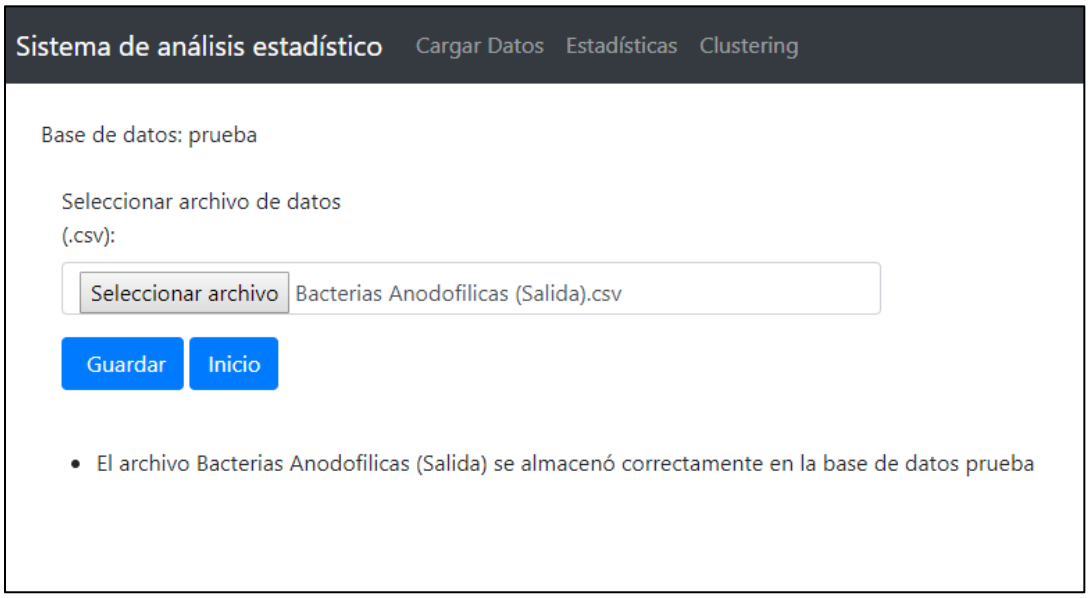


Figura 33.- Carga de archivo CSV en base de datos "prueba"

En la figura 34 se visualiza la una parte de la interfaz de *MongoDB Compass* como muestra de que los datos fueron almacenados correctamente en el formato establecido que se mencionó en la subsección 3.1.

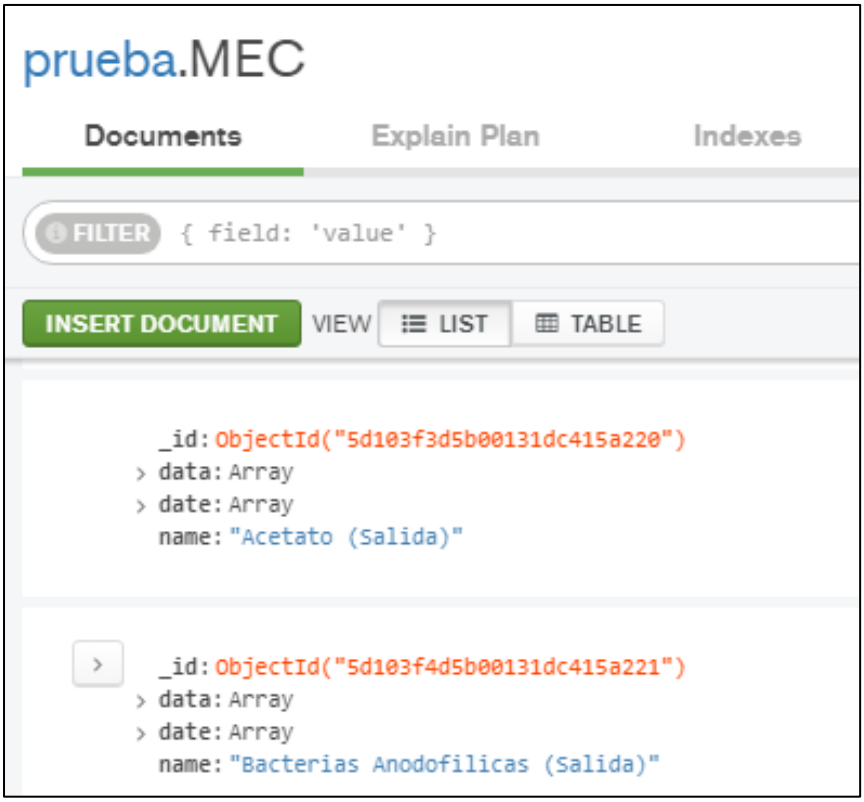


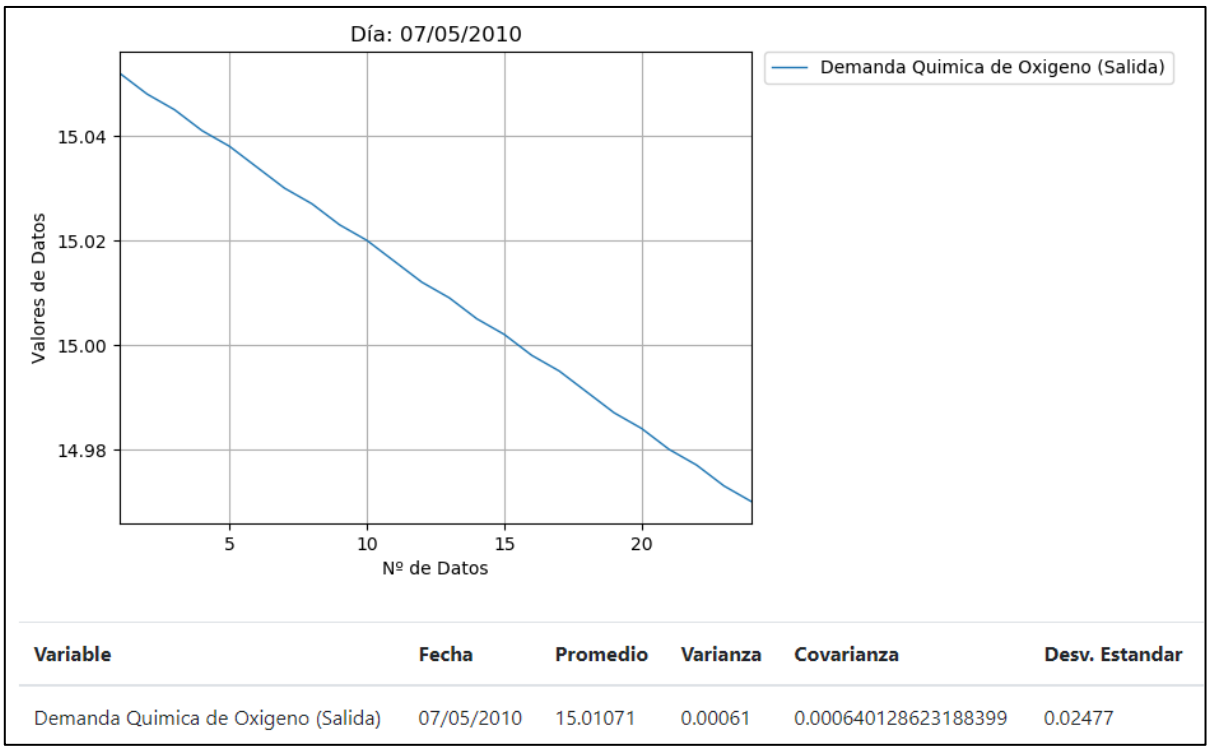
Figura 34.- Datos almacenados correctamente en la base de datos

4.2.- Análisis estadístico

En esta subsección se visualiza los diversos resultados obtenidos del uso de los formularios mencionados en el capítulo 3. Para estos ejemplos, se utilizó la base de datos “Prueba” para realizar cada uno de los cálculos estadísticos.

4.2.1.- Visualización de datos en día específico

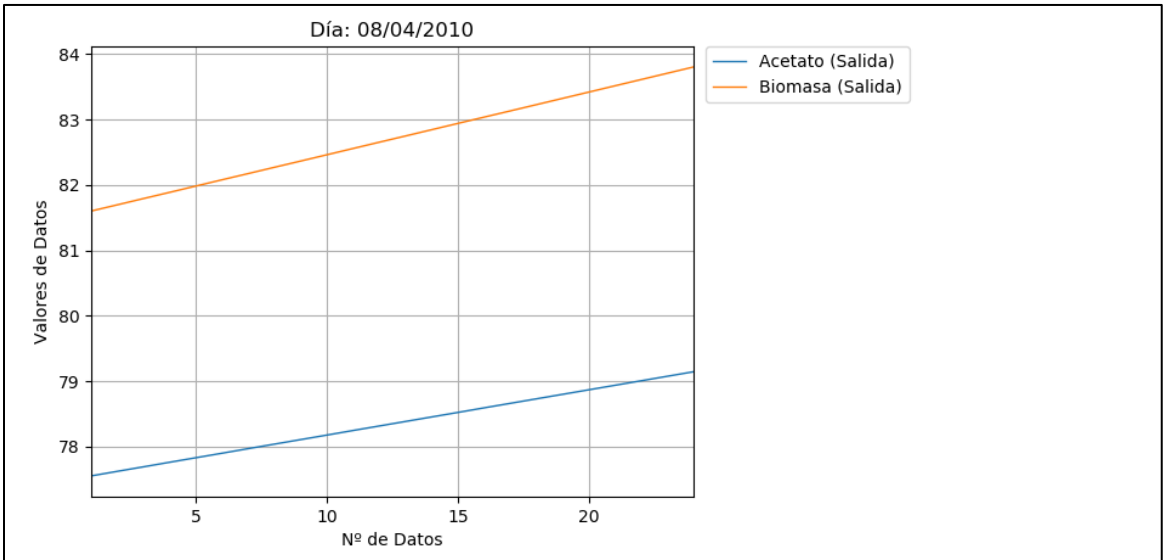
En este ejemplo se realizó el cálculo de promedio, varianza, covarianza y desviación estándar al igual que una gráfica que muestra los valores de los datos obtenidos en el día especificado y cuantos datos son en total.



Variable	Fecha	Promedio	Varianza	Covarianza	Desv. Estandar
Demanda Quimica de Oxigeno (Salida)	07/05/2010	15.01071	0.00061	0.000640128623188399	0.02477

Figura 35.- Análisis estadístico en un día específico con una variable

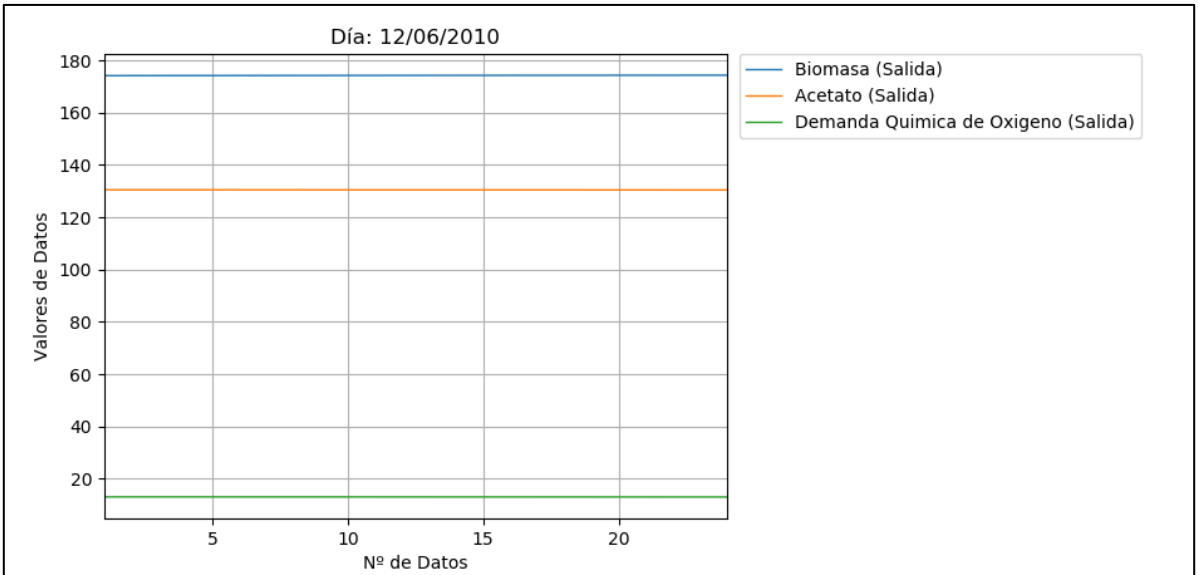
Como se logra apreciar en la figura 35, se utilizó la variable Demanda Química de Oxígeno a la salida del biorreactor y la fecha 01/03/2010 para realizar el cálculo. En la parte inferior del formulario, dentro de la misma página Web, aparece una gráfica mostrando los valores generados de ese dato en ese día especificado. Debajo de la gráfica se genera una tabla mostrando el nombre del dato utilizado, la fecha utilizada, el promedio, la varianza, la covarianza y la desviación estándar. Esos cálculos estadísticos se realizaron utilizando los datos generados en la fecha que se seleccionó. En los siguientes ejemplos se muestran más gráficas con una correlación entre 2 y hasta más datos a la vez.



Variable	Fecha	Promedio	Varianza	Covarianza	Desv. Estandar
Acetato (Salida)	08/04/2010	78.34804	0.22948	0.23945934601449312	0.47904
Biomasa (Salida)	08/04/2010	82.69992	0.43923	0.45832434057971216	0.66274

Figura 36.- Análisis estadístico en día especifico con dos variables

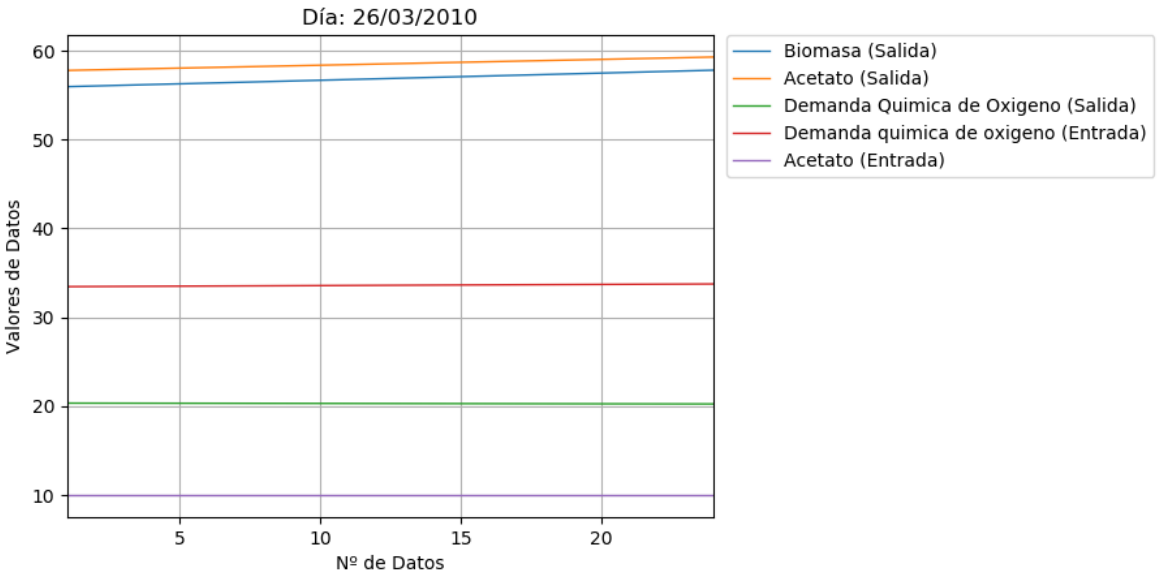
En la figura 36 se muestra un cálculo estadístico utilizando las variables Acetato y Biomasa a la salida del biorreactor. De igual manera que con un dato, se muestra la gráfica correspondiente mostrando los valores obtenidos en el día especificado y en la parte inferior la tabla con los cálculos estadísticos utilizando los mismos datos que se generaron ese día.



Variable	Fecha	Promedio	Varianza	Covarianza	Desv. Estandar
Biomasa (Salida)	12/06/2010	174.12042	0.00271	0.00283025362318813	0.05208
Acetato (Salida)	12/06/2010	130.37333	0.00045	0.0004666666666665922	0.02115
Demanda Quimica de Oxigeno (Salida)	12/06/2010	13.02317	0.00013	0.00013857971014492995	0.01152

Figura 37.- Análisis estadístico en día especifico con tres variables

En el ejemplo de la figura 37 se utilizaron tres variables, Biomasa, Acetato y Demanda Química de Oxígeno a la salida del biorreactor. De igual forma, con la gráfica correspondiente mostrando los datos y en la parte inferior los cálculos estadísticos.



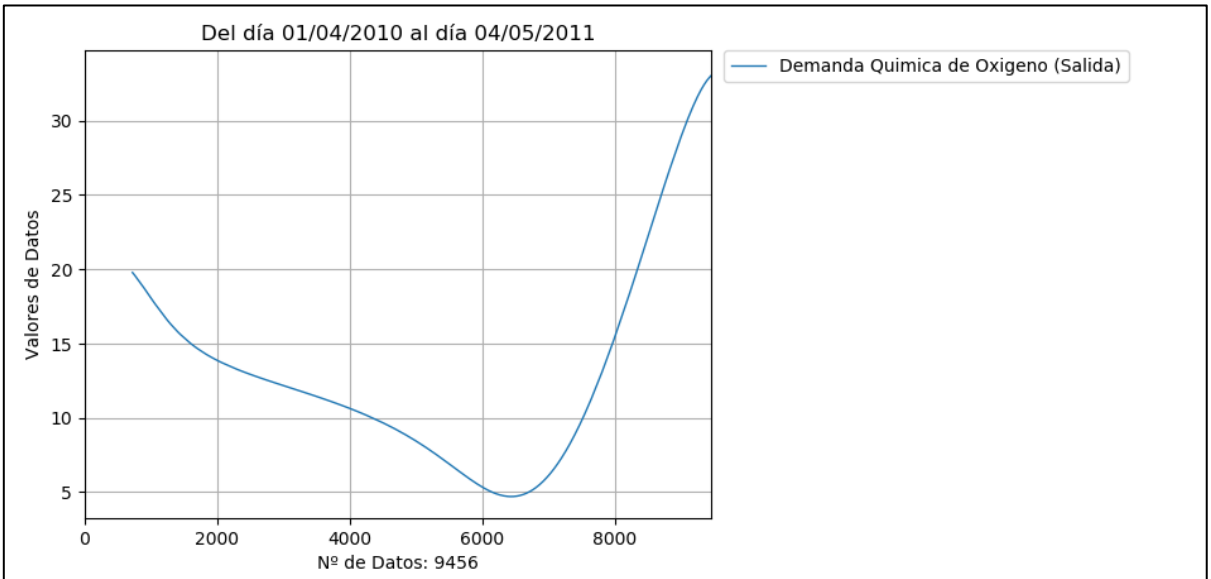
Variable	Fecha	Promedio	Varianza	Covarianza	Desv. Estandar
Biomasa (Salida)	26/03/2010	56.89658	0.31145	0.3249965144927539	0.55808
Acetato (Salida)	26/03/2010	58.55167	0.20567	0.21461597101449276	0.45351
Demanda Quimica de Oxigeno (Salida)	26/03/2010	20.28758	0.00074	0.0007705144927536258	0.02717
Demanda quimica de oxigeno (Entrada)	26/03/2010	33.60342	0.00826	0.008620514492753641	0.09089
Acetato (Entrada)	26/03/2010	10.0	0.0	0.0	0.0

Figura 38.- Análisis estadístico en día especifico con cinco variables

En la figura 38 se muestra el resultado de utilizar hasta 5 variables al mismo tiempo y generar la gráfica que muestra todos los valores de los datos en el día especificado y, la tabla en la parte inferior con el análisis estadístico correspondiente de cada uno de ellos.

4.2.2.- Visualización de datos en rango de días

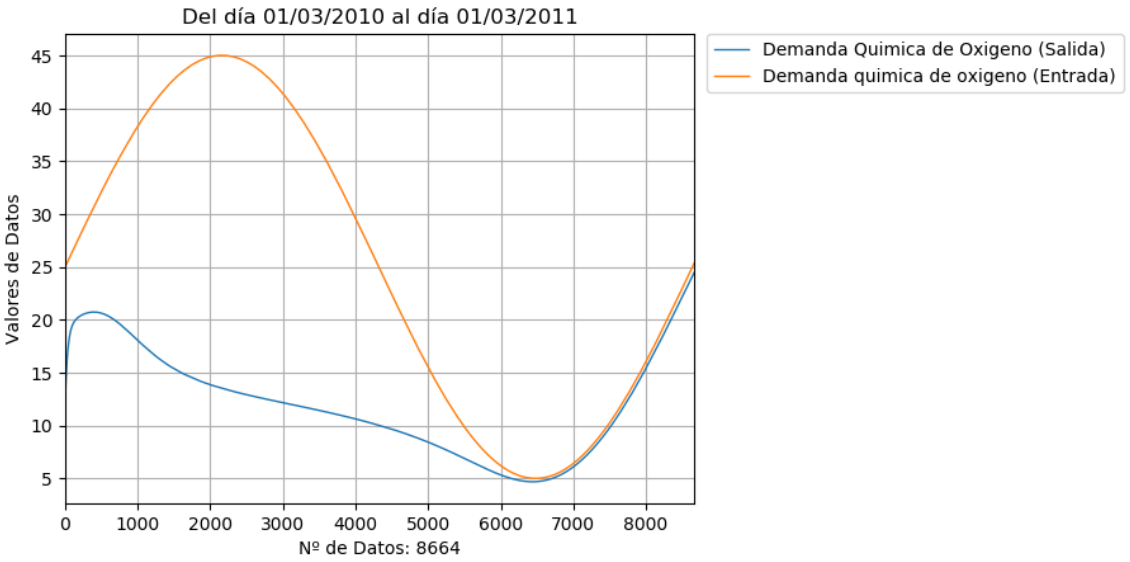
En el caso de querer especificar un rango de días específico el sistema Web muestra, de igual manera, la gráfica correspondiente y la tabla en la parte inferior con los cálculos estadísticos. En la figura 39 se muestra un ejemplo de visualización de datos en un rango de días especifico utilizando 1 variable.



Variable	Fecha inicio	Fecha final	Promedio	Varianza	Covarianza	Desv. Estandar
Demanda Quimica de Oxigeno (Salida)	01/04/2010	04/05/2011	14.33803	65.7711	65.77805862553338	8.10994

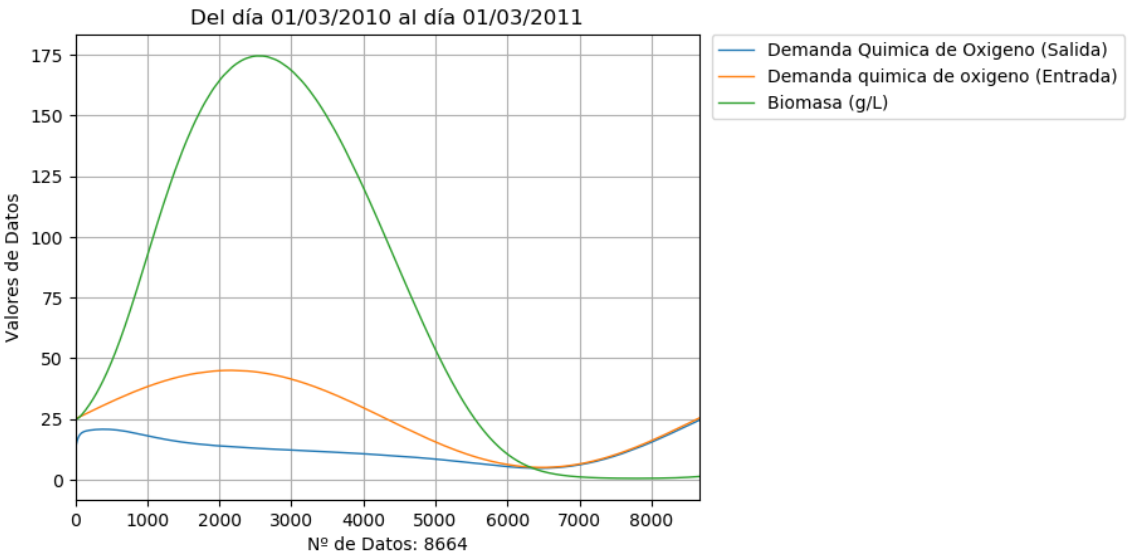
Figura 39.- Análisis estadístico en rango de días de una variable

En la misma figura, se muestra la gráfica correspondiente mostrando el rango de días que se asignó para realizar el cálculo estadístico, la cantidad de datos que se generaron en ese rango de días y los valores de esos datos. En la parte inferior, al igual que las gráficas anteriores, se muestra una tabla con los cálculos estadísticos correspondientes utilizando todos los datos generados en ese rango de días. En las Figuras siguientes, se muestran más ejemplos de análisis estadístico en rango de días de 2, 3 y 5 variables a la vez respectivamente.



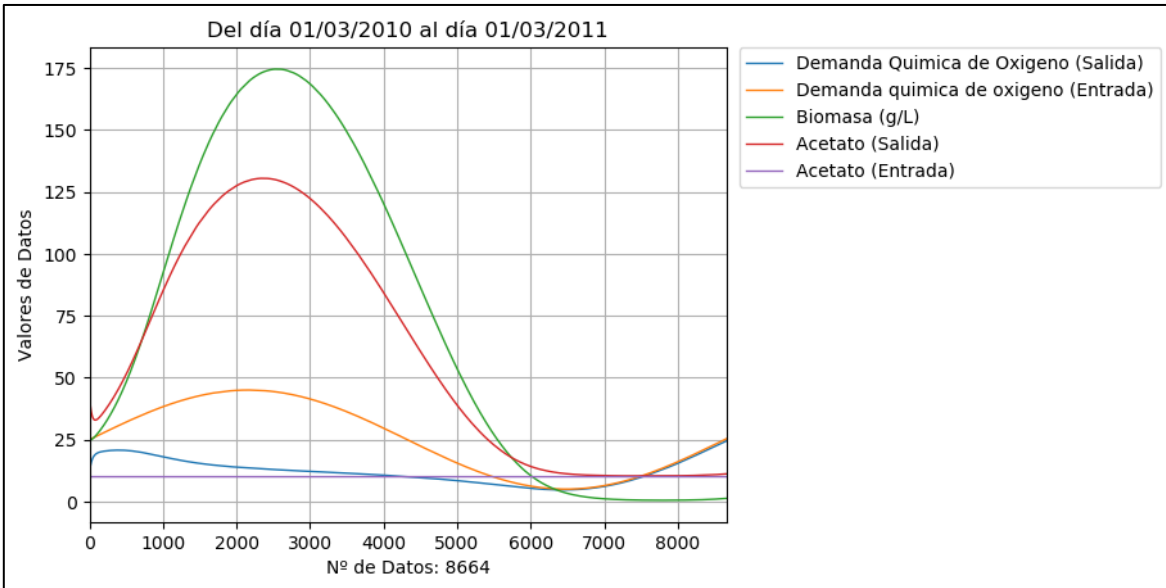
Variable	Fecha inicio	Fecha final	Promedio	Varianza	Covarianza	Desv. Estandar
Demanda Quimica de Oxigeno (Salida)	01/03/2010	01/03/2011	12.05667	24.77922	24.782081286145967	4.97787
Demanda quimica de oxigeno (Entrada)	01/03/2010	01/03/2011	25.00046	199.44939	199.47241452073234	14.12266

Figura 40.- Análisis estadístico en rango de días de dos variables



Variable	Fecha inicio	Fecha final	Promedio	Varianza	Covarianza	Desv. Estandar
Demanda Quimica de Oxigeno (Salida)	01/03/2010	01/03/2011	12.05667	24.77922	24.782081286145967	4.97787
Demanda quimica de oxigeno (Entrada)	01/03/2010	01/03/2011	25.00046	199.44939	199.47241452073234	14.12266
Biomasa (g/L)	01/03/2010	01/03/2011	72.20416	4230.35318	4230.841502853228	65.04117

Figura 41.- Análisis estadístico en rango de días de tres variables



Variable	Fecha inicio	Fecha final	Promedio	Varianza	Covarianza	Desv. Estandar
Demanda Quimica de Oxigeno (Salida)	01/03/2010	01/03/2011	12.05667	24.77922	24.782081286145967	4.97787
Demanda quimica de oxigeno (Entrada)	01/03/2010	01/03/2011	25.00046	199.44939	199.47241452073234	14.12266
Biomasa (g/L)	01/03/2010	01/03/2011	72.20416	4230.35318	4230.841502853228	65.04117
Acetato (Salida)	01/03/2010	01/03/2011	59.1182	2054.46607	2054.703222520947	45.32622
Acetato (Entrada)	01/03/2010	01/03/2011	10.0	0.0	0.0	0.0

Figura 42.- Análisis estadístico en rango de días de cinco variables

4.2.3.- Gráficas de cálculo estadístico en rango de días

A diferencia de la visualización de datos en días, en el apartado de análisis estadístico en rango de días se gráfica un tipo de análisis (promedio, varianza, covarianza o desviación estándar) utilizando los datos generados en un rango de días. En la figura 43 se muestra un ejemplo de una gráfica de la varianza de los datos generados en un día específico, dicha gráfica puede ser generada para cada uno de los análisis estadísticos, con cualquier rango de días y con correlación de hasta 5 variables a la vez.

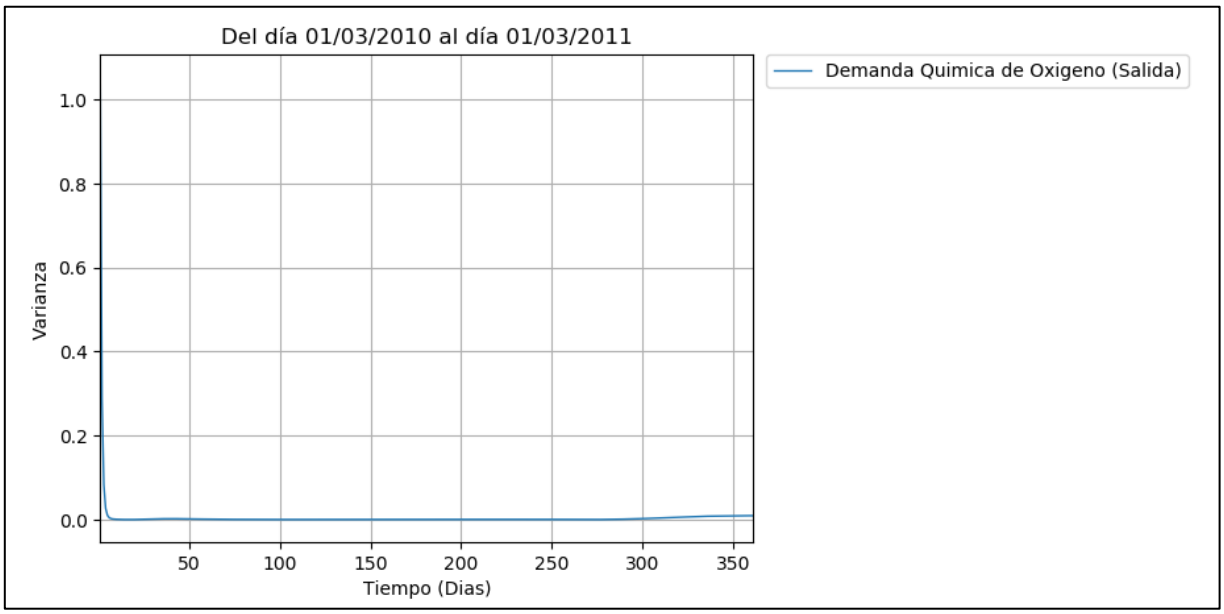


Figura 43.- Gráfica de varianza de una variable en rango de días

En las siguientes figuras, se utilizó la misma variable y el mismo rango de días, pero ahora utilizando el cálculo de promedio, covarianza y desviación estándar respectivamente.

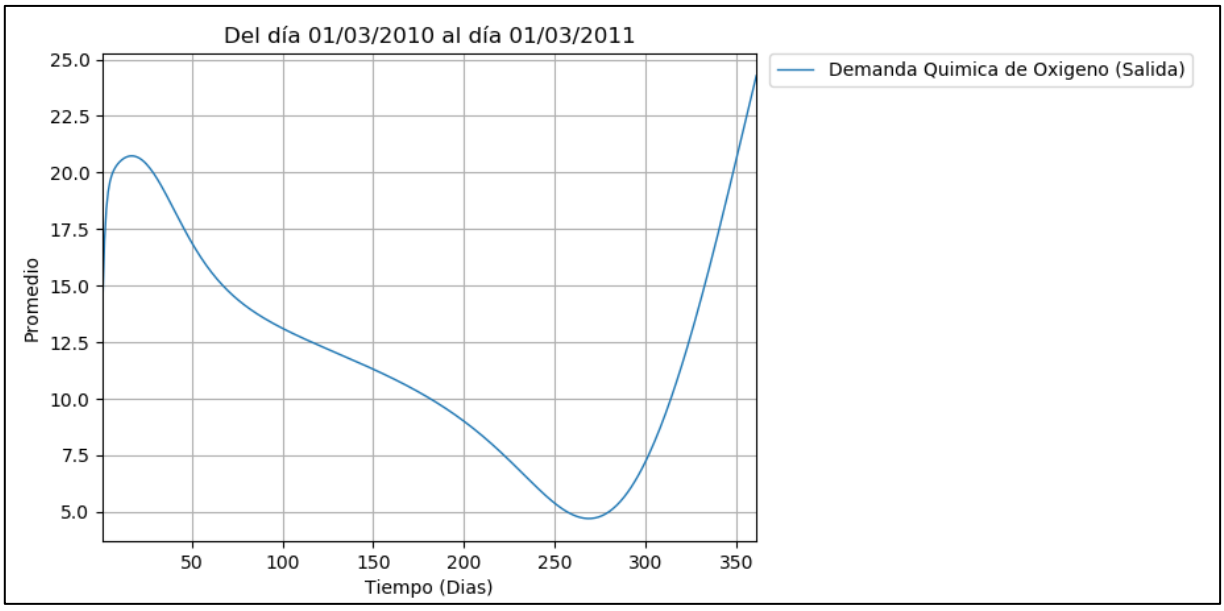


Figura 44.- Gráfica de promedio de una variable en rango de días

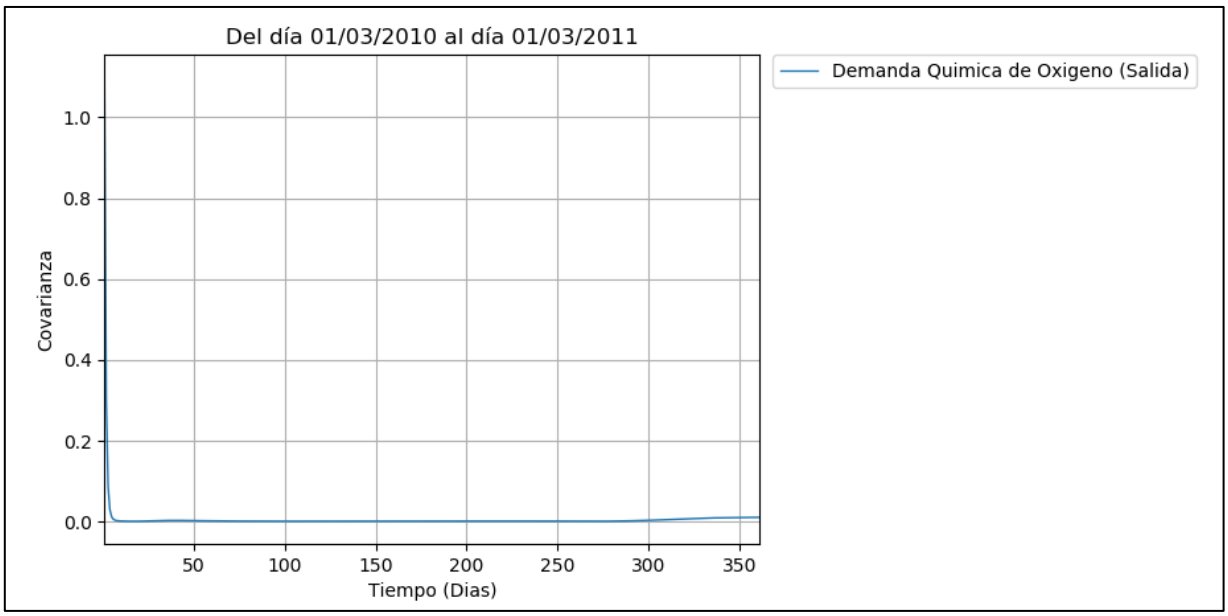


Figura 45.- Gráfica de covarianza de una variable en rango de días

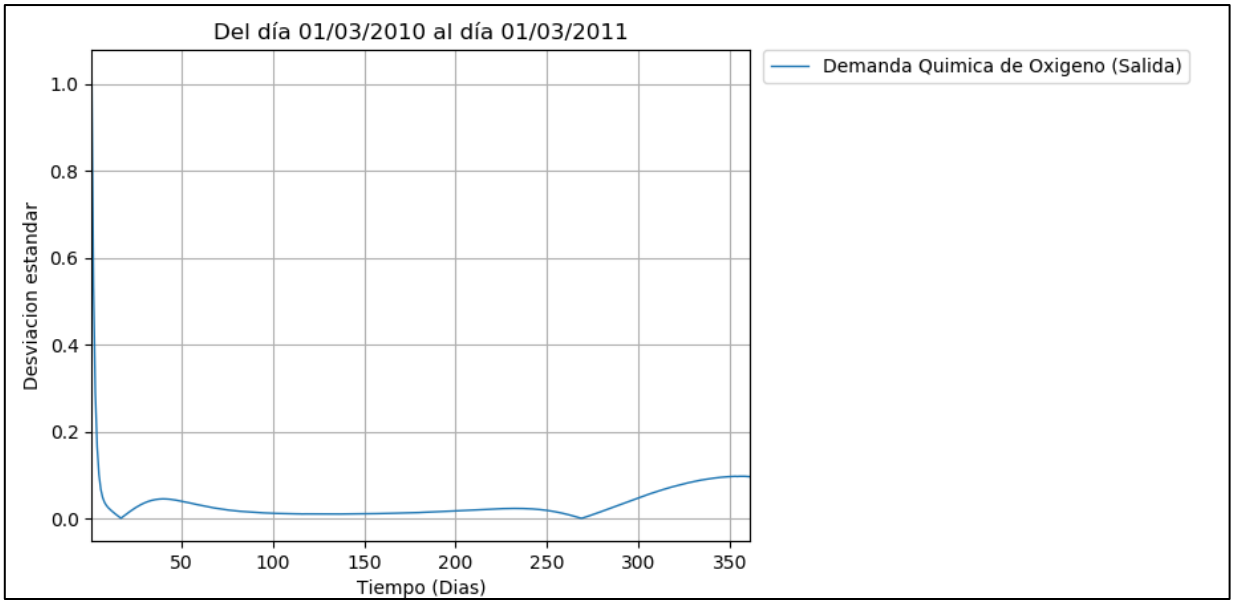


Figura 46.- Gráfica de desviación estándar de una variable en rango de días

En las siguientes figuras, se realizaron todos los cálculos estadísticos, pero ahora con cinco variables al mismo tiempo.

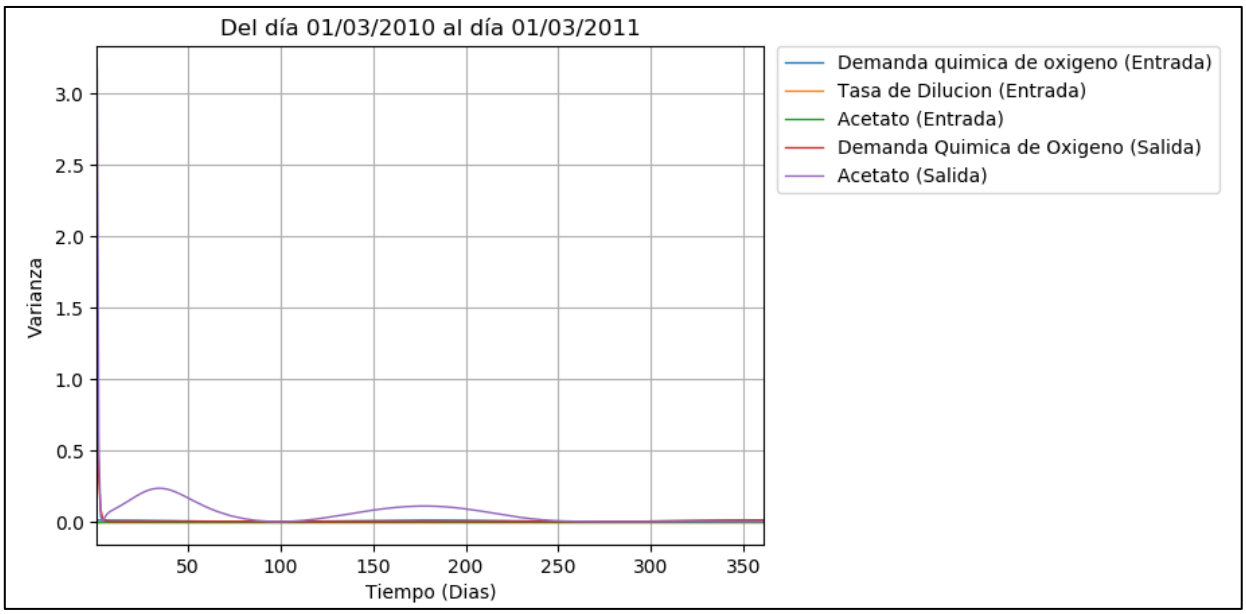


Figura 47.- Gráfica de varianza de cinco variables en rango de días

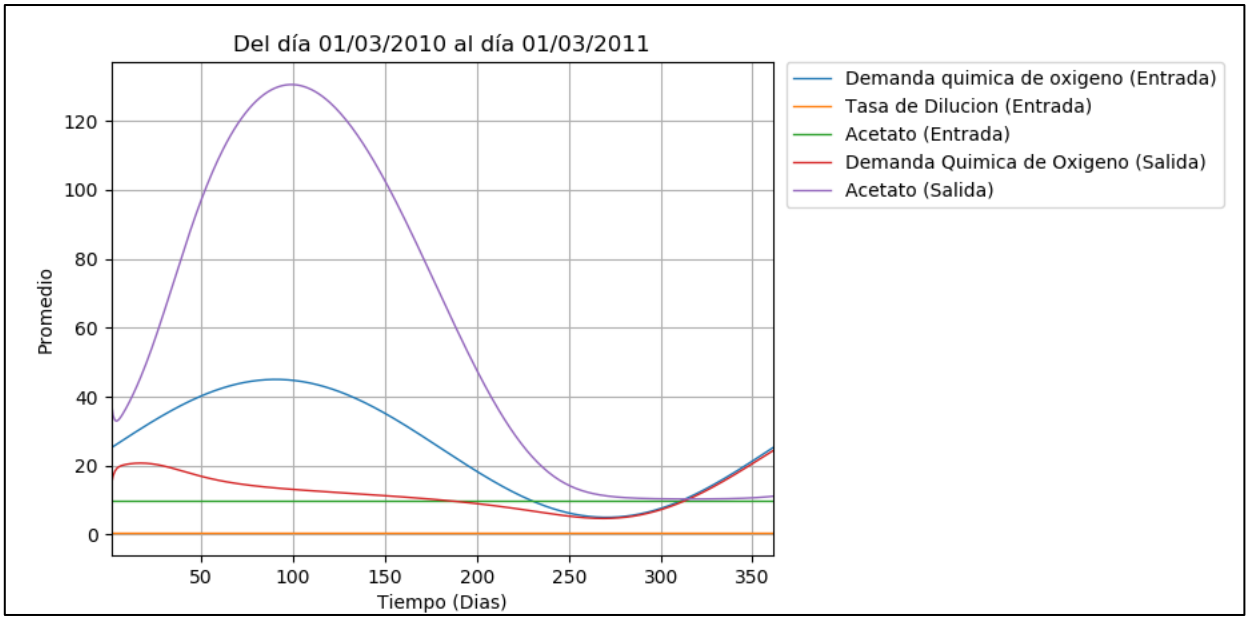


Figura 48.- Gráfica de promedio de cinco variables en rango de días

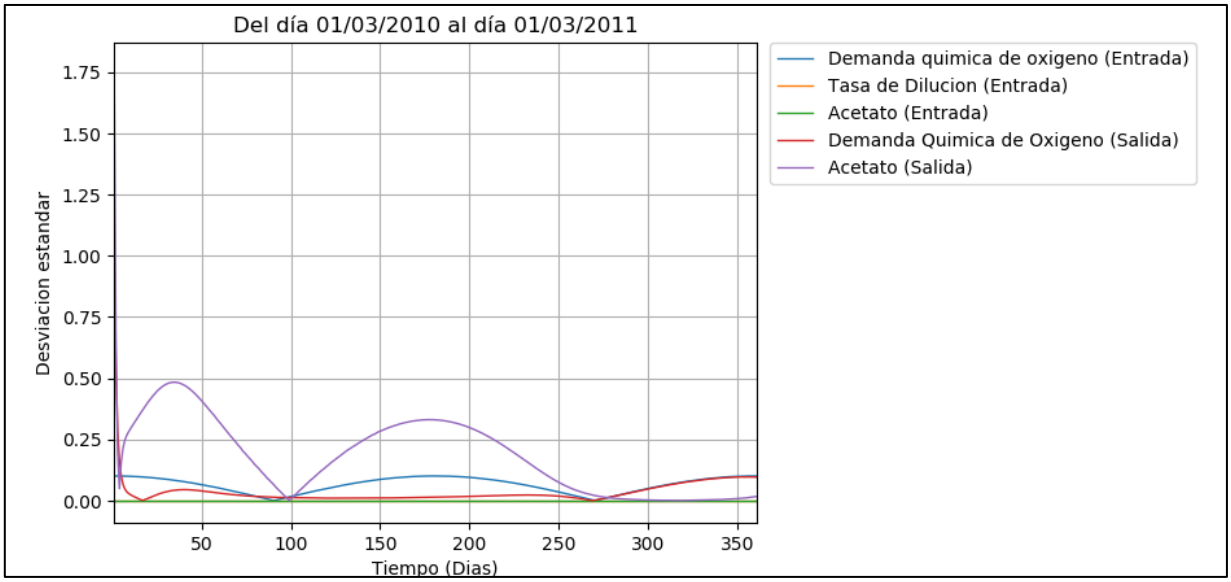


Figura 49.- Gráfica de desviación estándar de cinco variables en rango de días

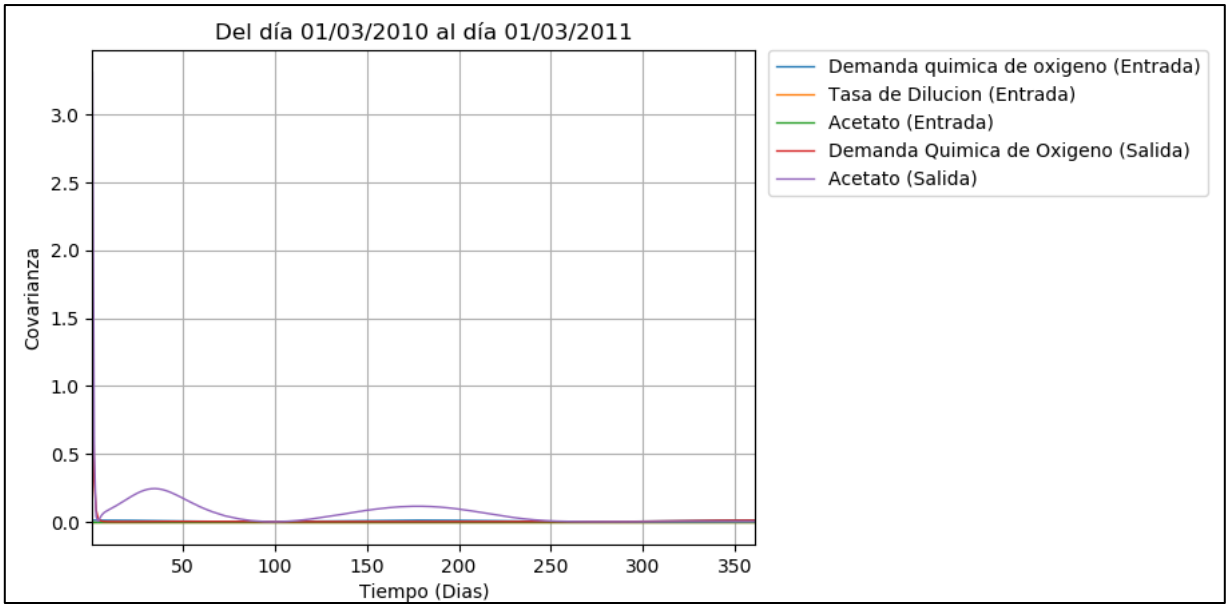


Figura 50.- Gráfica de covarianza de cinco variables en rango de días

4.2.4.- Prueba de Fisher

En la figura 51 se muestra un ejemplo de una gráfica generada utilizando el cálculo de la prueba de Fisher en un lapso de un mes utilizando las variables Demanda Química de Oxígeno a la entrada y Demanda Química de Oxígeno a la salida del biorreactor. Para realizar la gráfica es necesario especificar 2 variables y el rango de días en el cual el algoritmo tomará los datos.

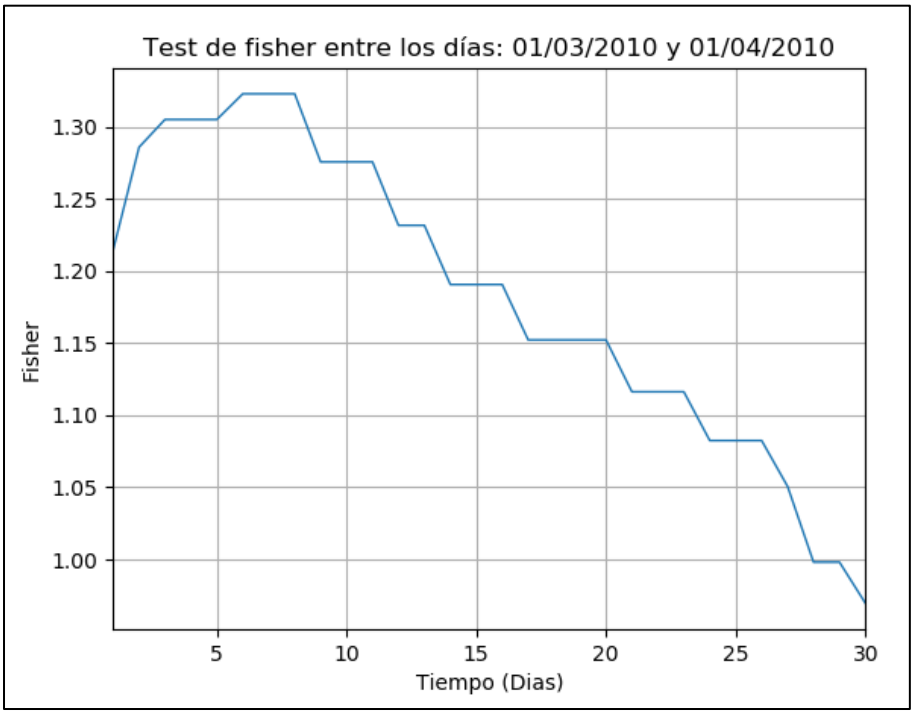


Figura 51.- Prueba de Fisher en lapso de un mes

En las figuras siguientes se muestran más ejemplos de gráficas de la Prueba de Fisher, pero ahora con lapsos de seis meses, un año y cinco años respectivamente utilizando las mismas variables.

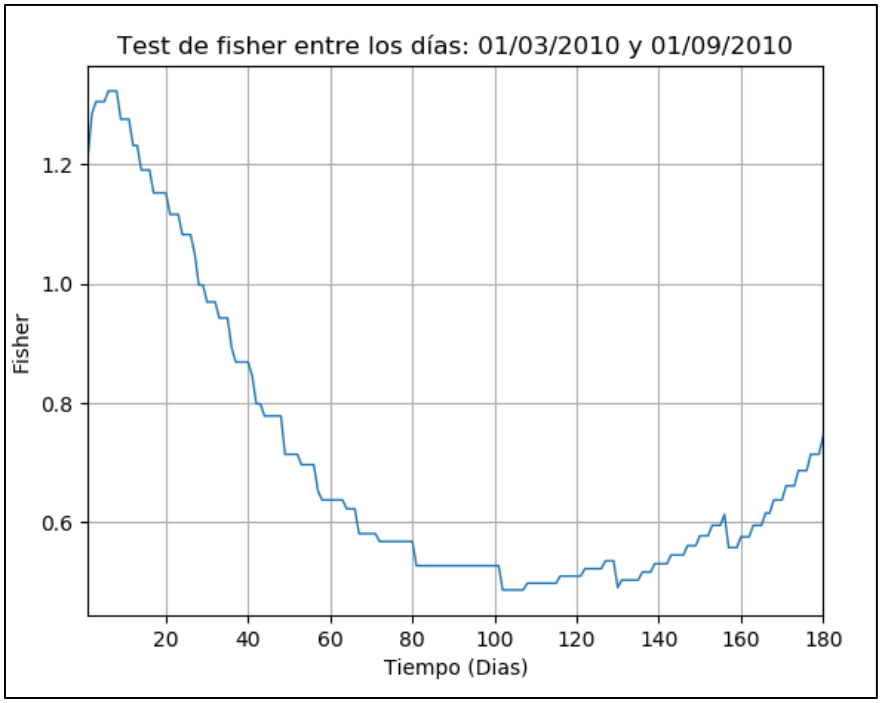


Figura 52.- Prueba de Fisher en lapso de seis meses

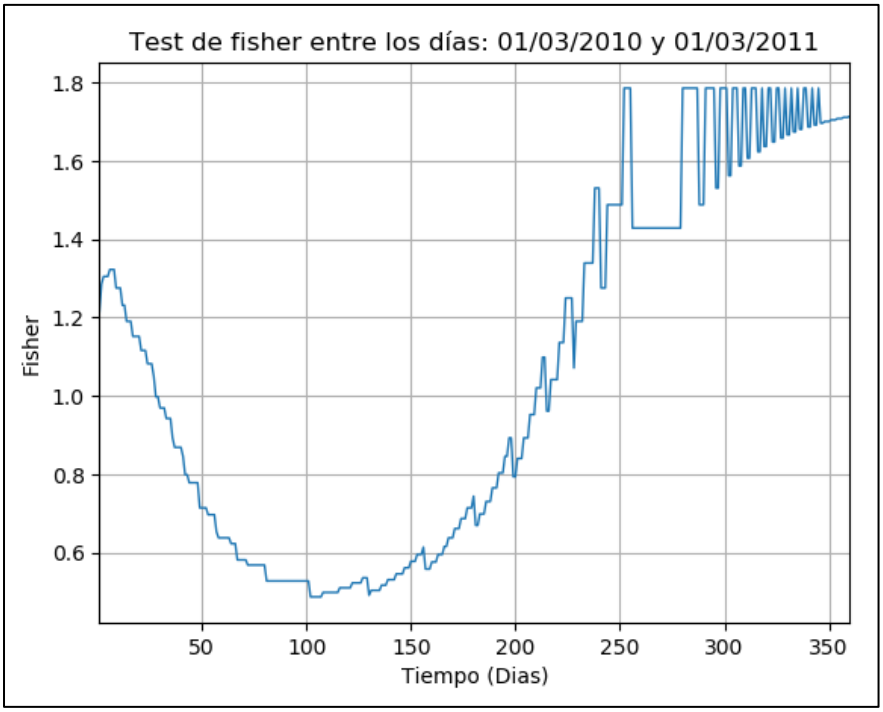


Figura 53.- Prueba de Fisher en lapso de un año

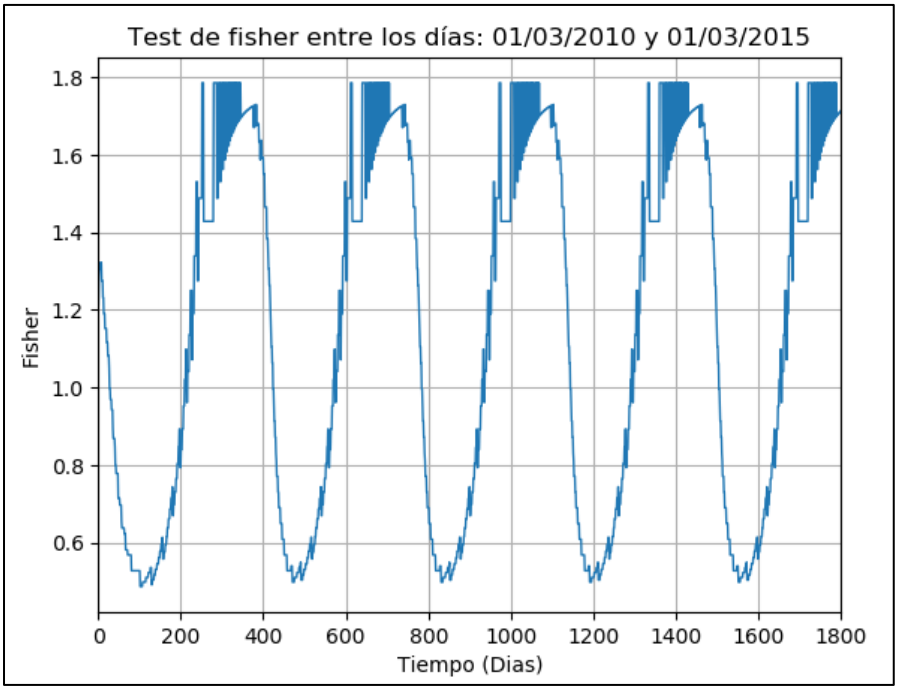


Figura 54.- Prueba de Fisher en lapso de cinco años

4.2.5.- ANOVA

El cálculo de ANOVA no genera gráficas, sino que genera únicamente la estadística y el *pvalue* de un rango especifico de días. En la figura 55 se muestra un ejemplo del cálculo de ANOVA de dos variables en un lapso de una semana.

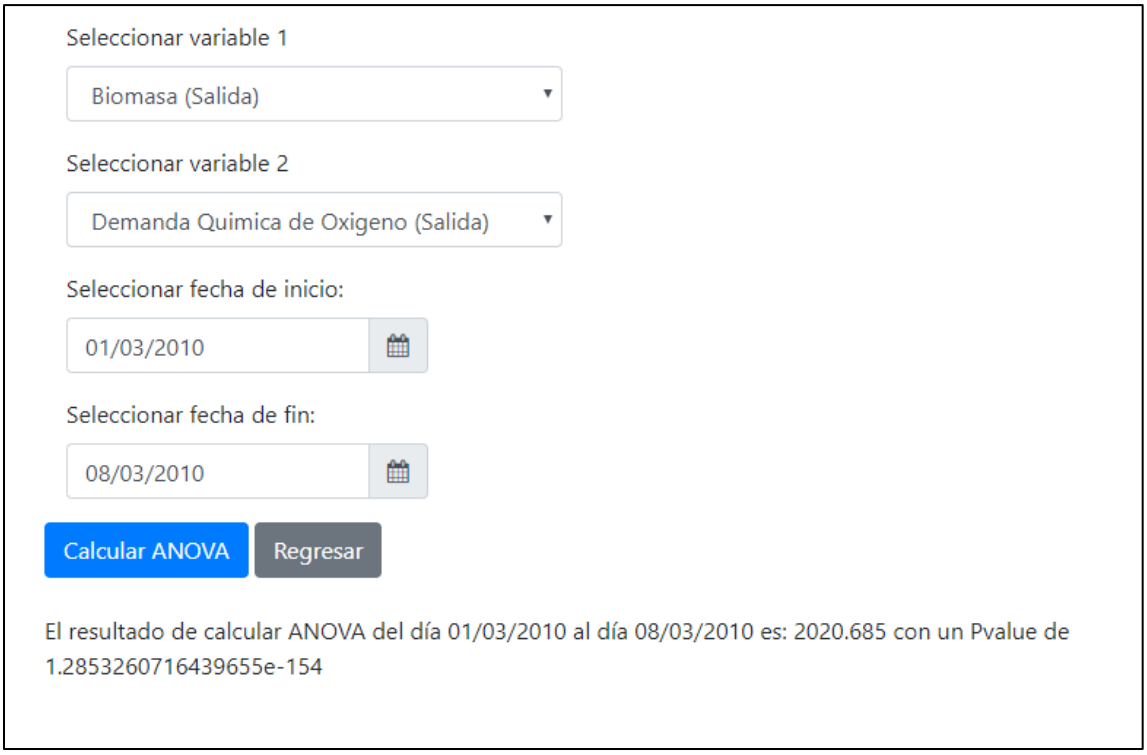


Figura 55.- Cálculo de ANOVA en lapso de 1 semana

En las figuras siguientes, se muestran cálculos de ANOVA, pero ahora con lapsos de un mes, seis meses y un año respectivamente.

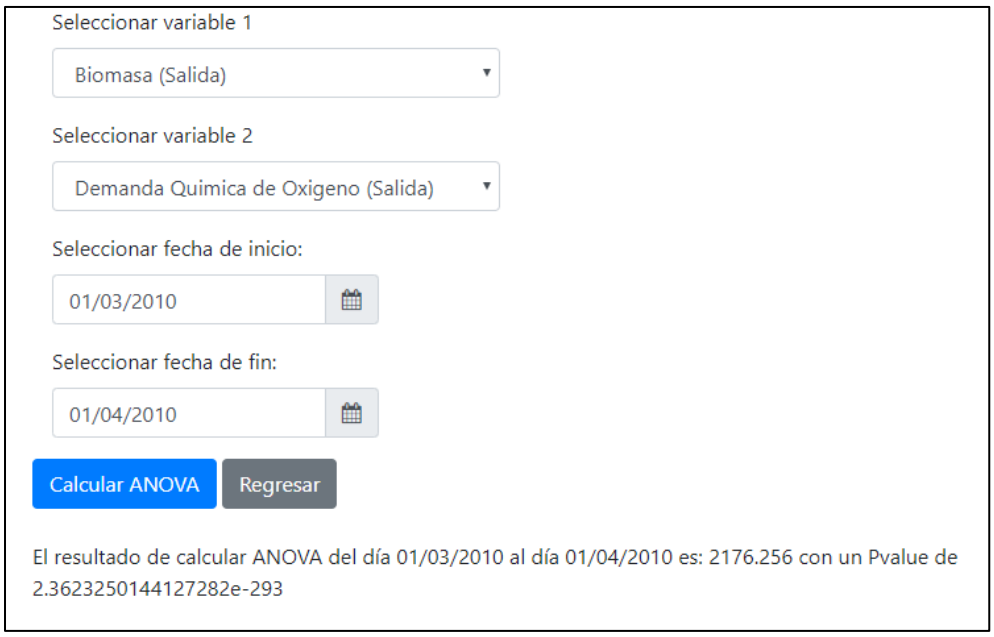


Figura 56.- Cálculo de ANOVA en lapso de un mes

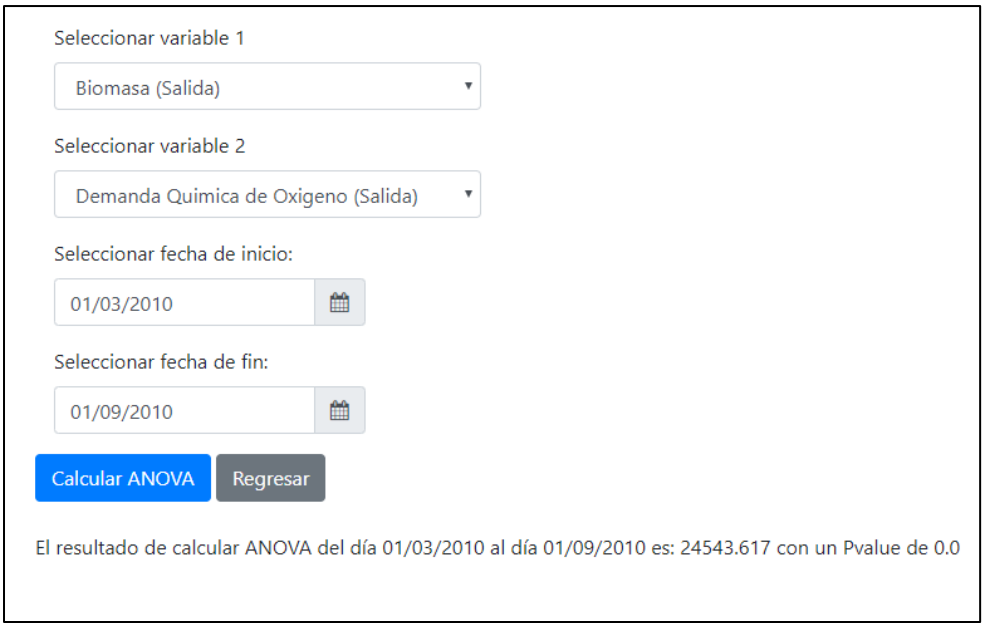


Figura 57.- Cálculo de ANOVA en lapso de seis meses

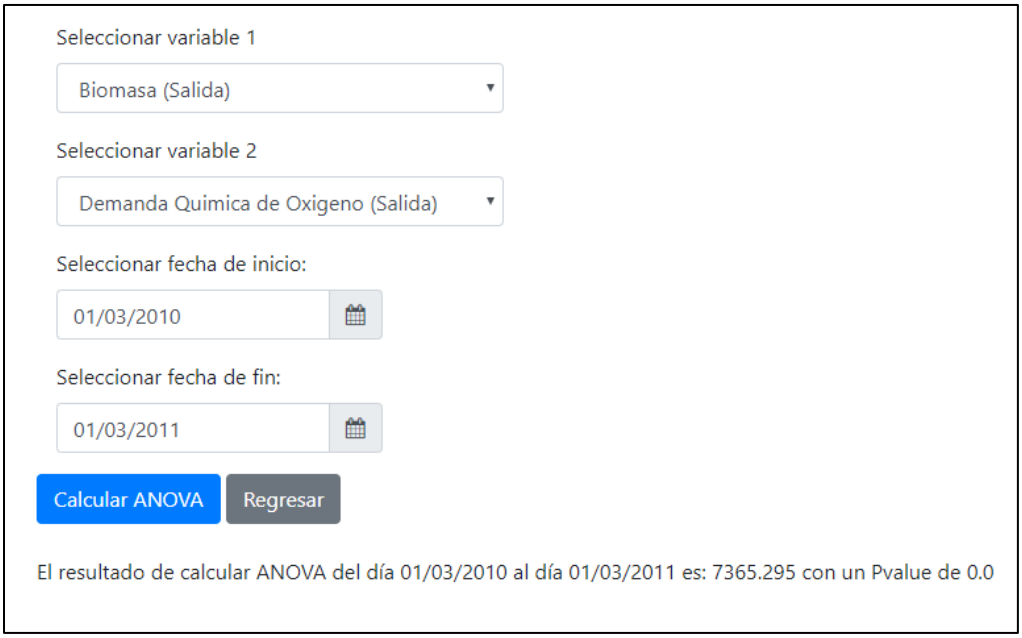


Figura 58.- Cálculo de ANOVA en lapso de un año

4.3.- Clustering

Para los ejemplos de clustering se utilizaron tres conjuntos de datos distintos con datos cercanos o excedidos de sus valores mínimos críticos o máximos críticos. Para el primer ejemplo se utilizó un dataset denominado “Errores1” y se generaron gráficas de cada una de sus variables de salida como se muestra en las figuras siguientes.

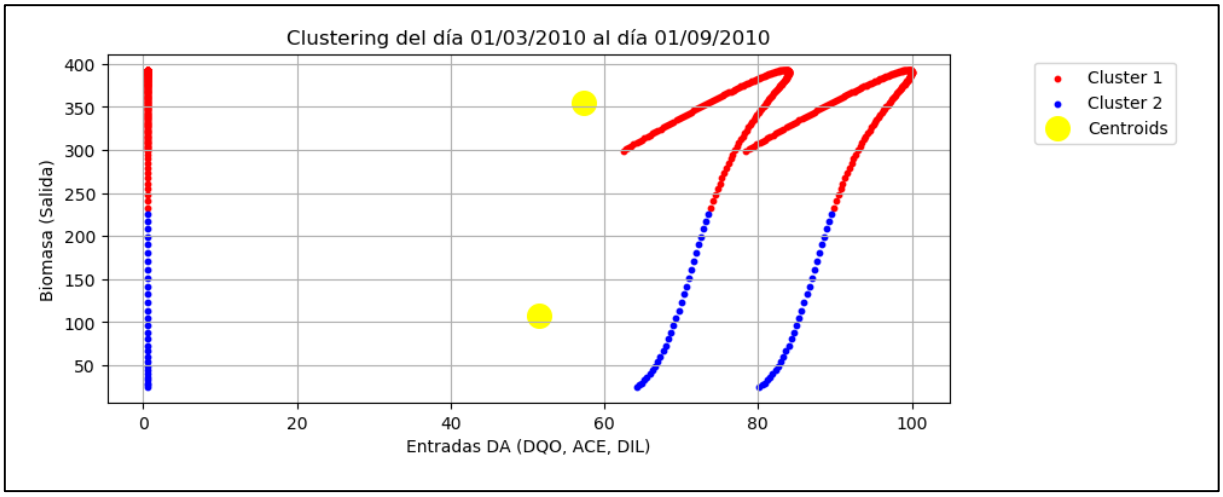


Figura 59.- Gráfica 1 de clustering del conjunto de datos"Errores1"

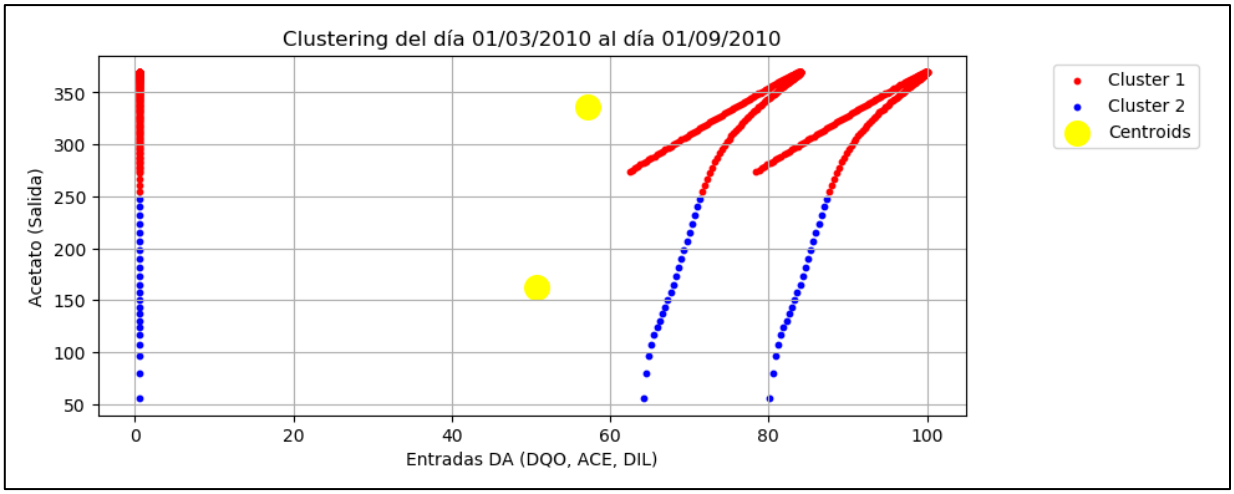


Figura 60.- Gráfica 2 de clustering del conjunto de datos "Errores1"

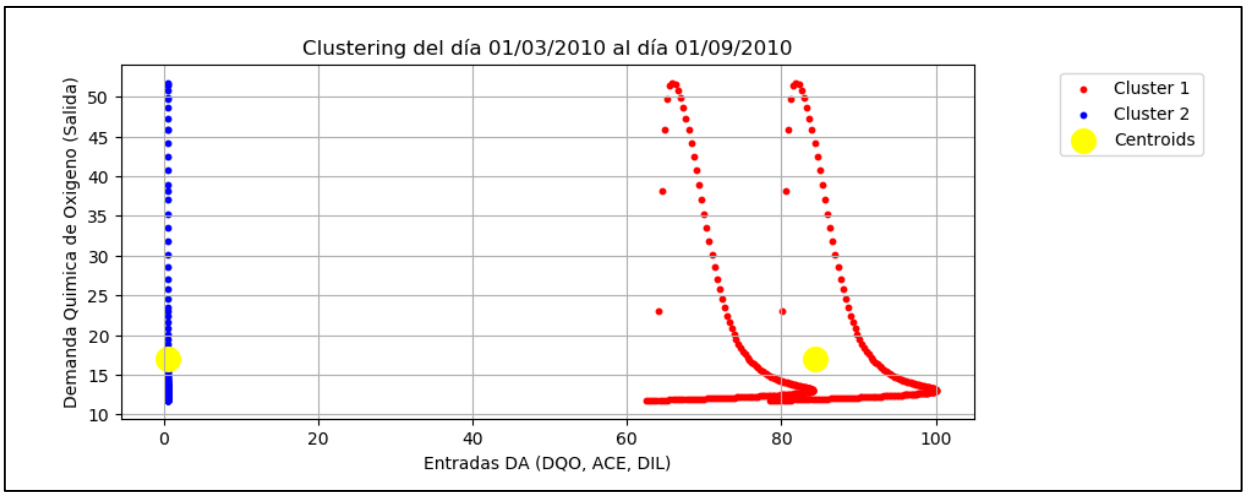


Figura 61.- Gráfica 3 de clustering del conjunto de datos "Errores1"

Como se muestra en las Figuras 59, 60 y 61, se realizó un cálculo de clustering del conjunto de datos “Errores1” utilizando 2 clústers, esto para determinar el rango que llegan a tener los valores de salida en cuando a los valores de entrada. Cada gráfica representa los valores de salida de un proceso de la biorrefinería y son graficados junto con los valores de entrada del mismo proceso. En la Tabla 23 se muestra la información más detallada y con algunas observaciones y recomendaciones de rendimiento.

Tabla de recomendaciones de rendimiento

Fecha	DQO (in)	AGV (in)	DIL (in)	Biomasa (out)	AGV (out)	DQO (out)	Observaciones	Recomendación
01/03/2010	64.16	80.16	0.6	25.158	55.868	23.014	DQO (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in),
02/03/2010	64.509	80.509	0.6	26.883	80.012	38.202	DQO (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in),
03/03/2010	64.858	80.858	0.6	29.414	96.34	45.932	DQO (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in),
04/03/2010	65.206	81.206	0.6	32.467	107.807	49.716	DQO (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in),
05/03/2010	65.555	81.555	0.6	35.967	116.55	51.349	DQO (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in),
06/03/2010	65.902	81.902	0.6	39.896	123.881	51.777	DQO (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in),

Tabla 23.- Tabla de recomendaciones de rendimiento del conjunto de datos "Errores1"

Como se muestra en la Tabla 24, hay información detallada del conjunto de datos utilizado para el cálculo de clustering. Esta información va desde la fecha en que se generaron los datos, los valores promedio de los datos generados ese día, observaciones que indican si un valor esta un 20% cercano al valor mínimo y/o máximo o 20% sobre el valor mínimo y/o máximo y las recomendaciones de rendimiento según sea la observación dada. Cuando se detecta una observación de un valor cercano a un valor mínimo o máximo, el recuadro de dicho valor es iluminado de color amarillo, en caso de ser un valor por encima del valor mínimo o máximo el recuadro se pondrá con color rojo como se muestra en la tabla 6.

Tabla de recomendaciones de rendimiento

Fecha	DQO (in)	AGV (in)	DIL (in)	Biomasa (out)	AGV (out)	DQO (out)	Observaciones	Recomendación
19/04/2010	79.198	95.198	0.6	332.625	339.022	14.534	DQO (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in),
20/04/2010	79.423	95.423	0.6	335.528	340.49	14.441	DQO (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in),
21/04/2010	79.643	95.643	0.6	338.304	341.907	14.356	DQO (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in),
22/04/2010	79.858	95.858	0.6	340.96	343.281	14.276	DQO (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in),
23/04/2010	80.068	96.068	0.6	343.504	344.609	14.202	DQO (in) encima del valor Máximo,	Disminuir el valor de DQO (in),
24/04/2010	80.273	96.273	0.6	345.941	345.895	14.134	DQO (in) encima del valor Máximo,	Disminuir el valor de DQO (in),

Tabla 24.- Tabla de recomendaciones de rendimiento con valores cercanos y por encima de valores mínimos y/o máximos

Para el ejemplo número 2 se utilizó un conjunto de datos con el nombre “Errores2” y, de igual manera, se generaron las siguientes gráficas tal y como se ven en las siguientes figuras.

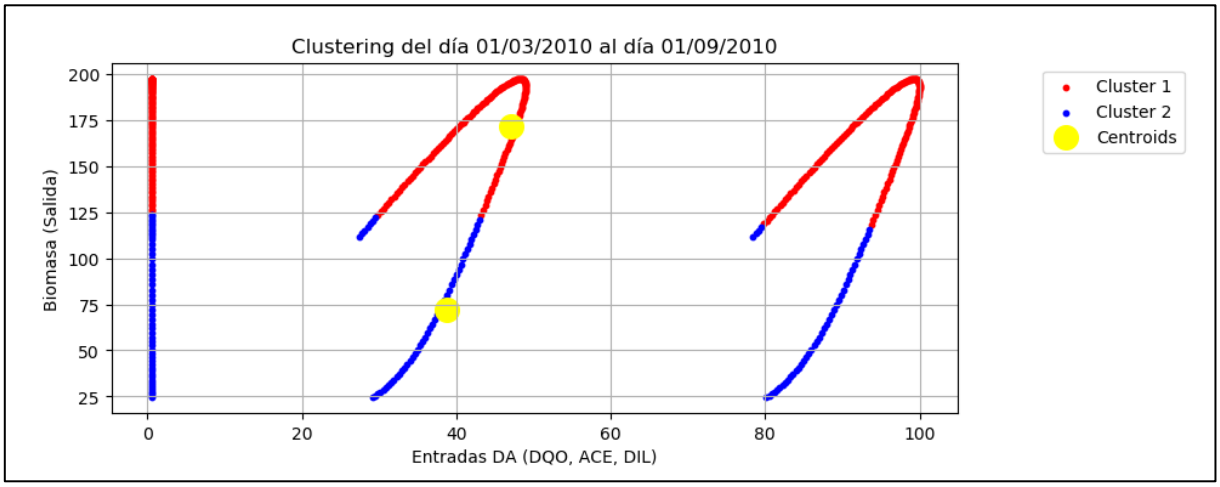


Figura 62.- Gráfica 1 de clustering del conjunto de datos"Errores2"

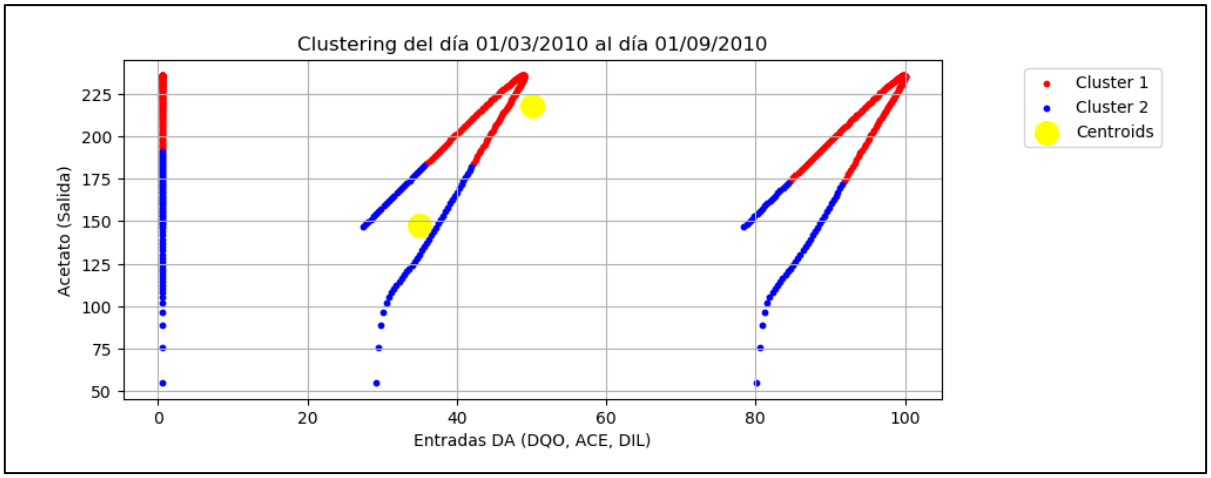


Figura 63.- Gráfica 2 de clustering del conjunto de datos "Errores2"

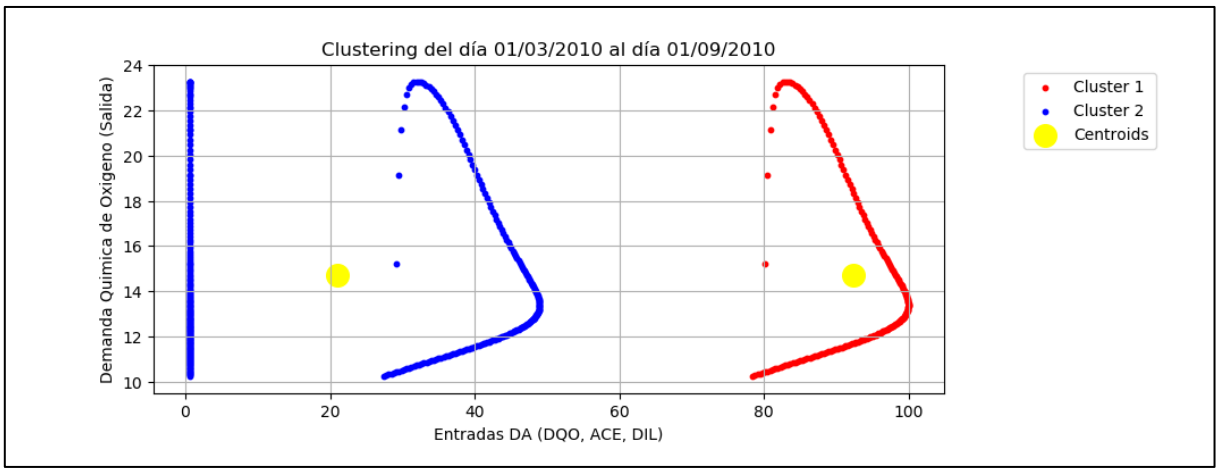


Figura 64.- Gráfica 3 de clustering del conjunto de datos"Errores2"

Para este ejemplo igual se ha generado una tabla como se muestra en la tabla 25. Como se puede observar en la tabla no muestra ninguna observación ni recomendación alguna, esto indica que los valores generados en esos días están en sus rangos normales y ninguno cercano al mínimo y/o al máximo.

Tabla de recomendaciones de rendimiento

Fecha	DQO (in)	AGV (in)	DIL (in)	Biomasa (out)	AGV (out)	DQO (out)	Observaciones	Recomendación
01/03/2010	29.16	80.16	0.6	24.895	54.977	15.205	Sin Observaciones	Sin Recomendaciones
02/03/2010	29.509	80.509	0.6	25.447	75.818	19.107	Sin Observaciones	Sin Recomendaciones
03/03/2010	29.858	80.858	0.6	26.32	88.532	21.113	Sin Observaciones	Sin Recomendaciones
04/03/2010	30.206	81.206	0.6	27.372	96.471	22.162	Sin Observaciones	Sin Recomendaciones
05/03/2010	30.555	81.555	0.6	28.543	101.668	22.715	Sin Observaciones	Sin Recomendaciones
06/03/2010	30.902	81.902	0.6	29.806	105.316	23.008	Sin Observaciones	Sin Recomendaciones
07/03/2010	31.25	82.25	0.6	31.147	108.101	23.161	Sin Observaciones	Sin Recomendaciones
08/03/2010	31.596	82.596	0.6	32.561	110.42	23.235	Sin Observaciones	Sin Recomendaciones
09/03/2010	31.942	82.942	0.6	34.044	112.495	23.262	Sin Observaciones	Sin Recomendaciones

Tabla 25.- Tabla de recomendaciones de rendimiento del dataset "Errores2"

En el ejemplo 3 se utilizó un conjunto de datos con nombre “Errores3” y las gráficas generadas se muestran en las figuras siguientes.

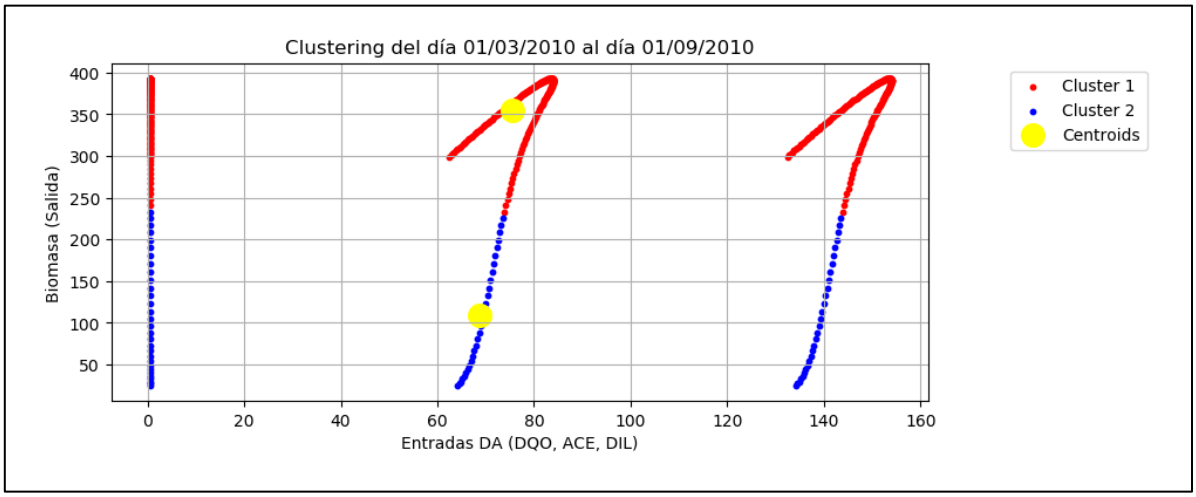


Figura 65.- Gráfica 1 de clustering del conjunto de datos "Errores3"

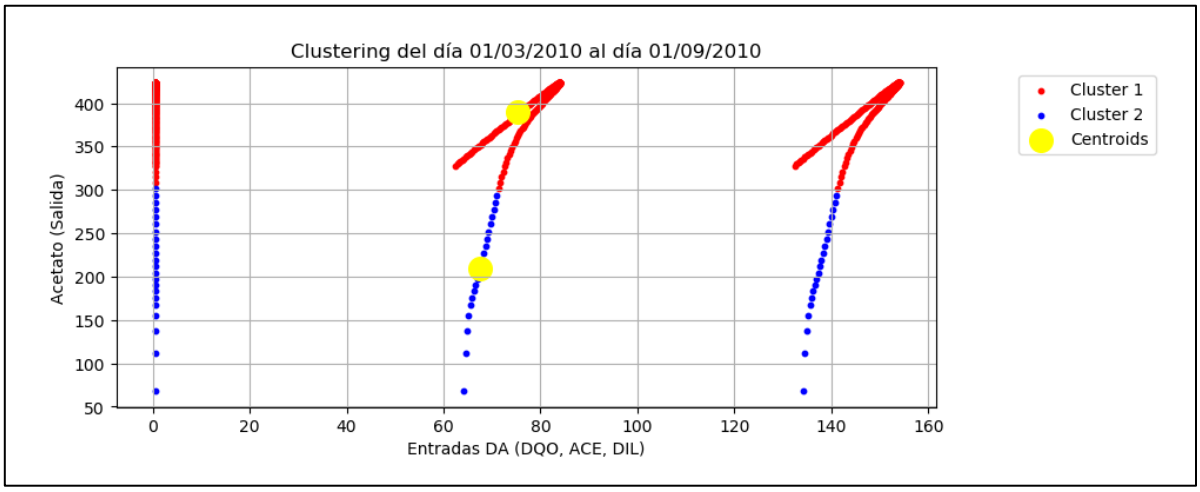


Figura 66.- Gráfica 2 de clustering del conjunto de datos "Errores3"

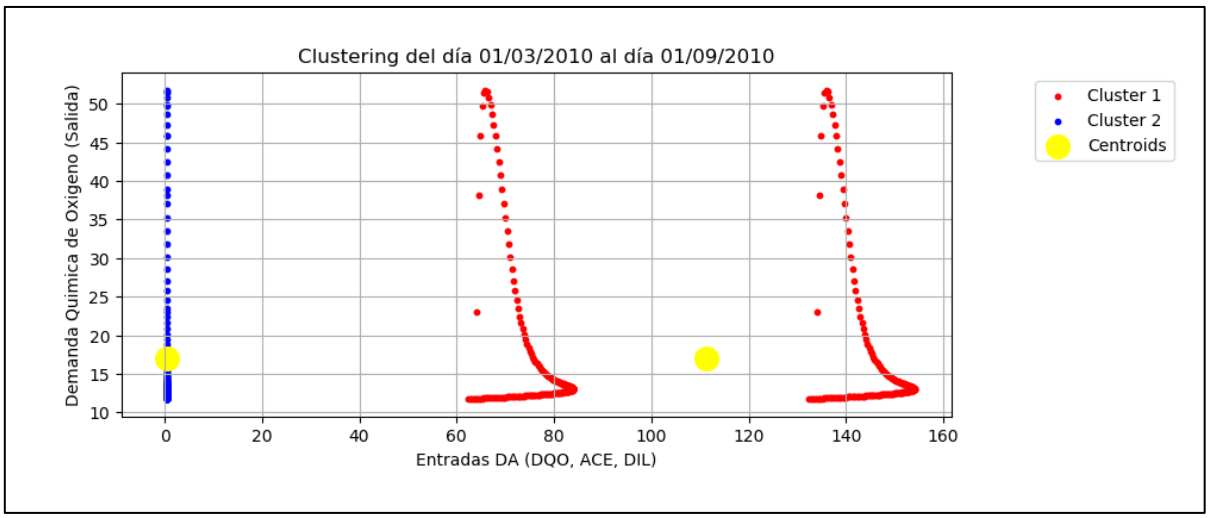


Figura 67.- Gráfica 3 de clustering del conjunto de datos "Errores3"

En la tabla 26 de recomendaciones de rendimiento del conjunto de datos “Errores3” se puede observar que hay dos variables que se encuentran un 20% cercano a su valor máximo y que sus celdas están iluminadas de color amarillo.

Tabla de recomendaciones de rendimiento

Fecha	DQO (in)	AGV (in)	DIL (in)	Biomasa (out)	AGV (out)	DQO (out)	Observaciones	Recomendación
01/03/2010	64.16	134.16	0.6	25.158	68.278	23.014	DQO (in) está 20% cercano al valor Máximo, AGV (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in), Disminuir el valor de AGV (in),
02/03/2010	64.509	134.509	0.6	26.883	111.174	38.202	DQO (in) está 20% cercano al valor Máximo, AGV (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in), Disminuir el valor de AGV (in),
03/03/2010	64.858	134.858	0.6	29.414	137.8	45.932	DQO (in) está 20% cercano al valor Máximo, AGV (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in), Disminuir el valor de AGV (in),
04/03/2010	65.206	135.207	0.6	32.467	154.923	49.716	DQO (in) está 20% cercano al valor Máximo, AGV (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in), Disminuir el valor de AGV (in),
05/03/2010	65.555	135.555	0.6	35.967	166.768	51.349	DQO (in) está 20% cercano al valor	Disminuir el valor de DQO

Tabla 26.- Tabla de recomendaciones del dataset "Errores3" con dos variables cercanos a su valor máximo

Si deslizamos hacia abajo la tabla podemos encontrar que algunos valores ya están encima de sus valores máximos y marcados con color rojo. Tal y como se muestra en la tabla 27.

Tabla de recomendaciones de rendimiento

Fecha	DQO (in)	AGV (in)	DIL (in)	Biomasa (out)	AGV (out)	DQO (out)	Observaciones	Recomendación
21/04/2010	79.643	149.643	0.6	338.304	395.907	14.356	DQO (in) está 20% cercano al valor Máximo, AGV (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in), Disminuir el valor de AGV (in),
22/04/2010	79.858	149.858	0.6	340.96	397.281	14.276	DQO (in) está 20% cercano al valor Máximo, AGV (in) está 20% cercano al valor Máximo,	Disminuir el valor de DQO (in), Disminuir el valor de AGV (in),
23/04/2010	80.068	150.068	0.6	343.504	398.609	14.202	DQO (in) encima del valor Máximo, AGV (in) encima del valor Máximo,	Disminuir el valor de DQO (in), Disminuir el valor de AGV (in),
24/04/2010	80.273	150.274	0.6	345.941	399.895	14.134	DQO (in) encima del valor Máximo, AGV (in) encima del valor Máximo,	Disminuir el valor de DQO (in), Disminuir el valor de AGV (in),
25/04/2010	80.474	150.474	0.6	348.278	401.14	14.07	DQO (in) encima del valor Máximo, AGV	Disminuir el valor de DQO

Tabla 27.- Tabla de recomendaciones del dataset "Errores3" con dos variables por encima de su valor máximo

Conclusiones

Hoy día al llevar a cabo una toma de decisiones acerca de diversos procesos y aplicar las observaciones correspondientes puede llevar mucho tiempo y al momento de querer realizar un cambio el proceso correspondiente ya cambio a otro estado de tiempo y puede que haya problemas de rendimiento debido a un cambio indebido. Lo más recomendable es construir un sistema de análisis de procesos que agilice la toma de decisiones y ya sea que aplique automáticamente las correcciones correspondientes o en otro caso, que genere recomendaciones en casi tiempo real para realizar un cambio a algún proceso antes de que llegue a otro estado de tiempo. En este trabajo de tesis vi que el proceso de toma de decisiones implicaba mucho tiempo y al momento de aplicar una corrección la biorrefinería ya había generado nueva información y no era correcto aplicar cambios para un estado anterior. Es por eso que al construir los módulos descritos en los capítulos de esta tesis ha ayudado a agilizar ese proceso e incluso a generar información y graficas adicionales para tener un mejor entendimiento de la información que esta generando la biorrefinería y tener un mejor entendimiento de sus procesos.

Referencias

[1]. King, D. (2010). The future of industrial biorefineries. World Economic Forum.

[2]. Baliban, R. C., Elia, J. A., Weekman, V., & Floudas, C. A. (2012). Process synthesis of hybrid coal, biomass, and natural gas to liquids via Fischer–Tropsch synthesis, ZSM-5 catalytic conversion, methanol synthesis, methanol-to-gasoline, and methanol-to-olefins/distillate technologies. *Computers & Chemical Engineering*, *47*, 29-56.

[3]. Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE.

[4]. Hand, D. J. (2007). Principles of data mining. *Drug safety*, *30*(7), 621-622.

[5]. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37.

[6]. M. James, C. Michael, B. Brad, and B. Jacques, Big Data: The Next Frontier for Innovation, Competition, and Productivity. New York, NY: McKinsey Global Institute, 2011.

[7]. M. Rouse. (2011). Machine Learning Definition. [Online]. Available: http://whatis.techtarget.com/definition/machine-learning

[8]. Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, *42*(11), 30-36.

[9]. Maadane, A., Merghoub, N., Ainane, T., El Arroussi, H., Benhima, R., Amzazi, S., ... & Wahby, I. (2015). Antioxidant activity of some Moroccan marine microalgae: Pufa profiles, carotenoids and phenolic content. *Journal of biotechnology*, *215*, 13-19.Jason Brownlee. (2013). *How to Prepare Data For Machine Learning.* [Online]. https://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/

[10]. El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In Machine Learning in Radiation Oncology (pp. 3-11). Springer, Cham.

[11]. Mitchell TM. Machine learning. New York: McGraw-Hill; 1997.

[12]. Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.

[13]. Apolloni B. Machine learning and robot perception. Berlin: Springer; 2005.

[14]. Ao S-I, Rieger BB, Amouzegar MA. Machine learning and systems engineering. Dordrecht/ New York: Springer; 2010

[15]. Györfi L, Ottucsák G, Walk H. Machine learning for fi nancial engineering. Singapore/London: World Scientifi c; 2012.

[16]. Gong Y, Xu W. Machine learning for multimedia content analysis. New York/London: Springer; 2007.

[17]. Fielding A. Machine learning methods for ecological applications. Boston: Kluwer Academic Publishers; 1999

[18]. Mitra S. Introduction to machine learning and bioinformatics. Boca Raton: CRC Press; 2008.

[19]. Cleophas TJ. Machine learning in medicine. New York: Springer; 2013.

[20]. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning, pages 98–227. MIT Press, 2016. http://www.deeplearningbook.org.

[21]. Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

[22]. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning, pages 98–227. MIT Press, 2016. http://www.deeplearningbook.org.

[23]. Ayodele, T. O. (2010). Types of machine learning algorithms. In *New advances in machine learning*. IntechOpen.

[24]. Pacheco, F., Exposito, E., Gineste, M., Baudoin, C., & Aguilar, J. (2018). Towards the deployment of machine learning solutions in network traffic classification: a systematic survey. *IEEE Communications Surveys & Tutorials*, *21*(2), 1988-2014.

[25]. MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

[26]. Dubes, R. C., & Jain, A. K. (1988). *Algorithms for clustering data.*

[27]. Orhan, U., Hekim, M., & Ozer, M. (2011). EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. *Expert Systems with Applications*, *38*(10), 13475-13481.

[28]. Biswas, D., Cranny, A., Gupta, N., Maharatna, K., Achner, J., Klemke, J., ... & Ortmann, S. (2015). Recognizing upper limb movements with wrist worn inertial

sensors using k-means clustering classification. *Human movement science*, *40*, 59-76.

[29]. Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., & Weerawarana, S. (2002). Unraveling the Web services Web: an introduction to SOAP, WSDL, and UDDI. *IEEE Internet computing*, *6*(2), 86-93.

[30]. W3C Working Group. (2004). Web Services Architecture Requirements. *URL http://www. w3. org/TR/wsa-reqs*.

[31]. Vicuña, J. N., Castillo, F. R., Urgilés, F. L. E., & Ríos, J. M. (2017). Raspberry analysis in the teaching of computer sciences. *International Journal of Applied Engineering Research*, *12*(7), 1182-1189.

[32]. Ríos, J. M., Mora, N. L., Ordóñez, M. P. Z., & Sojos, E. L. (2016). Evaluación de los Frameworks en el Desarrollo de Aplicaciones Web con Python. *Revista latinoamericana de Ingenieria de Software*, *4*(4), 201-207.

[33]. Cumba Armijos, P. D., & Barreno Pilco, B. A. (2013). *Análisis de PYTHON con Django frente a Ruby on Rails para desarrollo ágil de aplicaciones Web. Caso práctico: DECH* (Bachelor's thesis).

[34]. Carrazza, S., Ferrara, A., Palazzo, D., & Rojo, J. (2015). APFEL Web: a Web-based application for the graphical visualization of parton distribution functions. *Journal of Physics G: Nuclear and Particle Physics*, *42*(5), 057001.

[35]. US Department of Health and Human Services. (2004). Guidance for industry: PAT—A framework for innovative pharmaceutical development, manufacturing, and quality assurance. *Food and Drug Administration, Rockville, MD*.

[36]. Read, E. K., Park, J. T., Shah, R. B., Riley, B. S., Brorson, K. A., & Rathore, A. S. (2010). Process analytical technology (PAT) for biopharmaceutical products: Part I. Concepts and applications. *Biotechnology and bioengineering*, *105*(2), 276-284.

[37]. Mhatre, R., & Rathore, A. S. (Eds.). (2009). *Quality by Design for Biopharmaceuticals: Principles and Case Studies*. Wiley.

[38]. Read, E. K., Shah, R. B., Riley, B. S., Park, J. T., Brorson, K. A., & Rathore, A. S. (2010). Process analytical technology (PAT) for biopharmaceutical products: Part II. Concepts and applications. *Biotechnology and bioengineering*, *105*(2), 285-295.

[39]. De Beer, T. R. M., Allesø, M., Goethals, F., Coppens, A., Vander Heyden, Y., Lopez De Diego, H., ... & Baeyens, W. R. G. (2007). Implementation of a process

analytical technology system in a freeze-drying process using Raman spectroscopy for in-line process monitoring. *Analytical chemistry*, *79*(21), 7992-8003.

[40]. van den Berg, F., Lyndgaard, C. B., Sørensen, K. M., & Engelsen, S. B. (2013). Process analytical technology in the food industry. *Trends in food science & technology*, *31*(1), 27-35.

[41]. Boicea, A., Radulescu, F., & Agapin, L. I. (2012, September). MongoDB vs Oracle--database comparison. In *2012 third international conference on emerging intelligent data and Web technologies* (pp. 330-335). IEEE.

[42]. Fisher, R. A. (2006). *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

[43]. MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

[44]. Chris Piech. K means. Retrieved May 22, 2019 from https://stanford.edu/~cpiech/cs221/handouts/kmeans.html, 2013.