ISSN 1870-4069

Clasificación de Diabetes Mellitus tipo II detectando factores de riesgo en un conjunto de datos

Juan Manuel Cancino-Gordillo, Mireya Tovar-Vidal

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
México

juan@comitan.com, mireya.tovar@correo.buap.mx

Resumen. Una de las enfermedades más importantes a nivel mundial en salud pública es la Diabetes Mellitus (DM), ya que esta es una de las enfermedades no transmisibles más severa, frecuente y con diversas complicaciones crónicas. En este documento proponemos un método para detectar los factores de riesgo más comunes en pacientes que padecen la enfermedad conocida como diabetes mellitus tipo II, a través del análisis de componentes principales. Posteriormente comprobamos los resultados utilizando estos factores como características, por medio del algoritmo J48 mejorando los resultados de la clasificación. De acuerdo a los resultados experimentales se obtiene un 86.9% de precisión, la cual es una mejora en comparación con trabajos relacionados.

Palabras clave: Aprendizaje automático, PCA, treej48, clasificación.

Classification of Type II Diabetes Mellitus Detecting Risk Factors in a Data Set

Abstract. One of the most important diseases worldwide in public health is Diabetes Mellitus (DM), since this is one of the most severe and frequent non-communicable diseases with various chronic complications. In this document we propose a method to detect the most common risk factors in patients suffering from the disease known as type II diabetes mellitus, through principal component analysis. Later we check the results using these factors as characteristics, by means of the J48 algorithm, improving the classification results. According to the experimental results, 86.9% accuracy is obtained, which is an improvement compared to related works.

Keywords: Machine learning, PCA, treej48, classification.

1. Introducción

En el campo médico el diagnóstico es la parte más crítica al momento de tratar a una persona, ya que el médico utiliza sus conocimientos para detectar ciertos patrones en el

comportamiento o estudios médicos de un paciente y llegar a una conclusión que se traduce en el tratamiento o medicación. Entre la amplia gama de enfermedades, la Diabetes Mellitus (DM) y sus variantes son considerados severos por las diversas complicaciones crónicas y requiere la atención de diferentes especialistas para el tratamiento de una persona.

Existen dos variantes de la DM, tipo I y tipo II. La más común es la Diabetes Mellitus tipo II (DMT2), la cual es una enfermedad en la que el organismo no genera suficiente insulina para procesar la glucosa en la sangre dejando mucho de este material circulando en el sistema sanguíneo. En México se atribuye el 11.8% de muertes desde el 2005 [1] y un 63% como causa de muerte principal de enfermedades crónicas a nivel mundial en el 2015 [2].

El propósito de este trabajo es la detección de factores de riesgo que pueden originar la DMT2, a partir del análisis de componentes principales en un conjunto de expedientes clínicos. Posteriormente se aplica un algoritmo de aprendizaje automático para corroborar que esos factores contribuyen en la detección de DMT2, al mejorar la precisión en los resultados experimentales.

El artículo está distribuido en cuatro secciones. En la sección 2 se presentan los trabajos relacionados con esta investigación. En la sección 3 se incluye una breve explicación a la metodología, algoritmos y métricas utilizadas. En la sección 4 se presentan los resultados experimentales y finalmente se incluyen las conclusiones del trabajo realizado.

2. Trabajos relacionados

A continuación, se describen algunos trabajos relacionados con el uso de algoritmos de aprendizaje automático.

En el trabajo presentado por AlJarullah et al. [3] se utilizan árboles de decisión con un conjunto de datos enfocado en mujeres para la detección de la diabetes. Este trabajo está presentado en dos etapas, la primera que consiste en mejorar los datos con un pre-procesamiento de datos aplicando métodos como *garbage in*, *garbage out* utilizados generalmente en proyectos de minería de datos, eliminando instancias del conjunto de datos.

La segunda etapa consiste en el uso del nuevo conjunto de datos para generar un árbol de clasificación usando el algoritmo TreeJ48, mostrando la matriz de confusión para calcular métricas como precisión, exactitud y F_1. Al tener menos información irrelevante se genera un modelo de predicción más exacto, lo cual logró subir la precisión a un 78.17%. Demostrando que el pre-procesamiento de datos mejora la clasificación de instancias del conjunto de datos.

Los autores del trabajo [4] proponen varios algoritmos utilizados en la rama de minería de datos como: SMO (*Sequential Minimal Optimization*), *random forest*, *tree* J48 y *Naive Bayes* para comparar el rendimiento de los algoritmos de clasificación y poder determinar que algoritmo posee una mayor exactitud al realizar el diagnóstico.

Los autores aplican métodos para la limpieza de los datos no requeridos para el estudio, así mismo los autores interpretan los datos faltantes del conjunto de datos. Para la evaluación de los algoritmos utilizan el método conocido como *Cross-validation* junto a las métricas precisión, exactitud y F_1, en donde dividen su conjunto de datos

con una relación 50:50. Los autores mencionan que la relación que utiliza para su conjunto de datos no es ideal, ya que para este tipo de evaluaciones es mejor seccionar en tres partes el conjunto de datos. En los resultados y conclusiones del trabajo nos hace la mención de la exactitud del algoritmo J48, llegando en el mejor de los casos al 73.82% de exactitud al clasificar.

En el 2015 los autores del trabajo [5] presentaron una comparación entre diferentes algoritmos de minería de datos utilizando un conjunto de datos llamado *Pima Indians Diabetes Dataset*. El cual consiste de 768 registros con ocho atributos (edad, insulina, presión, entre otros) y un campo para clasificar (positivo y negativo).

Los algoritmos utilizados en el trabajo son: árboles de decisión J48, *Naive Bayes* y RBF (*Radial Basis Function*), donde utilizan tres métricas de evaluación (precisión, exhaustividad y F_1) para medir el resultado de la clasificación. En el trabajo la precisión más alta fue del algoritmo J48 con un 77.1% en promedio, pero con una clasificación de instancias menor al algoritmo de *Naive Bayes*.

Mencionan los atributos del conjunto, pero no muestran algún tipo de tratamiento de datos. El conjunto de datos fue dividido para realizar el entrenamiento y evaluación, contando con una cantidad aproximada de 230 registros para la evaluación.

En el 2017 Yamilé et al. [2] realizaron un estudio transversal con diseño muestral aleatorio, para detectar la prevalencia de enfermedades crónicas no trasmisibles y sus factores de riesgo. El trabajo utilizó un total de 2085 registros de personas entre 14 municipios, de diferentes edades (32-56 años) utilizando variables como: sexo, edad, perímetro abdominal, glucosa, insulina, triglicéridos, colesterol, entre otros.

Con el uso de medias y desviación estándar para generalizar los atributos presentaron las tablas a varios expertos del campo para diagnosticar cada registro. Llegando a la conclusión que a mayor edad ($\geq$ 50 aproximadamente) se producen cambios hormonales y metabólicos que afectan a varios sistemas. Consecuentemente, desarrollando intolerancia a la glucosa, DMT2 y obesidad abdominal.

En el año 2018 Orlando A. Chan et al. [6] presentaron una investigación realizada sobre un conjunto de datos de 768 pacientes, donde todos los registros son basados en mujeres para la detección de diabetes gestacional, mencionan que dichos atributos en el conjunto de datos son de alta importancia para la detección de DMT2.

Algunas de las variables consideradas son: glucosa, insulina, presión sanguínea y edad. El objetivo final de los autores fue crear un sistema experto para detectar diabetes a partir de los atributos seleccionados del conjunto de datos utilizando algoritmos de clasificación proporcionadas por la herramienta WEKA y BigML.

Los autores utilizan como clasificador árboles de decisión y obtienen un 70% de precisión en la clasificación de pacientes que no presentan DMT2, un 63% para los que si presentan y un 73.83% de exactitud.-En este trabajo realizaremos un tratamiento de datos faltantes a un conjunto de datos con el fin de utilizarlos en mejorar la clasificación usando el algoritmo J48.

Posteriormente se eliminan atributos no relevantes del conjunto utilizando análisis de componentes principales (PCA) y se realiza una comparativa de los resultados, para demostrar que es factible el uso de tratamiento de datos y la reducción de términos con PCA sin perder exactitud en la clasificación.

3. Metodología

El proyecto se divide en cuatro etapas que se resumen a continuación.

3.1. Extracción de datos y análisis de datos

Para esta sección se presenta el conjunto de datos conocido como *Pima Indians Diabetes Dataset*, el cual es utilizado en artículos dentro de la sección de trabajos relacionados. Este conjunto es una recopilación de datos clínicos de mujeres con ascendencia hindú.

Este conjunto de datos será utilizado para la extracción de factores de riesgo, el procedimiento utilizado se compone de tres fases: La primera fase es la recopilación del conjunto de datos, detección de datos anormales y análisis de dichos datos. La segunda fase es explicada en la siguiente sección y en la fase final se hace uso de un análisis de componentes principales con el objetivo de determinar los factores de riesgo con mayor relevancia para llevar a cabo una clasificación.

3.2. Pre-Procesamiento

Con la finalidad de evitar una clasificación pobre que sacrifique la exactitud o precisión de nuestros resultados, el pre-procesamiento de datos es utilizado para eliminar atributos no relevantes, descartando la variedad de estos mismos, ya que los algoritmos empiezan a clasificar erróneamente al tener una gran cantidad de atributos no relevantes dentro de la información recopilada.

Una propuesta muy recurrida de aplicar un pre-procesamiento, es tomar en cuenta los registros completos de datos completos, pero esto provocaría la omisión de muchos datos y no proporcionaría una buena clasificación.

Una mejor propuesta para esta fase es la sustitución por media, la cual funciona calculando el promedio de cada atributo y remplazando a los datos faltantes, dependiendo del atributo que clasifica al registro. Para la aplicación de esta fase se hace uso del lenguaje de programación Python el cual nos ofrece varias herramientas para manipular grandes cantidades de datos de manera eficiente para concluir en un archivo separado por comas (*csv*) donde se guardarán los nuevos datos.

3.3. Algoritmos de clasificación

El objetivo de la implementación de los algoritmos de clasificación es descubrir que atributos son los encargados de realizar la clasificación de los datos para comparar estos atributos con los resultados de PCA y determinar una lista de atributos principales proveniente de los factores de riesgo.

Algunos algoritmos de clasificación utilizados más frecuentemente son: SVM (*Support Vector Machine*) [7], *Random Forest* [8], Arboles de decisión J48 [9] y *Naive Bayes* [10]. Una vez obtenido los factores de riesgos más relevantes del conjunto de datos se realizará nuevamente la clasificación del conjunto de datos para ver si existe una mejora en su clasificación con menor cantidad de atributos.

Tabla 1. Descripción del conjunto de datos.

Atributo	*Descripción*	*Min-Max*
Embarazos	Cantidad de embarazos	0-17
Glucosa	Concentración de glucosa en plasma a dos horas en una prueba de tolerancia a la glucosa oral	0-199
Presión sanguínea	Presión arterial diastólica (mm Hg)	0-122
Grosor de la piel	Espesor del pliegue cutáneo del tríceps (mm)	0-99
Insulina	Insulina sérica de 2 horas (mu U / ml)	0-846
IMC	Índice de Masa Corporal (Kg/m^2)	0-67.1
PedigreeFunction	Función de árbol genealógico de la diabetes	0.08-2.42
Edad	Edad en años	21-81
Resultado	Resultado del paciente a la enfermedad de diabetes	0-1

3.4. Evaluación

La evaluación se puede realizar de dos formas diferentes. La primera consiste en el uso de métricas de evaluación para medir los resultados obtenidos de la clasificación. La segunda consiste en utilizar un experto en el campo, el cual se encargará de revisar los estudios del paciente y dar el veredicto si la clasificación es correcta.

Para la evaluación es necesario tener presente la matriz de confusión generada por los algoritmos de clasificación, la cual nos compara la predicción de las clases con los resultados etiquetados, resultando en cuatro métricas [11]:

- Precisión: Es el número de casos relevantes recuperados entre el número de casos recuperados (ver Ecuación 1).
- Exhaustividad: El cual nos informa sobre la capacidad para identificar nuevos registros usando el modelo matemático generado (ver Ecuación 2).
- Exactitud: Mide el porcentaje de casos que el modelo ha acertado o clasificado correctamente (ver Ecuación 3).
- F_1: Es utilizado para combinar las medidas de precisión y exhaustividad en un solo valor (ver Ecuación 4):

$$Precisión = \frac{VP}{VP+FP}, \tag{1}$$

$$Exhaustividad = \frac{VP}{\mathrm{VP}+FN}, \tag{2}$$

$$Exactitud = \frac{VP+VN}{VP+VN+FP+FN}, \tag{3}$$

$$F_1 = 2 \cdot \frac{Precisión \cdot \mathrm{Exhaustividad}}{Precisión + \mathrm{Exhaustividad}}. \tag{4}$$

donde:

Verdaderos positivos (VP): Son resultados positivos clasificados correctamente.

Verdaderos negativos (VN): Son resultados negativos clasificados como positivos.

Falsos positivos (FP): Son resultados negativos clasificados como positivos.

Falsos negativos (FN): Son resultados negativos clasificados correctamente.

Tabla 1. Resultados experimentales.

Resultados	*Datos Completos*	*Tratamiento de datos*	*Tratamiento de datos + PCA*
Correctamente clasificado	567	666	**668**
Incorrectamente clasificado	201	102	**100**
Error absoluto (promedio)	31.58%	16.58%	**16.41%**
Precisión	63.24%	82.42%	**83.87%**
Exhaustividad	59.70%	**78.73%**	77.61%
Exactitud	73.83%	86.72%	**86.98%**
F_1	61.42%	80.53%	**80.62%**

El objetivo de la evaluación es de medir la eficiencia del modelo creado en registros nuevos. Esta eficiencia es medida en porcentajes que pueden variar dependiendo del conjunto de datos utilizado.

4. Resultados

En esta sección se presenta el conjunto de datos utilizados en los experimentos y los resultados obtenidos.

4.1. Conjunto de datos

El conjunto de datos utilizado en el estado del arte está basado en mediciones medicas provenientes del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales, donde el objetivo es predecir si un paciente tiene diabetes o no basándose en determinadas medidas de un diagnóstico. En particular, todos los pacientes en este conjunto de datos son mujeres de al menos 21 años y de ascendencia hindú, con un total de 500 registros no diabéticos y 268 diabéticos; también conocida como *Pima Indians Diabetes Dataset* [12].

En la Tabla 1 se describe el conjunto de datos que incluye ocho atributos y un resultado que indica si el paciente padece diabetes (clase 1) o no (clase 0).

4.2. Resultados experimentales

Antes de tratar los datos se realiza una ejecución del algoritmo de clasificación Treej48, para obtener las métricas de clasificación antes de cualquier modificación al conjunto de datos. Una vez realizado la clasificación se aplica el pre-procesamiento de los datos y se ejecuta nuevamente una ejecución del clasificador, dando dos conjuntos de datos extras para la comparación de los resultados al reducir los atributos.

Para aplicar la compresión de datos (PCA) se hace uso de la herramienta *Pandas* del lenguaje *Python*, el cual nos proporciona un *DataFrame* que representa una estructura ordenada de los datos y podemos crear arreglos de una manera más sencilla a partir de

I	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	169.5	33.6	0.627	50	1
1	1	85	66	29	102.5	26.6	0.351	31	0
2	8	183	64	32	169.5	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Fig. 1. Orden del archivo *CSV*.

```
Nombre del archivo (.csv): Diabetes
Columna de inicio: 1
Columna final: 9
Encabezados (Separados por comas): Pregnancies,Glucose,BP,ST,Insulin,BMI,DPF,Age
```

Fig. 2. Ejecución de algoritmo de reducción de términos.

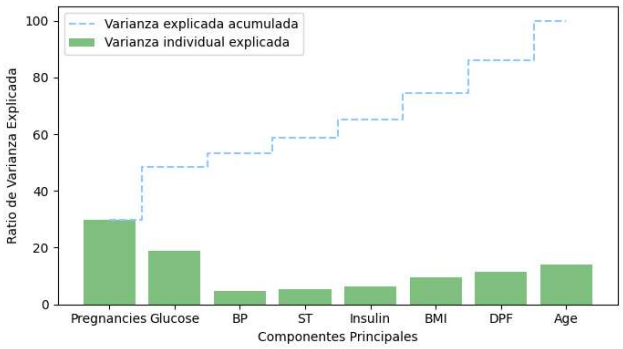

Fig. 2. Resultado final del algoritmo PCA.

un *csv* [13]. El archivo es ordenado respetando una única regla, la etiqueta clasificadora (*Outcome*) debe ser la última columna del archivo como se muestra en la Fig. 1.

Una vez que el archivo fichero cumpla con esa única regla será la entrada en un *script* de *Python*, donde se realizará la ejecución del algoritmo de análisis de componentes principales. También se solicitan datos como: Columna de inicio, columna final y los encabezados para la ejecución de los algoritmos. Posteriormente el *script* muestra una salida con la ponderación de cada atributo (ver Fig. 2).

En la Fig. 3 se puede ver que el algoritmo PCA utiliza la varianza explicada para representar la importancia de cada atributo dentro del conjunto utilizado y la varianza acumulada, la cual es la suma de cada columna, la cual al final siempre tiene que ser 100%. Los datos con mayor porcentaje son los atributos de embarazos y la glucosa, representado un 50% de la varianza dentro del conjunto de datos, aproximadamente. Los siguientes atributos a tomar en cuanta: edad, la función de árbol genealógico de la diabetes, índice de masa corporal e Insulina a con la finalidad de lograr un 90% de representación de los datos. Dejando a la presión arterial y grosor de la piel un 10% de representación.

Para llevar a cabo la clasificación del conjunto de datos se hace uso de la herramienta *WEKA*, ya que esta nos da una gran gama de algoritmos utilizados en el entorno de minería de datos, los cuales utilizan algoritmos de clasificación [14].

En las columnas de la Tabla 2, se presentan los resultados obtenidos por cada conjunto de datos usando validación cruzada de 10 *folds*, en donde la primera columna

Tabla 2. Comparativa de clasificación y métricas.

Tratamiento	Fuente	Precisión	Exhaus-tividad	Exactitud	F_1
Ninguno	[6]	63.24%	59.70%	73.83%	61.42%
Desconocido	[5]	78.60%	77.20%	N/R	77.10%
Pre-procesa-miento	Propia	82.42%	**78.73%**	86.72%	80.53%
Pre-procesa-miento + PCA	Propia	**83.87%**	77.61%	**86.98%**	**80.62%**

(datos completos) es el conjunto de datos sin retirar ningún atributo; la segunda columna (tratamiento de datos) representa el conjunto de datos tratando usando sustitución por media y la última columna representa el análisis de componentes principales sobre el conjunto de datos con tratamiento de datos y retirando atributos con el ~~los~~ análisis de PCA.

Como se puede ver en la Tabla 2, el aumento en los datos ***correctamente clasificados con tratamiento de datos*** (99 instancias) es un indicativo de una mejora en la clasificación del conjunto de datos utilizado que se aprecia en el incremento en la ***precisión***, ***exactitud*** y F_1.

Al comparar los resultados de las columnas tres y cuatro se observa que el incremento es mínimo (solo 2 instancias), pero demuestra que los atributos eliminados (grosor de piel y presión sanguínea), quienes dieron un bajo porcentaje reportado en el análisis de componentes principales realizado con anterioridad (>10%) no representan importancia al momento de utilizar el algoritmo de clasificación.

La Tabla 3 nos da una vista de los resultados en donde la primera columna se presentan la fuente, y las columnas restantes corresponde a cada métrica explicada con anterioridad. Se puede observar que la diferencia en las métricas reportadas por los autores del trabajo [6] y las presentadas en este trabajo existe un aumento en cada métrica con un tratamiento de datos basado en promedios, llegando a tener una precisión del 82.42% y una exactitud del 86.72%.

Mientras que la reducción de dos atributos en el conjunto de datos ayudó significativamente a mejorar la precisión llegando a un 83.87% y la exactitud al 86.98%. Mostrando que los atributos eliminados no tenían relevancia para la clasificación.

5. Conclusiones

El uso del algoritmo PCA, que es de suma importancia en el campo de la minería de datos, permitió reducir la cantidad de atributos utilizados en la clasificación. Esto provocó una ejecución más rápida dado que hay una cantidad menor de información en el conjunto de datos. Los resultados experimentales muestran una mejora en los resultados de las métricas de exactitud (86.92%), precisión (83.87%) y F1 (80.62%); en un total de 668 registros.

Al utilizar los algoritmos de clasificación con un pre-procesamiento de datos podemos ser capaces de ver que atributos son relevantes al momento de realizar una clasificación, pero al comprimir los datos después de tratar la información nos permite

eliminar atributos sin perder confianza en el modelo generado, logrando el 86.98% de exactitud.

El usar un tratamiento de datos ayuda a mejorar la clasificación del algoritmo por si solo en la mayoría de los casos, en este trabajo se utilizó un método simple para tratar datos no válidos, dejando la posibilidad de utilizar diferentes métodos como *k* vecinos más próximos o realizar discretización de datos para mejorar aún más la clasificación.

Las mejoras no simplemente se encuentran en el tratamiento de datos, sino que también aplican a los algoritmos de clasificación, en donde métodos más avanzados como son las redes neuronales o clasificadores ingenuos como *Naive Bayes* pueden aumentar la precisión de registros correctamente clasificados, ya que algunos se basan en reglas establecidas por el conjunto de datos sacando conclusiones de los resultados.

Referencias

1. Rodríguez-Rivera, N. S., Cuautle-Rodríguez, P., Castillo-Nájera, F., Molina-Guarneros, J. A.: Identification of genetic variants in pharmacogenetic genes associated with type 2 diabetes in a mexican-mestizo population. Biomed. Reports, vol. 7, no. 1, pp. 21–28 (2017) doi: 10.3892/br.2017.921
2. Miguel, P. E., Sarmiento, Y., Mariño, A. L., Llorente, Y., Rodríguez, T., Peña, M.: Prevalencia de enfermedades crónicas no transmisibles y factores de riesgo en adultos mayores de Holguín. Rev. Finlay, vol. 7, no. 3, pp. 155–167 (2017)
3. Al-Jarullah, A. A.: Decision tree discovery for the diagnosis of type II diabetes. In: Proceeding of International Conference on Innovations in Information Technology, pp. 303–307 (2011) doi: 10.1109/INNOVATIONS.2011.5893838
4. Hemant, P., Pushpavathi, T.: A novel approach to predict diabetes by cascading clustering and classification. In: Proceedings of Third International Conference on Computing Communication and Networking Technologies (2012) doi: 10.1109/ICCCNT.2012.6396069
5. Sa'di, S., Maleki, A., Hashemi, R., Panbechi, Z., Chalabi, K.: Comparison of data mining algorithms in the diagnosis of type II diabetes. International Journal on Computational Science & Applications (IJCSA), vol. 5, no. 5, pp. 1–12 (2015) doi: 10.5121/ijcsa.2015.5501
6. Chan, O., Peña, J., Vianne, J., Zapata, M.: Construcción de un modelo de predicción para apoyo al diagnóstico de diabetes (construction of a prediction model to support the diabetes diagnosis). vol. 40, no. 130, pp. 2105–2122 (2018)
7. Joaquín-Amat, R.: Máquinas de vector soporte (SVM) con python. Cienciadedatos.net (2020) https://www.cienciadedatos.net/documentos/py24-svm-python.html
8. Joaquín-Amat, R.: Random forest con python. Cienciadedatos.net (2020) https://www.cienciadedatos.net/documentos/py08_random_forest_python.html
9. Joaquín-Amat, R.: Árboles de decisión con python: Regresión y clasificación (2020) https://www.cienciadedatos.net/documentos/py07_arboles_decision_pyt hon.html
10. Pedregosa, F.: Scikit-learn: Machine learning in python. The Journal of machine Learning research, vol. 12, pp. 2825–2830 (2011)
11. Heras, J. M.: Precision, recall, F1, accuracy en clasificación. IArtificial.net (2020) https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/
12. Machine Learning: Pima indians diabetes database. Predict the onset of diabetes based on diagnostic measures (2016) https://www.kaggle.com/uciml/pimaindians-diabetes-database
13. Wade, R.: Reading CSV files. Advanced analytics in power BI with R and python: ingesting, transforming, visualizing, apress (2020) pp. 151–175

14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA data mining software: An update. ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10–18 (2009) https://www.kdd.org/exploration_files/ p2V11n1.pdf