Ciclo Autonómico de Análisis de Datos para el Diseño de Descriptores para Algoritmos de Aprendizaje Automático

Ricardo Vargas
Posgrado en Computación-CEMISID
Universidad de Los Andes
Mérida, Venezuela
ricardo.servitechs@gmail.com

Jose Aguilar
CEMISID Universidad de Los Andes
Universidad de Los Andes
Mérida, Venezuela
aguilar@ula.ve

Eduard Puerto
Grupo de Investigación GIDIS
Universidad Francisco de Paula Santander
Cúcuta, Colombia
eduardpuerto@ufps.edu.co

***Resumen*—Varios trabajos en la literatura han determinado que, para obtener buenos resultados con Algoritmos de Aprendizaje Automático, se requieren excelentes descriptores del fenómeno estudiado. En particular, para el proceso de reconocimiento de patrones es importante tener buenos descriptores. En ese sentido, en este trabajo se propone un ciclo autonómico de tareas de Analítica de Datos (AdD) que faciliten la obtención de descriptores para el proceso de reconocimiento de patrones. El ciclo autonómico provee las características /descriptores ideales a ser usado por los Algoritmos de Aprendizaje Automático en sus tareas de construcción de modelos de conocimiento (clasificadores, predictores, etc.). Los experimentos iniciales con el ciclo autonómico han mostrado resultados alentadores.**

Palabras Clave—Ciencias de Datos, Ingeniería de Características, Analítica de Datos, Aprendizaje Automático

I. Introducción

El área de Ingeniería de Características/Descriptores consiste en el proceso de búsqueda de descriptores en un conjunto de datos, y está compuesto por los procesos de extracción, construcción, reducción y/o selección de descriptores/características [1, 2, 3, 4]. La Ingeniería de Características es muy importante para los algoritmos de Aprendizaje Automático, también conocidos como Aprendizaje de Máquinas (o Machine Learning, en inglés), ya que la precisión de los mismos depende de la calidad de los descriptores. Existen un importante número de trabajos que han estudiado estos problemas individualmente, proponiendo diferentes técnicas específicas para cada uno de esos procesos [2, 5, 6, 7], y, además, normalmente enfocándose en específicas áreas de aplicación, como procesamiento de imágenes [2,8], análisis de tráfico de redes [5], o análisis de identidad urbana y puntos de interés en ciudades [6,7].

Por otro lado, la analítica de datos (AoD o Data analytics, en inglés) es un área que comienza a tener cierto grado de madurez, con un importante número de aplicaciones en diferentes ámbitos [7, 9]. Ahora bien, la utilización de la AdD se ha entendido como un proceso de diseño aislado (una tarea para un problema específico), y los trabajos actuales no consideran la integración de un conjunto de tareas de analítica de datos para resolver problemas complejos.

Recientemente, se ha propuesto el concepto de Ciclo Automático de Tareas de AdD para el ámbito de aprendizaje [9, 10, 11, 12], para organizar los diferentes tipos de tareas de Análisis de Datos que se integran en ese entorno, para alcanzar diferentes objetivos de aprendizaje (por ejemplo, para optimizar las condiciones ambientales, o mejorar el proceso de enseñanza-aprendizaje). Un ciclo autónomo es un ciclo cerrado de tareas de análisis de datos, que supervisa constantemente el proceso bajo estudio, tal que las tareas de análisis de datos tienen diferentes roles: observar el proceso, analizarlo, y tomar decisiones.

En este trabajo se propone la construcción de un ciclo autonómico de tareas de AdD para el proceso de ingeniería de características, basándonos en la metodología MIDANO [13, 14]. MIDANO es una metodología para desarrollar tareas de AdD, que parte por el análisis exhaustivo del ámbito de estudio, para determinar en qué procesos es posible extraer conocimiento desde los datos.

En particular, por medio del ciclo autonómico se busca simplificar el proceso de ingeniería de características, subdividiéndolo en tareas de AdD más simples, las cuales son fáciles de implementar y evaluar. Cada tarea es definida por un grupo de técnicas de minería, requeridas para el proceso de ingeniería de características, según el problema específico que se esté considerando en un momento determinado.

Así, este trabajo es del ámbito de las ciencias de los datos, y en particular, de la Ingeniería de Características, por lo cual se organiza de la siguiente manera. En la Sección II se ampliará el contexto teórico, y se presentan los trabajos de referencias para nuestra propuesta. La Sección III introduce el ciclo autonómico (CA, por sus siglas en español) propuesto, sus tareas y técnicas que lo componen, y el modelo de datos que lo acompaña. La Sección IV detalla la implementación de este CA y se presenta el caso de estudio para probarlo. En la Sección V se realizan experimentos, y en la siguiente Sección se presentan las conclusiones y trabajos futuros.

II. Contexto Teórico

A. *Ingeniería de Características*

Al realizar procesos de reconocimiento de patrones, ya sea mediante métodos supervisados, semi-supervisados o no supervisados, uno de los elementos más importantes a tener en cuenta son los descriptores o características que se usan para representar el fenómeno a estudiar. Estos descriptores son de suma importancia, ya que la precisión del modelo depende de que los descriptores representen lo mejor posible los objetos a reconocer, además de que también repercuten en los recursos, tiempo, entre otras cosas, requeridos por los modelos.

Es por esto que existen diferentes tipos de técnicas usadas para procesar los datos del experimento, y obtener los descriptores que se usarán en el proceso de reconocimiento. A esta fase se le denomina *Ingeniería de Características*, y consiste en la extracción, construcción, reducción y/o selección, de descriptores/características desde los datos. Cada una usa diferentes técnicas de minería y tiene diferentes objetivos [1, 2, 3, 4 y 15]. A continuación, se describen los procesos principales de la *Ingeniería de Características* considerados en este trabajo.

1) Extracción de Características (Feature Extraction, FE): Estas técnicas buscan identificar los descriptores que mejor describan el fenómeno estudiado desde los datos disponibles, que, a su vez, sean de mayor utilidad para construir los modelos de conocimiento. Para ello, se aplican transformaciones sobre los datos usando funciones específicas. Estas funciones varían dependiendo de distintos criterios, como el tipo de variable/característica sobre la cual se implementará, el tipo de algoritmo de aprendizaje a usar, etc. Ahora bien, los criterios más comunes a usar para escoger las funciones de transformación son: a) maximizar varianza o variación, es decir, buscar características que tomen valores diferentes en cada instancia, ya que características que puedan tomar los mismos valores en diferentes instancias no aportarían ningún valor como discriminantes, b) reducir correlaciones o evitar características redundantes, usualmente lográndolo por medio de funciones que reducen la dimensionalidad entre los datos. Existen diferentes tipos de funciones que se pueden usar, dependiendo del tipo de datos al cual se aplicará. En este tipo de técnicas se puede incluir el Análisis de Componentes Principales (PCA, por sus siglas en inglés), el cual es una de las más comunes [2, 3, 4, 6], y busca reducir la dimensionalidad al proyectar valores en componentes que mejor reflejen la información de los datos originales, con mayor independencia entre ellos.

En general, las características/descriptores pueden ser:

a) Características Estadísticas [5,8]*:* Son características que se extraen de datos de tipo numéricos, mediante valores estadísticos como la media, mediana, moda, desviación estándar, etc.

b) Basados en Grafos [5,8]*:* Son características que usan algún tipo de representación en grafos, de tal manera de poder aplicar la teoría de grafos. En general, los grafos pueden representar muchos problemas importantes, como redes sociales, interacciones entre bacterias, redes de computadoras, etc. Sobre estas representaciones se pueden aplicar diferentes técnicas de la teoría de grafos, para determinar cuáles son los mejores descriptores. Por ejemplo, la técnica de detección de comunidades, en la cual se busca agrupar nodos fuertemente conectados en clusters, mientras que los nodos que están en diferentes clusters son los que están más débilmente conectados. Estos tipos de técnicas varían según si se aplican a grafos dirigidos, no dirigidos, con pesos, sin pesos, etc. En general, se pueden usar métricas como la distancia, centralidad, densidad, además de la ya mencionada, entre otras

c) Espectrales o Basadas en Series de Tiempo: Estos son tipos de características que se derivan de datos secuenciales o continuos, como por ejemplo una serie de eventos que ocurren unos detrás de otros [14]. Se puede tomar en cuenta la periodicidad, magnitud del espectro, variaciones temporales, etc., y se usan técnicas como la transformada de Fourier [15].

En general, las técnicas usadas en "Extracción de Características" se pueden clasificar en:

a) Métodos Lineales: Usualmente, los problemas de clasificación pueden resolverse mediante separaciones lineales en un plano o hiperplano, es decir, las clases pueden delimitarse fácilmente "trazando" líneas que las separen. Entre estas técnicas se encuentran la ya mencionada PCA, y también el Análisis de Discriminantes Lineales (LDA, por sus siglas en inglés), el cual es un método que también busca reducir dimensionalidad, pero a diferencia de PCA que es no supervisado (no toma en cuenta clases), LDA es supervisado y funciona buscando proyecciones óptimas que maximicen las distancias entre las clases y minimicen distancias de los datos dentro de esas clases.

b) Métodos no Lineales: Cuando la clasificación es más compleja y las clases no son linealmente separables, es decir, los límites entre ellas son discontinuos, se utilizan métodos no lineales, para realizar esta separación de clases. Entre estos métodos pueden usarse técnicas como realizar una transformación de los datos a dimensiones más altas, donde sí sea posible realizar una separación lineal para aplicar un modelo lineal. Veamos el ejemplo de la Figura 1 de datos no linealmente separados.

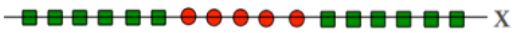

Figura 1. Método Lineal

En la Figura 1 se aprecian dos clases (representadas por los cuadros verdes y círculos rojos), los cuales no están linealmente separados. Se puede realizar una transformación a dos dimensiones mediante la función $X \rightarrow \{X, X2\}$, resultando en un doble valor por cada elemento del ejemplo original, que al representarlo en un plano se puede apreciar que sí es linealmente separable en esa nueva representación (ver Figura 2). Entre las técnicas que realizan esto están las redes

neuronales multicapa, k-NN, máquinas de soporte vectoriales no lineales (SVM por sus siglas en inglés), etc.

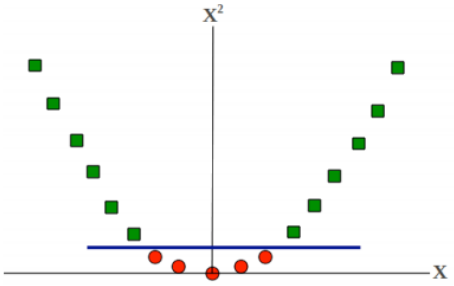

Figura 2. Partición de Datos

Existen algunos trabajos recientes en el área de extracción de características, como [16, 17]. En [16] se propone un enfoque para reconocimiento facial bajo condiciones de variación de luz. Este enfoque consta de tres fases, la primera en la cual se realiza una normalización de los componentes de iluminación. En segundo lugar, se realiza la extracción de descriptores. Para esto, los autores usan dos grupos de técnicas: holísticas y locales. Entre las técnicas holísticas se encuentran PCA y LDA, y entre las locales los métodos de patrones binarios locales (LBP, por sus siglas en inglés) o patrones direccionales locales (LDP, por sus siglas en inglés). Finalmente, en la tercera fase usan SVM para realizar la clasificación. En [17] se propone un proceso de extracción de características en textos, el cual se divide a nivel sintáctico y a nivel semántico. A nivel sintáctico proponen una versión de $\chi 2$ (Chi-squared statistics) mejorada (ICHI), aplicada sobre una matriz de palabras, y a nivel semántico una técnica llamada asignación latente de Dirichlet, sobre una representación de documentos por tópicos. Finalmente, se realiza una fusión entre ambos resultados, para obtener el conjunto de características final.

2) Construcción de Características (Feature Construction, FC): Estos tipos de técnicas buscan generar nuevas características [1, 4]. Su uso principal es cuando se tienen datos incompletos, y se desea buscar la información faltante que complemente la descripción del objeto, o la relación que existe entre las características que ya se tienen. Esta nueva información se puede obtener por medio de inferencia o creación de nuevos datos, pero siempre usando los datos ya existentes para derivarla [4], buscando incrementar el poder de expresión de los datos ya existentes. Algunas de las técnicas que se podrían usar para generar estos nuevos datos son las que aplican algún operador sobre estos datos, como por ejemplo operadores algebraicos en datos numéricos.

En cuanto a propuestas de construcción de características, el trabajo [18] propone un método basado en programación genética (GP por sus siglas en inglés), para obtener características faltantes en datos incompletos. En esta técnica se utilizan funciones de intervalos como parte del conjunto de funciones del GP, resultando en una sustitución del valor faltante por un intervalo de posibles valores, o si es un valor no faltante, se sustituye por un intervalo que incluye el valor original. En [19] se propone un proceso de construcción de características basadas en entropía, para la detección de ataques en redes de comunicación. Este proceso consta de tres partes: primero, se extraen características de cada cabecera de los paquetes de red. Datos como direcciones IP origen y destino, puertos origen y destino, y protocolos, son los más comúnmente usados. En segundo lugar, se calcula la entropía de Shannon, usando un intervalo de tiempo específico, y, por último, se construyen nuevas características con variaciones de entropía, usando combinaciones de características generadas en los pasos previos.

Selección de Características (Feature Selection, FS): Con estos tipos de técnicas, al contrario de las descritas antes, se busca reducir el conjunto de datos a utilizar. Como lo definen Blum y Langley, citado en [1], "se diferencia de las transformaciones de características en que no se generan nuevas, si no que se selecciona un subconjunto de ellas". Esta reducción puede ser realizada ya sea mediante una selección entre las características disponibles, o mediante la agregación/fusión de ellas. La selección se puede realizar [4] generando subconjuntos aleatoriamente, incrementalmente, etc. En el caso incremental, podría iniciarse con un subconjunto vacío, y se van agregando características una a una; o empezar con el conjunto completo, e ir eliminando características. Se suelen usar diferentes criterios para evaluar si un subconjunto generado es óptimo o no, estos criterios se pueden agrupar dentro de dos grandes modelos: Filtros (en adelante Filtering) y Envoltorios (en adelante Wrapping) [15].

Los Filtering constan de criterios que son independientes del modelo de aprendizaje que se usará luego. Es una aproximación eficiente, pero debido a esta separación del proceso de aprendizaje posterior, puede descartar información que hubiera sido útil para el mejor rendimiento del algoritmo de aprendizaje usado. El proceso usado suele realizarse en dos pasos; primero la aplicación del criterio de selección, entre las cuales pueden estar: dependencia de una característica de su clase, correlaciones de característica-característica, característica-clase, medidas de distancia, medidas de consistencia, etc. En el segundo paso, simplemente se selecciona el subconjunto con mejor ranking.

En el modelo de Wrapping sí se toma en cuenta el algoritmo de aprendizaje a usar, es decir, se mide el rendimiento obtenido por ese algoritmo, usando el subconjunto de características seleccionadas. Ese proceso se suele realizar usando una estrategia de búsqueda entre las características o subconjunto de características, las cuales se pasan al algoritmo para medir su rendimiento, y este resultado se vuelve a pasar al componente de búsqueda en un proceso iterativo, hasta que se selecciona el conjunto de características con el mejor rendimiento. Para evitar una búsqueda exhaustiva se pueden usar distintas estrategias como algoritmos genéticos, Best-first, Hill-climbing, etc.

Existe un tercer modelo que es una especie de unión entre Filtering y Wrapping para obtener lo mejor de ambos mundos [20, 21]. Esto porque Filtering no toma en cuenta el algoritmo de aprendizaje a usar, pudiendo descartar características útiles

para ese algoritmo, y Wrapping debe evaluar el rendimiento de los subconjuntos de característica en el modelo, haciéndolo bastante costoso computacionalmente. Este tercer modelo, llamado Embedding, busca incrustar el proceso de selección en la construcción del modelo de aprendizaje, logrando mejorar el costo computacional, ya que no es necesario correr el proceso de aprendizaje repetidas veces.

Finalmente, en el área de selección de características se pueden mencionar algunos trabajos, tales como [22], que propone un enfoque de selección de características para previsión de precios y cargas en sistemas de suministro eléctrico. Este enfoque mezcla características de modelos de Filtering y Wrapping. De este modo, el método propuesto selecciona un subconjunto de características tomando en cuenta criterios como relevancia, redundancia e interacción entre características, considerando el algoritmo de aprendizaje que se usará. En [23] se proponen mejorar la identificación de patologías del pecho mediante el análisis de radiografías. En este enfoque se extraen características de estas imágenes radiológicas usando redes neuronales convolucionales, luego, establecen una fase de reducción de características realizando una combinación de ellas, por medio de una técnica de fusión lineal ponderada basada en los pesos de probabilidad de cada clase. Finalmente, realizan una fase de selección de características, en la cual usando pruebas de Kruskal-Wallis sobre la varianza de los datos, determinan las características más informativas.

B. *MIDANO*

MIDANO es una metodología para el Desarrollo de Aplicaciones de Minería de datos (MD) basada en el Análisis Organizacional. MIDANO es diseñada para el desarrollo de aplicaciones de Minería de Datos para un proceso de cualquier empresa o institución, sin embargo, esta puede ser utilizada en procesos de AdD [13,14]. MIDANO está compuesta por tres fases (ver Figura 3).

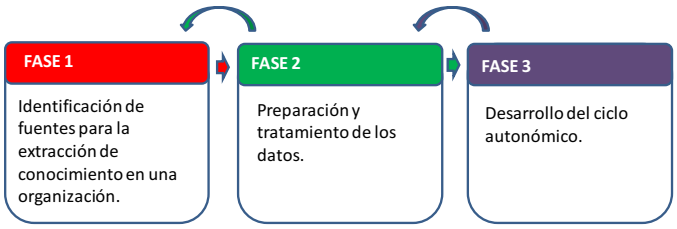

Figura 3. MIDANO

Fase 1. Identificación de Fuentes para la Extracción de Conocimiento en una Organización: Esta fase tiene como finalidad realizar un proceso de ingeniería de conocimiento, orientado a organizaciones/empresas, de las cuales no se conoce o se tiene poca información del (de los) problema(s) o los procesos a estudiar. El principal objetivo de esta fase es conocer la organización, sus procesos, sus expertos, entre otros aspectos, para definir el objetivo de la aplicación de AdD en la organización.

Fase 2. Preparación y Tratamiento de los Datos: Para aplicar AdD sobre un problema en específico, es necesario contar con un historial de datos asociado al problema de estudio. Esto conlleva realizar distintas operaciones con los datos, con la finalidad de prepararlos. Ese proceso se basa en el paradigma ETL (por sus siglas en inglés): extracción de los datos desde sus fuentes, transformación de los datos, y carga de los mismos en el almacén de datos del CA. Para realizar este proceso se crea una vista minable, que básicamente contiene información sobre las variables y sus históricos. En específico, se crea una vista minable conceptual (VMC), que detalla cada una de las variables a ser tomadas en cuenta para las tareas de AdD. La misma está compuesta por la descripción de todas las variables de interés, y algunos campos adicionales de importancia para realizar el proceso de tratamiento de datos (por ejemplo: dependencias con otras variables, transformaciones a realizar, entre otras características). Con esta VMC se crea el modelo de datos a ser usado por el medio de almacenamiento (almacén de datos). Finalmente, el medio de almacenamiento es cargado con los datos. Al medio de almacenamiento cargado con los datos, lo llamaremos vista minable operativa. Así, en esta fase se construye el modelo de datos requerido por el ciclo autonómico de tareas de AdD. Por lo tanto, en esta fase se realiza la preparación y tratamiento adecuado de los datos, que serán utilizados por el ciclo autonómico de AdD.

Fase 3. Desarrollo del Ciclo Autonómico de Tareas de AdD: En esta fase se implementan las tareas de AdD. Así, esta fase tiene como objetivo implementar las diferentes tareas de AdD del ciclo autonómico, que generan los modelos de conocimientos requeridos (por ejemplo, modelos predictivos, modelos descriptivos, etc.). Esta etapa culmina con la implementación de un prototipo del ciclo autonómico. Para el desarrollo de las tareas de AdD, se puede usar cualquiera de las metodologías existentes de desarrollo de tareas de MD. Además, durante esta fase se realizan experimentos para validar los modelos de conocimiento generados por las tareas de AdD.

La utilización de esta metodología permite el desarrollo sistemático de aplicaciones de software especializado basada en técnicas inteligentes, para la extracción de conocimiento a partir de los datos almacenados en las bases de datos de cualquier industria o proceso.

III. Especificación de las Tareas de Análisis de Datos

En esta sección se describe el ciclo autonómico a ser utilizado, con sus respectivas tareas de AdD. El objetivo de este ciclo autonómico es buscar la obtención de las características óptimas que mejor describan el objeto estudiado, para su uso en un proceso de reconocimiento de patrones.

A. *Ciclo Autonómico*

El ciclo autonómico propuesto para este trabajo consta de dos etapas:

- Monitoreo: En esta etapa del ciclo autonómico se realiza la recolección de los datos desde sus fuentes.
- Análisis y toma de decisiones: Se ejecutan las tareas de Análisis de Datos sobre los datos obtenidos en la etapa anterior, para determinar los descriptores que se usarán en la clasificación.

Las tareas de AdD que se aplicarán en cada etapa se describen en la Tabla I.

TABLA I. ESPECIFICACIÓN DE TAREAS DE ADD PARA EL CICLO AUTONÓMICO

Tarea	Nombre	Fuentes generales de datos requeridas	Indicadores generados	Efectos esperados sobre el objetivo estratégico
Monitoreo	Captura de datos	Tablas VMC, ETL (Extracción, Tratamiento y Carga) y CCA (colección, curetaje y agregación)	Datos del Experimento	Se obtienen los datos recogidos en etapas anteriores sobre los cuales se quiere realizar la clasificación
Análisis y Toma de decisiones	Construcción de características	Datos obtenidos del paso anterior	Datos tratados y depurados	Se emplean los primeros métodos y técnicas de preparación y tratamiento de datos
	Extracción de características	Datos depurados	Medias, medianas, modas, mínimos, máximos, entre otros valores numéricos necesarios	Conjunto de técnicas para extraer valores numéricos, métricas, etc. que mejor representen los datos
	Selección y reducción de características	Representación numérica de los datos	Conjunto final de características	Se terminan de depurar las características extraídas, reduciendo descriptores redundantes, o descartando algunas carac excedentes

En las siguientes tablas (Tablas II al V), se especifican en detalle cada una de estas tareas:

TABLA II. DETALLES DE LA TAREA DE CAPTURA DE DATOS

Nombre de la tarea:	Captura de datos
Descripción	Obtener los datos del caso de estudio sobre el que se desea realizar la clasificación.
Fuente de datos	Tablas de VMC, ETL, CCA
Tipo de tarea de analítica de datos	Descubrimiento
Tipo de modelo de conocimiento	Descriptivo
Tareas relacionadas de analítica de datos	Construcción de características
Tipo de tarea del ciclo (rol)	Monitoreo

TABLA III. DETALLES DE LA TAREA DE FC

Nombre de la tarea:	Construcción de características
Descripción	Preparar los datos obtenidos
Fuente de datos	Datos obtenidos en la etapa anterior
Tipo de tarea de analítica de datos	Clasificación, Agrupamiento, Asociación
Técnicas de analítica de datos	Normalización, estandarización, transformación, reducción, discretización, etc.
Tipo de modelo de conocimiento	Descriptivo, Predictivo
Tareas relacionadas de analítica de datos	Captura de datos Extracción de características
Tipo de tarea del ciclo (rol)	Análisis / Toma de decisiones

TABLA IV. DETALLES DE LAS TAREAS DE FE

Nombre de la tarea:	Extracción de características
Descripción	Obtener representaciones numéricas de los datos procesados para usarlas luego en el proceso de clasificación
Fuente de datos	Datos depurados
Tipo de tarea de analítica de datos	Clasificación, Agrupamiento, Asociación, Regresión
Técnicas de analítica de datos	Cálculos estadísticos, KNN, random forest, regresión, grafos, etc.
Tipo de modelo de conocimiento	Descriptivo, Predictivo
Tareas relacionadas de analítica de datos	Construcción de características Selección y reducción
Tipo de tarea del ciclo (rol)	Análisis / Toma de decisiones

B. *Modelo de Datos*

a) Vista Minable Conceptual para el CA: Los datos se obtienen desde las Tablas generadas por el proceso que se está examinando, principalmente de su propia VMC y Tablas ETL y CCA. Además de estos datos, también se requieren otras variables derivadas de cada tarea del ciclo autonómico. Una posible VMC podría ser la Tabla VI.

TABLA V. DETALLES DE LAS TAREAS DE FS

Nombre de la tarea:	Selección y reducción de características
Descripción	Asegurar que se obtienen solo las características más relevantes para realizar la clasificación

Fuente de datos	Características extraídas en la etapa anterior
Tipo de tarea de analítica de datos	Clasificación, Agrupamiento, Asociación
Técnicas de analítica de datos	Filtering, Wrapping, análisis bivariantes, etc.
Tipo de modelo de conocimiento	Descriptivo, Predictivo
Tareas relacionadas de analítica de datos	Extracción de características
Tipo de tarea del ciclo (rol)	Análisis / Toma de decisiones

TABLA VI. POSIBLE TABLA VMC PARA EL CICLO AUTONÓMICO

Variable	Descripción	Procedencia	Descripción
variable_vmc	Variables de la VMC del proceso estudiado	Vmc	Estos son los datos pertenecientes a la tabla VMC sobre la cual se realiza el análisis la cual posee los datos de entrada del caso de uso.
descripcion_vmc	Descripción de las variables de la VMC	Vmc	
procedencia_vmc	Procedencia de las variables de la VMC	Vmc	
observaciones_vmc	Observaciones de las variables de la VMC	Vmc	
variable_etl	Variables de la tabla ETL del proceso estudiado	Etl	Estos son los datos de la tabla etl la cual posee los datos con un primer pre-procesamiento sobre los datos de entrada.
extraccion_etl	Fuente de datos de donde fueron extraídas	Etl	
transformacion_etl	Procesos de pre-procesamiento realizados en ellas	Etl	
carga_etl	Dimensión del modelo donde irán	etl	
variable_cca	Variables de la tabla CCA del proceso estudiado	cca	Datos curados sobre fuentes de datos externas.
coleccion_cca	Fuente externas de datos de donde fueron extraídas	cca	
curacion_cca	Procesos de pre-procesamiento realizados en ellas	Cca	
analisis_cca	Criterios de calidad y dimensión donde irán	cca	
medias	Cálculos de la media sobre datos de entrada	Datos de entrada	Datos estadísticos calculados a partir de los datos curados de entrada necesarios para aplicar tareas de AdD para extracción de características.
medianas	Cálculos de la mediana sobre datos de entrada	Datos de entrada	
modas	Cálculos de la moda sobre datos de entrada	Datos de entrada	
desviacion_estandar	Cálculos de la moda sobre datos de entrada	Datos de entrada	

b) Modelo de Datos Multidimensional: De la VMC propuesta, se obtiene el siguiente modelo multidimensional con las variables agrupadas por temas (ver Figura 4).

Cada dimensión de la tabla multidimensional corresponde a los datos de las diferentes tablas usadas para el análisis. Una dimensión tiene la información de la tabla ETL, otra de las tablas VCM, CCA, y finalmente, una contiene los datos estadísticos generados a partir de los datos del caso de estudio. El contenido de las Tablas ETL y CCA contienen los datos de los procesos de preparación de datos, y la de VCM los detalles de las variables que conforman la fuente de datos (ver [14,15] para más detalles).

IV. IMPLEMENTACIÓN DEL CA

En esta sección, vamos a dar un ejemplo de instanciación de nuestro CA.

A. Flujo Asociado al CA

Para la implementación del CA, en primer lugar, se define el conjunto de etapas del flujo asociado al CA, a ser usadas al instanciar el modelo propuesto con datos reales del caso de estudio seleccionado. Estas etapas se muestran en la Tabla VII.

TABLA VII. ETAPAS PARA EL CICLO AUTONÓMICO

Etapa	Detalle	Tipo de Tarea	Producto
1	Captura de datos	Descubrimiento	Datos del Experimento
2	Aplicar técnica de construcción de características	Cálculos, inferencias, procedimientos que producen nuevos datos	Conjunto de características extendido
3	Aplicar transformaciones básicas en los datos	Transformación	Primer conjunto de Características con valores numéricos usables en técnicas de FE
4	Aplicar técnicas de extracción de características (FE)	Transformación, Clasificación, Clustering, etc.	Conjunto de características procesadas
5	Aplicar técnicas de selección	Filtrado, Agrupamiento, Unión, etc.	Conjunto final de características

Estas etapas están basadas en las tareas propuestas para el CA, por lo que en la Tabla se puede ver que cada paso corresponde a una de estas tareas. Además, se especifica el(los) tipo(s) de tarea(s) de análisis de datos que se podrían aplicar, y los datos que se producen en cada paso, siendo el resultado final del flujo, el conjunto de características óptimas seleccionadas.

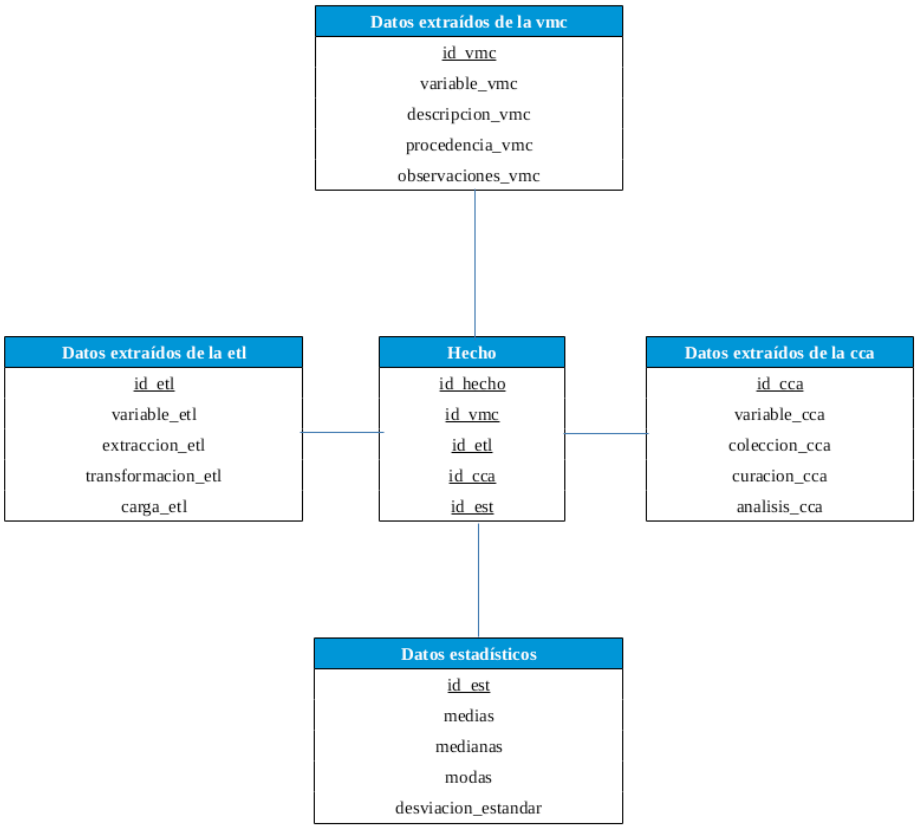

Figura 4. Modelo Multidimencional

B. *Caso de Estudio*

El caso de estudio seleccionado para esta implementación consiste en una base de datos [24] de gestos de letras escritos a mano en una tableta digital, por niños que se encuentran aprendiendo a escribir. Esta base de datos consiste en 6118 registros con diferentes características, obtenidas al momento en el que el niño traza la letra. Entre los datos obtenidos se encuentran [25]:

- Repeats: número de veces que el niño ha intentado escribir la letra
- Class: la letra escrita
- ID: id del niño
- Gender: genero del niño (masculino/femenino)
- Age: edad del niño
- Laterality: lateralidad del niño (zurdo/derecho)
- Duration: el tiempo en ms que el niño tardó en escribir la letra
- letraNbStrokes: número de trazos usados para escribir la letra
- AveragePressure: presión promedio aplicada por el niño al realizar el gesto
- VariancePressure: varianza de la presión aplicada por el niño al realizar el gesto

Se usará esta base de datos para instanciar el CA, aplicando las técnicas correspondientes a cada tarea, para determinar las características que mejor representen las letras dibujadas.

V. Experimentos

Para realizar los experimentos se usará la librería del lenguaje de programación Python, *scikit-learn*, que cuenta con un gran número de implementaciones de algoritmos para ciencia de los datos, lo cual facilita la experimentación, permitiendo concentrarse solo en los posibles resultados del CA.

A. *Especificación de los Pasos*

a) *Captura de Datos:* En este paso, simplemente se cargan los datos iniciales necesarios para comenzar el proceso de ingeniería de características. En este caso, son los datos obtenidos de INTUIDOC, como se muestra en la Figura 5 [24, 25]:

	A	B	C	D	E	F	G	H	I	J	K
1	Repeats	Class	ID	Gender	Age	Laterality	Duration	NbStrokes	AveragePressure	VariancePressure	SavedScoreGlobal
2	1	d	anonstudent1475048982665	BOY	ND	RIGHT	1300	1	0,580593467	0,127687503	0,619901053
3	1	n	anonstudent1475048868521	BOY	ND	RIGHT	1514	1	0,932148755	0,165912816	1
4	2	n	anonstudent1475048868521	BOY	ND	RIGHT	1299	1	0,897993743	0,182596353	1
5	1	u	anonstudent1475048868521	BOY	ND	RIGHT	1567	1	0,955063939	0,125530824	1
6	1	d	anonstudent1475049326525	GIRL	ND	RIGHT	773	1	0,620547235	0,09371386	1
7	1	u	anonstudent1475049326525	GIRL	ND	RIGHT	840	1	0,568899095	0,083730014	0,83106493
8	1	n	anonstudent1475049326525	GIRL	ND	RIGHT	2690	1	0,7090469	0,112996955	1
9	1	u	anonstudent1475048868521	BOY	ND	RIGHT	2609	1	0,933333337	0,109634476	1
10	1	j	anonstudent1478764526427	BOY	6	RIGHT	4130	2	0,898688734	0,116078498	1
11	2	j	anonstudent1478764526427	BOY	6	RIGHT	3293	2	0,886679053	0,104820331	1
12	3	j	anonstudent1478764526427	BOY	6	RIGHT	3285	2	0,881134212	0,132600282	1
13	1	m	anonstudent1478763974290	BOY	ND	RIGHT	1157	1	0,666666687	0,106140371	1
14	2	m	anonstudent1478763974290	BOY	ND	RIGHT	1488	1	0,772387564	0,095200617	0,74480688
15	1	v	anonstudent1478764526427	BOY	6	RIGHT	2238	1	0,542212009	0,063159453	0,840620845
16	2	v	anonstudent1478764526427	BOY	6	RIGHT	2826	1	0,501985133	0,061357192	0,921001711
17	3	v	anonstudent1478764526427	BOY	6	RIGHT	2276	1	0,535986364	0,061060099	1

Figura 5. Datos del Experimento Obtenidos de INTUIDOC

b) *Aplicar Técnicas de Construcción de Características:* Una vez obtenidos estos datos, se aplican las primeras técnicas de construcción de características. Estos nuevos datos con los que se expandirá el espectro de características disponibles, son datos estadísticos generados desde los datos del experimento. Para este experimento se calcula la media del tiempo de duración que tomó el niño al dibujar la letra, y el número de veces promedio que los niños escribieron cada letra, además de sus desviaciones estándares, agregando así nuevas variables al conjunto total de ellas.

c) *Aplicar Transformaciones Básicas en los Datos:* Como primer paso de las tareas de extracción de características, se realizan algunas transformaciones básicas necesarias para obtener datos más relevantes. En este experimento se puede notar que, de los datos disponibles, los datos de Class, Gender y Laterality no son óptimos, ya que se encuentran en formato de cadenas de texto, por lo que se realiza una transformación sobre ellos, para obtener un formato numérico, más útil para usar otras técnicas. Los resultados de esta transformación se muestran en la Figura 6, en donde ya todos los valores son numéricos.

```
      Repeats  Class  Gender  Laterality  Duration  NbStrokes  \
0           1      4       0           1      1300          1
1           1     14       0           1      1514          1
2           2     14       0           1      1299          1
3           1     21       0           1      1567          1
4           1      4       1           1       773          1
5           1     21       1           1       840          1
6           1     14       1           1      2690          1
7           1     21       0           1      2609          1
8           1     10       0           1      4130          2
9           2     10       0           1      3293          2
10          3     10       0           1      3285          2
11          1     13       0           1      1157          1
```

Figura 6. Transformación de los Datos a Valores Numéricos

d) *Aplicar Técnicas de Extracción de Características:* En este caso, se aplican dos técnicas de extracción de características. La primera es Random Forest (RF), con la cual se determinan los pesos de las características disponibles, y así la relevancia de cada una de ellas. Esta técnica consiste en generar un conjunto de árboles de decisión, cada uno usando aleatoriamente un subconjunto de las características disponibles como nodos, y se van asignando pesos a cada característica de acuerdo a que tan relevante fue en el resultado del árbol al cual pertenecía [5]. La segunda técnica a aplicar es PCA. Así, se definieron tres escenarios, uno donde solo se usa RF, en otro solo PCA, y en el tercero se aplica una mezcla de ambos, aplicando RF a los datos obtenidos al aplicar PCA.

e) *Aplicar Técnicas de Selección de Descriptores:* Finalmente, se aplican técnicas de selección y reducción sobre este conjunto de datos final, para descartar las características menos relevantes. Para este experimento se usan diferentes técnicas proporcionadas por scikit learn. La primera es *Recursive Feature Elimination* (RFE), la cual realiza la selección considerando recursivamente conjuntos de características cada vez más pequeños, seleccionadas basadas en sus importancias calculadas en el paso anterior, hasta conseguir el número de características deseadas (ver Figura 7). En segundo lugar, se usa la técnica *VaraianceThreshold*, la cual elimina las características cuya varianza no alcancen un valor de umbral propuesto, para lo cual se prueba con distintos valores de umbrales calculados con diferentes porcentajes de varianza aceptable para las características.

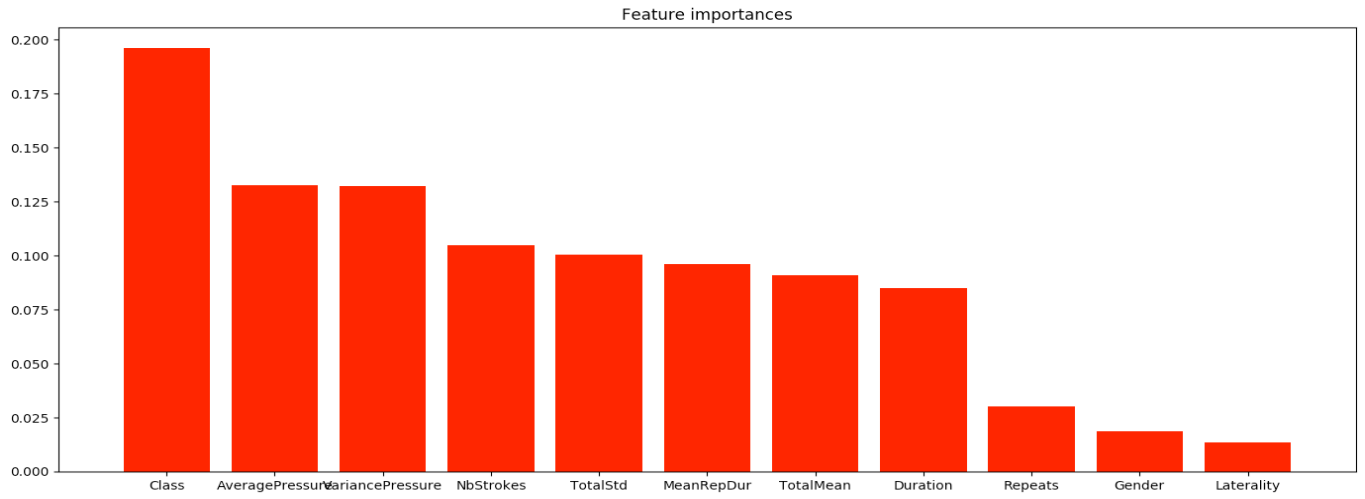

Figura 7. Importancia de Descriptores con Random Forest

B. Resultados

En esta sección se mostrarán los resultados obtenidos al aplicar las técnicas descritas previamente, y se usarán las características finales obtenidas en un algoritmo de clasificación, para estudiar y comparar sus comportamientos.

La Figura 7, muestra el resultado de aplicar RF para extracción de características, en la cual se muestran las características ordenadas por su importancia. Luego de calcular estas importancias, se utilizan para realizar la selección a RFE. Este proceso se realiza un par de veces, para calcular con esta técnica las características óptimas, obteniéndose dos listas: la primera con 8 características: "Class, Duration, NbStrokes, AveragePressure, VariancePressure, MeanRepDur, TotalMean, TotalStd"; y la segunda con 6 características "Class, NbStrokes, AveragePressure, VariancePressure, TotalMean, TotalStd".

Por otro lado, se aplica la técnica de VarianceThreshold, seleccionando las características con varianzas de 50%, 80% y 90%.

Para validar los resultados obtenidos, se usa nuevamente la técnica de RF, en este caso como clasificador, para verificar la precisión de la clasificación de las letras usando los diferentes grupos de características obtenidas por las diferentes técnicas y estrategias definidas en el CA. Para calcular la métrica de precisión se usó la técnica de validación "k-fold cross", con k = 10, tal que el 90% de los datos son usados para entrenamiento y 10% para probar. Estos resultados se muestran en la Tabla VIII, donde los nombres para indicar las técnicas de selección de descriptores son:

- DNP: Datos no procesados, no se aplicó ninguna selección, se usaron todas las características obtenidas hasta el paso de FE.
- RFE_6: Aplicar RFE obteniendo 6 características óptimas.
- RFE_8: Aplicar RFE obteniendo 8 características óptimas.
- RFE_6_V_50: Aplicar RFE obteniendo 6 características óptimas, y luego VarianceThreshold con umbral de 50% de varianza.
- RFE_6_V_80: Aplicar RFE obteniendo 6 características óptimas y luego VarianceThreshold con un umbral de 80% de varianza.
- RFE_6_V_90: Aplicar RFE obteniendo 6 características óptimas y luego VarianceThreshold con un umbral de 90% de varianza.
- RFE_8_V_50: Aplicar RFE obteniendo 8 características óptimas y luego VarianceThreshold con un umbral de 50% de varianza.
- RFE_8_V_80: Aplicar RFE obteniendo 8 características óptimas y luego VarianceThreshold con un umbral de 80% de varianza.

TABLA VIII. COMPARACIÓN DE RESULTADOS

Técnica FS vs FE	RF	PCA	RF + PCA
DNP	0.8155	0.8109	0.8008
RFE_6	0.8097	0.6471	0.8187
RFE_8	0.8090	0.8017	0.8090
RFE_6_V_50	0.8173	0.6023	0.8109
RFE_6_V_80	0.8133	0.6068	0.8068
RFE_6_V_90	0.8203	0.6221	0.8078
RFE_8_V_50	0.8155	0.7848	0.8080
RFE_8_V_80	0.8157	0.7887	0.8042

RFE_8_V_90	0.8090	0.7936	0.8094

La Tabla VIII muestra los valores obtenidos después de aplicar las diferentes técnicas de extracción de características y selección de descriptores usando las métricas de validación descritas anteriormente. Los valores más altos indican un mejor rendimiento en la clasificación. Como se puede observar en la Tabla VIII, los casos donde se usa la técnica de RF en la fase de extracción de características, es donde se obtienen los mejores resultados. Además, los mejores resultados se obtuvieron cuando se seleccionaron 6 características por medio de RFE.

VI. Conclusiones y Trabajos Futuros

En general, este trabajo demuestra la importancia de conseguir buenos descriptores, con la finalidad de mejorar el comportamiento de los algoritmos de Aprendizaje Automático. En particular, la calidad de las métricas se ve influenciado por los descriptores usados (en este caso, se usó la métrica de *precisión* en tareas de clasificación).

Por otro lado, también el trabajo muestra como las diferentes etapas de la Ingeniería de Características están vinculadas. Según las técnicas que se usen en las fases de extracción de características y selección de descriptores, los resultados varían. Lo anterior también indica la necesidad de ver a la Ingeniería de Características desde un proceso de CA, donde las tareas de AdD se engranen entre ellas. En ese sentido, nuestro CA es pertinente para el descubrimiento de óptimos descriptores para procesos de reconocimiento de patrones.

Así, el CA mostró cómo al aplicar distintas técnicas de FC, FE y FS, y combinaciones de ellas, se puede obtener una gran variedad de resultados con calidades distintas, lo cual significa que debe seleccionarse correctamente la técnica que mejor ayude a determinar los descriptores relevantes de acuerdo al área de aplicación que se esté analizando en cada caso.

Este trabajo realiza la prueba de concepto sobre el CA, pero deriva en un importante número de trabajos futuros. Uno de los trabajos futuros es realizar muchas más pruebas con el CA (para diferentes contextos de aplicación), de tal manera de definir el perfil de técnicas adecuadas para el CA según el contexto de aplicación. Eso implica utilizar un mayor número de técnicas en cada una de las fases del CA (FC, FE y FS), no solamente RF, PCA, y RFE, ya que se demostró que el CA es muy sensible a las técnicas usadas en cada tarea de AdD. Otro trabajo futuro es usar más métricas de calidad, además de la precisión (*precision*), tales como la exactitud (*accuracy*) y la memorización (*recall*), en el caso de tareas de clasificación, para determinar si el comportamiento de calidad se mantiene con las diferentes métricas. También, otro trabajo es hacer el estudio del comportamiento del CA para problemas con datos con diferentes características: desbalance entre clases, datos etiquetados y no etiquetados, datos con ruidos, o muchas variables con comportamiento que se solapan entre ellas. Finalmente, en este trabajo solo se usó para el problema de clasificación a RF, pero hay que hacer pruebas con otras técnicas de clasificación, para determinar si se mantiene el comportamiento de las métricas de calidad en los descriptores seleccionados.

Referencias

[1] H. Liu and H. Motoda, "Feature Extraction, Construction and Selection: A Data Mining Perspective". Springer Science & Business Media Vol. 453, 1998.

[2] S. Khalid, T. Khalil and S. Nasreen, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning". In *Science and Information Conference (SAI),* pp. 372-378, 2014.

[3] B. Yoshua, O. Delalleau, N. L. Roux, J. Paiement, P. Vincent, M. Ouimet. "Spectral Dimensionality Reduction". In Feature Extraction: Foundations and Applications (I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh, Eds.), Springer, pp. 519-549, 2003.

[4] H. Motoda and H. Liu "Feature Selection, Extraction and Construction". *Communication of IICM (Institute of Information and Computing Machinery), Vol* 5, pp. 67-72, 2002.

[5] F. Pacheco, E. Exposito, M. Gineste, C. Budoin, J. Aguilar, "Towards the Deployment of Machine Learning Solutions in Traffic Network Classification: A Systematic Survey", accepted to publication, IEEE Communications Surveys and Tutorials, 2018.

[6] M. Chang, P. Buš, G. Schmitt, "Feature Extraction and K-means Clustering Approach to Explore Important Features of Urban Identity". In *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1139-1144. 2017.

[7] M. Sánchez, J. Aguilar, J. Cordero, and P. Valdiviezo, "Basic Features of a Reflective Middleware for Intelligent Learning Environment in the Cloud (IECL)", In *Proceeding of the Asia-Pacific Conference on Computer Aided System Engineering (APCASE)*. 2015

[8] G. Kumar, P. Bhatia, "A Detailed Review of Feature Extraction in Image Processing Systems" In *Fourth International Conference on Advanced Computing & Communication Technologies (ACCT),* pp. 5-12. 2014.

[9] J. Aguilar, O. Buendia, K. Moreno, D. Mosquera. "Autonomous Cycle of Data Analysis Tasks for Learning Processes", In Technologies and Innovation (R. Valencia-García, et al., Eds.), Communications Computer and Information Science Series, Vol. 658, Springer, pp. 187-202, 2016.

[10] J. Aguilar, J. Cordero L, Barba, M. Sanchez, P. Valdiviezo, L. Chamba, "Learning Analytics Tasks as Services in Smart Classroom", Universal Access in the Information Society Journal, Vol. 17, No. 4, pp. 693–709, 2018.

[11] J. Aguilar, J. Cordero, O. Buendia, "Specification of the Autonomic Cycles of Learning Analytic Tasks for a Smart Classroom", Journal of Educational Computing Research, vol 56, no. 6, pp. 866-891, 2018.

[12] M. Sánchez, J. Aguilar, J. Cordero, P. Valdiviezo-Díaz, L. Barba-Guamán, and L. Chamba-Eras, "Cloud Computing In Smart Educational Environments: Application in Learning Analytics as Service". In *New Advances in Information Systems and Technologies* (A. Rocha, M., Correia, H., Adeli, P. Reis, M. Mendonca, Eds.), Springer, pp 993-1002, 2016.

[13] C. Rangel, F. Pacheco, J. Aguilar, M. Cerrada, J. Altamiranda, "Methodology for Detecting the Feasibility of Using Data Mining in an Organization". Proceedings of the XXXIX Conferencia Latinoamericana en Informática (CLEI), Vol. 1, pp. 502-513, 2013.

[14] F. Pacheco, J. Aguilar, C. Rangel, M. Cerrada, J. Altamiranda, "Methodological Framework for Data Processing Base on the Data Science Paradigm", Proceedings of the XL Conferencia Latinoamericana en Informática (CLEI), 2014.

[15] K. Igor, and M. Kukar. "*Machine Learning and Data Mining"*. Horwood Publishing, 2007.

[16] C. Tran, C. Tseng, P. Chao, C. Shieh, L. Chan, T. Lee, "Face Recognition under Varying Lighting Conditions: A Combination of Weber-face and Local Directional Pattern for Feature Extraction and Support Vector Machines for Classification", Journal of Information

Hiding and Multimedia Signal Processing, Vol. 8 No 5, pp. 1009-1019, 2017.

[17] F. Wang, T. Xu, T. Tang, M. Zhou, H. Wang, "Bilevel Feature Extraction-Based Text Mining for Fault Diagnosis of Railway Systems", IEEE Transactions on Intelligent Transportation Systems, Vol. 18, No. 1, pp. 49-58, 2017

[18] C. Tran, M. Zhang, P. Andreae, B. Xue, "Genetic Programming based Feature Construction for Classification with Incomplete Data", Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1033-1040, 2017.

[19] A. Koay, A. Chen, I. Welch, W. K. G. Seah, "A New Multi Classifier System using Entropy-based Features in DDoS Attack Detection", International Conference on Information Networking (ICOIN), 2018.

[20] L. Talavera, An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering. In *International Symposium on Intelligent Data Analysis.* Springer, Berlin, Heidelberg, pp. 440-451, 2005.

[21] F. Peres, F. Fogliatto, Variable Selection Methods in Multivariate Statistical Process Control: A Systematic Literature Review. *Computers & Industrial Engineering*, *115*, 603-619, 2018.

[22] O. Abedinia, N. Amjady, H. Zareipour, "A New Feature Selection Technique for Load and Price Forecast of Electrical Power Systems", IEEE Transactions on Power Systems, Vol. 32, No. 1, 2017.

[23] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, H. Greenspan, "Chest Pathology Identification Using Deep Feature Selection with Non-Medical Training", Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, Vol. 6, No. 3, pp. 259-263, 2016.

[24] Children Handwriting DataBase https://goo.gl/R1d6ZK

[25] Project-Team IntuiDoc, "IntuiScript Project: Handwriting Quality Analysis", In Intuitive user interaction for document": Technical Report. pp. 19 – 23, 2016. https://goo.gl/KsC9Xo