**Machine Learning aplicado a la escala de satisfacción de la vida (SWLS): Estudiantes de la Universidad Francisco José de Caldas.**

Reinaldo Espinel Torres - 20182020015

Miguel Ángel Romo -20182020027

Cristian Martínez Blanco 20182020155

Universidad Distrital Francisco José de Caldas

**Resumen - En el siguiente artículo está dirigido a conocedores en el tema de machine learning, en el artículo se describen los pasos realizado en el estudio realizado por los autores el cual tiene que ver con la felicidad y/o bienestar y su aplicación en algoritmos de machine learning, se citarán diferentes artículos relacionados con SWLS, machine learning y bienestar, los cuales serán la base en diferentes pasos del estudio en diferentes etapas del estudio. Se describe el proceso de construcción del modelo desde la obtención de los datos hasta el entrenamiento del modelo y su validación.**

**Abstract - In the following article is directed for the ones who has some kind of knowledge about machine learning, in the article are described the steps done in the studie made by the authors that has to be with happiness and wellness, different studies related with SWLS, machine learning and wellness will be quoted, this articles will be the bases in different stages of the studie, the process of building of the model will be described from the obtainment of the data to the model training and testing of itself.**

**Índice de Términos -** SWLS, Machine Learning, Random Forest,

Correlación.

# introducción

La felicidad ha sido durante décadas el objeto de estudio de autoridades y académicos que tienen como propósito el mejorar el bienestar no económico de la sociedad, tomando como ejemplo [(Oxford Happiness Inventory (OHI), construido por Argyle et al.(1989); Hills & Argyle, (2002))](#bookmark=id.mq5t4j7k452o), adicionalmente es de notar que diversos estudios han encontrado suficiente evidencia empírica para demostrar que los individuos tienen conceptos subjetivos con respecto a la misma y que son diversos los factores que influyen en la misma.

El modelo de bienestar subjetivo descrito según Ed Diener [(Diener,2000)](#bookmark=id.xbhpu7tngfwa) se encuentra compuesto por diversos factores tales como:

* La satisfacción por la propia vida (juicio global)
* Satisfacción en diferentes ámbitos (ejemplo, estudio).
* Junto a un nivel bajo de emociones negativas.

De esta forma es posible definir la satisfacción en la vida como un juicio impartido por la persona hacia su propia vida. De esta misma manera es necesario aclarar que la evaluación no es objetiva dado que se encuentra condicionada por las circunstancias del entorno vital en conjunto con los patrones internos de la persona.([Diener,1984](#bookmark=id.e1uqlgfwyfts);[Diener,2000](#bookmark=id.xbhpu7tngfwa)).

Durante la última década se han realizado numerosos estudios haciendo uso de esta escala; véase España ([Vázquez, Duque & Hervás, 2013](#bookmark=id.yeu7d81trqyw)) en el cual contó con una muestra de 2964 la cual constaba con personas de diferentes grupos de edades, genero,educación y empleo; Brasil ([Gouveia, Milfont, Nunes & Peçanha de Miranda, 2009](#bookmark=id.bap2h5d10h6b)) contando con una muestra de 2180 personas de distintos grados de educación; en Arabia [(Taisir Abdallah, 2012)](#TaisirAbdallah) con 864 personas entre los 18 y 31 años; en Brasil [(Soares, A. K., da Silva Nascimento, B., da Silva, J. S., da Cruz Serejo Barbosa, N., & Fernandes Kamazaki, D.,2021)](#anaKSilvaSoaresBrunaDaSilvaNascimiento). En los anteriores estudios se concluyó que la escala SWLS es válida para los contextos estudiados. En Colombia se ha usado esta escala poblaciones universitarias como en ([Gomez,Villegas, Barrera y Cruz, 2017](#bookmark=id.oifbtoc7dg9b)), ([Salazar,2009](#bookmark=id.jqwqascmxq6x)) y ([Vinaccia Alpi, E., Parada, N., Quiceno, J.M., Riveros Munévar, F., & Vera Maldonado, L.A,2019](#bookmark=id.u1908jfwj84b)), en el último en mención es el primer estudio desarrollado en Colombia con la escala SWLS en el cual concluyen que la mencionada es una herramienta de medición confiable para universitarios en la ciudad de Bogotá.

Adicional a la encuesta SWLS, las pruebas SABER realizadas por el estado han mostrado tener un uso más allá que evaluar las capacidades individuales de quienes las presentan, también han permitido evaluar los niveles de eficiencia de las diferentes instituciones de educación superior y sus respectivos programas académicos [(Melo-Becerra, Ligia & Ramos, Enrique & Hernández Santamaría, Pedro,2014)](#a15)

# Objetivos

* Determinar el bienestar y/o felicidad de los estudiantes de la Universidad Distrital Francisco José de Caldas mediante el uso de algoritmos de aprendizaje automatizado, basados en diferentes estudios realizados sobre felicidad, bienestar y machine learning.
* Corroborar si la escala de satisfacción de la vida (SWLS) es aplicable al modelo de la Universidad Distrital específicamente.
* A través del análisis de correlación y alpha de Cronbach, detectar si el dataset de las pruebas saber pro puede ser útil para predecir la satisfacción de estudiantes de la universidad Distrital Francisco José de Caldas

# Estado de la técnica

* 1. *Aprendizaje Automático y Algoritmos de Clasificación*

El aprendizaje automático es una forma de inteligencia artificial en la que se accede a una big data y se interpretan para un estudio determinado, entrenando el sistema para predecir una nueva información de entrada a partir de unos algoritmos de aprendizaje.

La clasificación es una tarea que realiza el sistema para determinar la clase que pertenece alguna serie de datos en específico, dependiendo de sus características, comportamientos y patrones. Así, estos algoritmos se dividen en dos tipos principales de clasificaciones: clasificación binaria y clasificación multiclase. La clasificación binaria son aquellos algoritmos que tan solo se pueden asignar dos clases diferentes (0 o 1), por lo general se utilizan para obtener un resultado de un sí o no, positivo o negativo (Forero, Piñero, Molano, 2019). La clasificación multiclase difiere a los algoritmos donde se pueden asignar múltiples categorías a las observaciones.

* 1. *Algoritmo de Clasificación Multiclase*

En este artículo se utilizaran los algoritmos de aprendizaje de clasificación multiclase, ya que el objetivo del modelo y desarrollo del estudio es predecir la satisfacción de un estudiante a partir de los rangos establecidos en el SWLS para un estudiante cuyos valores de las variables permiten su clasificación; estos algoritmos son Boque de decisión aleatoria (Random Forest), Máquinas de Vectores de Soporte (SVM) y K Vecinos Más Próximos (KNN).

* 1. *Índice De Satisfacción Segun SWLS*

Entre los variados componentes de satisfacción subjetiva, la escala SWLS se centra rigurosamente en tasar el bienestar global de las personas sin utilizar elementos en específico. La escala consta de 5 preguntas tipo Likert, cada pregunta proporciona una puntuación entre 1 y 7 donde la puntuación máxima posible es 35 y mínima 5.[(Diener,1985)](#bookmark=id.1rqo2piuhvx0)

Se resalta los siguientes puntos de corte, una valoración de 30-35 expresa que el individuo está muy satisfecho con su vida, 25-29 corresponde a que el individuo está satisfecho, 20-24 indica levemente satisfecho, 15-19 indica levemente insatisfecho, 10-14 indica insatisfecho y 5-9 indica que el individuo está extremadamente insatisfecho con su vida actual.

Teniendo en cuenta que la escala SWLS genera 6 rangos, se ha establecido una categoría numérica por cada rango de 1 a 6, donde 1 representa el menor rango siendo 5-9 y 6 figura el rango más alto posible que es 30-35.

# Procedimiento en el estudio

* 1. *Recolección de datos*

Se realizó una encuesta entre el 21 de Enero y el 8 de Febrero de 2021 en la cual un grupo de personas responden la adaptación y traducción al español de la encuesta SWLS de William Pavot y Ed Diener, complementando dicha información con componentes socioeconómicos adaptados al contexto de la Universidad Distrital “Francisco José de Caldas”, en total participaron 933 estudiantes de las diferentes facultades de la Universidad Distrital, cada uno respondiendo de forma voluntaria a 17 variables que se presentan en la [tabla 1](#tabla1).

* 1. *Tratamiento de datos*

Luego de la obtención del dataset con los datos recogidos en la encuesta, se procedió a limpiar los errores que los estudiantes hayan podido cometer a la hora de llenar la encuesta, para posteriormente convertir de datos categóricos a datos numéricos debido a que los algoritmos de machine learning son netamente matemáticos por lo que es importante realizar la conversión, luego de realizar lo anterior se sacará un puntaje acerca de las preguntas relacionadas con la encuesta SWLS dado que las preguntas proporcionan una clasificación de los datos mediante el puntaje obtenido se puede obtener la etiqueta relacionada con el bienestar y/o la felicidad de la persona encuestada al momento de diligenciar la encuesta. La conversión numérica puede encontrarse en el repositorio creado para el artículo.

* 1. *Validación de la Encuesta.*

Para la preparación de datos se realizó un [análisis exploratorio](#bookmark=id.7dwn7tmnjjzp) ,mediante las librerías de python Numpy y Matplotlib en la que se utilizó estadística descriptiva relacionando las diferentes variables del modelo, posteriormente se procedió a desarrollar la matriz de correlaciones entre las variables incluyendo la correlación de cada variable con el puntaje obtenido en la encuesta SWLS ya que es importante para [(Silva, Da Silva, Silvestre, Da Cruz, Fernandez,2021)](#anaKSilvaSoaresBrunaDaSilvaNascimiento) relacionar las variables sociodemográficas con la encuesta para revisar cómo estas influyen en el puntaje obtenido y para tener una idea global del peso de las variables en el estudio y se utilizó para cada columna de la tabla el [coeficiente de correlación de Pearson](#correlaciondePearson), posterior para medir la confiabilidad del instrumento estadístico utilizado se procedió a hallar el [coeficiente alfa de Cronbach](#alfadecronbach) en su forma normal y estandarizada (esta última se hace desde la correlación de las variables) ya que [(Gallardo,Sanchez, Leiva)](#bookmark=id.9aw22d6jpe0q) utilizan el coeficiente para evaluar la consistencia interna de la encuesta en su respectivo estudio, este coeficiente arrojó una confiabilidad del 68% en sus dos formas por lo que se puede afirmar que posee una alta fiabilidad para la encuesta realizada por los autores del estudio.

* 1. *Modelos Escogidos.*

Para la selección del algoritmo de clasificación se tuvo en cuenta el tipo de machine learning, en nuestro caso se presenta machine learning supervisado, además los resultados se caracterizan por ser varias clasificaciones con diversas variables independientes, por ende los algoritmos de clasificación que se adaptan más a estas circunstancias son el SVM (Máquinas de Vectores de Soporte) y random forest para realizar el entrenamiento [(Kotsiantis, Zaharakis, Pintelas, 2016),](#a17) y la posterior validación del modelo, se escoge Random forest debido a que este método presenta diversas ventajas, al tratarse de un método no paramétrico, y no se encuentra basado en un modelo estocástico en particular, tanto como que no se asume una distribución específica de la variable objetivo, y no requiere ningún tipo de información acerca del tipo de relación (lineal o no lineal) entre la variable objetivo y sus atributos, además para los datos ruidosos como el dataset que se ha obtenido se genera una normalización para evitar conflictos en el estudio [(Reis, Baron, Shahaf, 2018)](#a18). Finalmente este modelo permite alcanzar resultados más precisos en términos de predicción además de permitir evaluaciones más robustas sobre la importancia de las variables respecto a los demás métodos basados en árboles [(Corduas & Piscitelli,2016)](#a16), y SVM presenta ventajas debido a que su variedad de kernels hace que este pueda realizar modelos lineales, no lineales y combinados [(Babić,2017)](#a11)

* 1. *Evaluación del modelo.*

Posterior a la preparación de datos, se procede a la [evaluación](#bookmark=id.4kzcmpxezkcb) del modelo, para esto la librería utilizada (Scikit Learn) la cual realiza particiones en los datos, de las cuales una es usada para entrenamiento del modelo y la otra para el testeo del mismo, el procedimiento del test arroja las métricas necesarias para obtener el grado de confiabilidad del modelo, fue del 71% en precisión para random forest, y del 71% también para SVM lo que indica que los modelos escogidos son efectivos al momento de predecir bienestar y/o felicidad en los estudiantes de la Universidad Distrital. Para asegurarnos de las clasificaciones calculamos la estabilidad del puntaje obtenido en el score así garantizamos el valor otorgado, en nuestro caso se obtuvo un porcentaje de 82% de estabilidad y a fin de afianzar lo recolectado generamos una matriz de confusión que nos describe el desempeño de los modelos de clasificación por clase [(ver grafica 1).](#grafica1) Desde este punto de vista, los valores de la diagonal principal representan el grado de clases correctamente predichas.

* 1. *Medida o indicador de éxito.*

El caso de estudio obedece a un caso de aprendizaje supervisado, por lo que los indicadores que serán más utilizados serán las matrices de confusión y el porcentaje de precisión, esto nos indicará el grado de confiabilidad del algoritmo implementado a la hora de clasificar la felicidad de los estudiantes de la Universidad Distrital Francisco José de Caldas.

* 1. *Analizando los datasets de saber-PRO*

Se escogió el Dataset de las pruebas saber-pro el cual contenía 11 variables de las 17 que escogieron los autores del presente estudio, para realizar a validación de la prueba saber pro se realizó la matriz de correlación para las variables del dataset, con el ánimo de ver cómo se comportan unas respecto a otras, posteriormente se procede a realizar el cálculo del coeficiente alfa de Cronbach el cual va a indicar la fiabilidad de la encuesta, con esto se garantiza que el Dataset de las pruebas saber-pro con un 75% de fiabilidad es lo suficientemente óptimo para realizar estudios de bienestar y felicidad en los estudiantes.

1. CONCLUSIONES

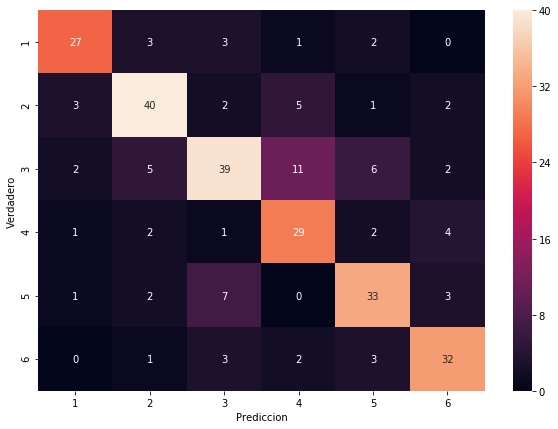
El alpha de cronbach aplicado al dataset obtenido a partir del formulario creado nos ayuda a determinar la confiabilidad generada en la encuesta. En este caso se ha obtenido una fiabilidad del dataset del 68% siendo interpretada como una encuesta “buena”, asimismo se consideran que las variables de sexo y programa académico no son necesarias para predecir si una persona está satisfecha o no en la universidad Distrital Francisco José de Caldas ya que a partir de una prueba Tstudent y Pvalue nos dio un resultado menor al 5% donde se acepta la hipótesis nula y se afirma que los parámetros no tienen relación entre sí, además, gracias a esta cantidad de datos recolectados y credibilidad de la encuesta se produjo un modelo con una precisión del 71% donde la estabilidad de puntaje es del 82% generando una gran seguridad en el sistema creado. Con base a esta información al realizar el análisis del dataset de las pruebas saber pro del año 2019 se decidio analizar 11 variables de las 17 relacionadas a nuestra encuesta, dándonos una confiabilidad del 75% que se interpreta como un dataset “muy bueno”, por ende, las 11 variables de las pruebas saber pro pueden ser consideradas para obtener la satisfacción de los estudiantes de la Universidad Distrital Francisco José de Caldas en la facultad de ingeniería, igualmente pueden ser utilizadas para futuros estudios relacionados al SWLS.

# Indicaciones útiles

## Figuras y tablas

|  |  |
| --- | --- |
| **Variable** | **Definición** |
| Programa Académico | Carrera que cursa el estudiante |
| Numero de Matricula | Semestre actual del estudiante |
| Sexo | Género del estudiante: Masculino (M) o Femenino (F) |
| Edad | Edad del estudiante al momento de diligenciar el formulario |
| Estado Civil | Estado civil del estudiante al momento del estudio |
| Valor de la Matrícula | Costo por semestre en el momento del estudio |
| Vive con la Familia | Indica si el encuestado vive con su familia en el momento del estudio |
| Cabeza de Familia | Da una noción del estado de responsabilidad del encuestado. |
| Personas a Cargo | Cantidad de personas a cargo del encuestado |
| Estrato | Clasificación socioeconómica el cual se evalúa entre 0 y 6 |
| Ingreso Familiar | Cantidad de ingresos mensuales |
| Trabaja Actualmente | Estado Laboral del encuestado |
| Horas Trabajadas | Horas trabajadas por el encuestado |
| Educación Padre | Nivel de educación alcanzado por el padre |
| Educación Madre | Nivel de educación alcanzado por la madre |
| Ocupación Padre | Estado laboral del padre |
| Ocupación Madre | Estado laboral de la madre. |

Tabla 1. Definición De Las Variables Empleadas



Grafica 1. Descripción Del Desempeño A Través De La Matriz De Confusión

* 1. Ecuaciones  
     + 1. Formula Random Forest

= La importancia de la característica i calculada a partir de todos los árboles en el modelo Random Forest.

= La importancia de la característica normalizada para i en el árbol j.

* + - 1. Coeficiente de Correlación de Pearson

Es un coeficiente (no tiene unidades) -1 .

Si existe una correlación directa fuerte

Si existe una correlación inversa fuerte

Si existe una correlación funcional

* + - 1. Alfa de Cronbach
         1. Forma Normal

= Varianza del ítem i

Varianza de los valores totales observados

* + - * 1. Forma Estandarizada

Es el número de ítems.

Es el promedio de las correlaciones lineales entre cada uno de los ítems (se tendrán pares de correlaciones

Reconocimiento

References

[1]Hills, P., & Argyle, M. (2002). The Oxford Happiness Questionnaire: A compact scale for the measurement of psychological well-being. Personality and individual differences, 33(7), 1073–1082. Recuperado a partir de <http://www.sciencedirect.com/science/article/pii/S0191886901002136>

[2] Diener, E. (2000). Subjective well - being the science of happiness and a proposal for a national index. American Psychologist, 55(1), 34-43. [http://dx.doi. org/10.1037/0003-066X.55.1.5](http://dx.doi.)

[3]Diener, E. (1984). Subjective well-being. Psychological Bulletin, 95, 542-575. Recuperado de [https://internal.psychology.illinois.edu/~ediener/Documents/ Diener\_1984.pdf](https://internal.psychology.illinois.edu/~ediener/Documents/)

[4]Vázquez, C., Duque, A., & Hervás, G. (2013). Satisfaction with Life Scale in a Representative Sample of Spanish Adults: Validation and Normative Data. Spanish Journal of Psychology, 16(82), 1-15. <https://www.ncbi.nlm.nih.gov/pubmed/24230945>

[5]Gouveia, V., Milfont, T., Nunes, P., & Peçanha de Miranda, J. (2009). Life satisfaction in Brazil: Testing the Psychometric Properties of the Satisfaction with Life Scale (SWLS) in Five Brazilian Samples. Social Indicators Research, 90, 267-277. <https://doi.org/10.1007/s11205-008-9257-0>

[6] Gómez, V., Villegas, C., Barrera, F., & Cruz, J. (2007). Factores predictivos de bienestar subjetivo en una muestra colombiana. Revista Latinoamericana de Psicología, 39(2), 311-325. Recuperado de <http://www.redalyc.org/pdf/805/80539208.pdf>

[7]Salazar, F. (2009). Felicidad en estudiantes de la facultad de salud de la Universidad Sur Colombiana. Revista Facultad de Salud, 1(2), 17-23. Recuperado de https://[www.journalusco.edu.co/index.php/rfs/article/view/42/59](http://www.journalusco.edu.co/index.php/rfs/article/view/42/59)

[8]Vinaccia Alpi, E., Parada, N., Quiceno, J.M., Riveros Munévar, F., & Vera Maldonado, L.A. (2019). Escala de satisfacción con la vida (SWLS): análisis de validez, confiabilidad y baremos para estudiantes universitarios de Bogotá. Psicogente 22(42), 1-20. <https://doi.org/10.17081/psico.22.42.3468>

[9]Ed Diener,Robert A. Emmos, Randy J. Larsen & Sharon Griffin,(1985). The Satisfaction With Life Scale

<http://labs.psychology.illinois.edu/~ediener/Documents/Diener-Emmons-Larsen-Griffin_1985.pdf>

[10] Leidy D. Forero, Yudy F. Piñeros, José I. Molano, (2019). Machine Learning For The Identification Of Students At Risk Of Academic Desertion.

<https://www.researchgate.net/publication/333407047_Machine_Learning_for_the_Identification_of_Students_at_Risk_of_Academic_Desertion>

[11] Ivana Đ. Babić (2017) Machine learning methods in predicting student academic motivation.

<https://doi.org/10.17535/crorr.2017.0028>

[12] Ismael Gallardo, Juan Sanchez, Marcelo Leiva (2011) Midiendo calidad en educación superior: análisis de confiabilidad y validez de una encuesta de certificación de calidad del pregrado, versión estudiantes.

http://dx.doi.org/10.4067/S0718-07052011000200006

[13] Ana k. Silva Soares, Bruna da Silva Nascimento, Jacqueline Silvestre da Silva, Nataly da Cruz Serejo Barbosa, Daniely Fernandes Kamazaki(2021)Psychometric properties of the Academic Major Satisfaction Scale (AMSS) in Brazilian College Students.

<https://doi.org/10.18800/psico.202101.010>

[14]Taisir Abdallah (1998) The Satisfaction with Life Scale (SWLS): Psychometric Properties in an Arabic-speaking Sample, International Journal of Adolescence and Youth, 7:2, 113-119.

<https://doi.org/10.1080/02673843.1998.9747816>

[15]Melo-Becerra, Ligia & Ramos, Enrique & Hernández Santamaría, Pedro. (2014). La educación superior en Colombia: Situación actual y análisis de eficiencia. Borradores De Economía, Banco de la Republica. 808. <https://www.researchgate.net/publication/260434677_La_educacion_superior_en_Colombia_Situacion_actual_y_analisis_de_eficiencia>

[16]Marcela Corduas, Alfonso Piscitelli.(2016). Modeling university student satisfaction: the case of the humanities and social studies degree programs.

<https://doi.org/10.1007/s11135-016-0428-9>

[17]Sotiris Kotsiantis, Id Zaharakis, Pe Pintelas, (2016). Machine Learning: a review of classification and combining techniques.

<https://www.researchgate.net/publication/226525180_Machine_learning_A_review_of_classification_and_combining_techniques>

[18] Itamar Reis, Dayla Baron, Sahar Shahaf, 2018, Probabilistic random forest: a machine learning algorithm for noisy data sets.

<https://iopscience.iop.org/article/10.3847/1538-3881/aaf101>

**Enlaces Externos**

[1] Análisis Exploratorio Sobre El Dataset De Estudio

<https://github.com/Cristian-Blanco/investigacion-operaciones/blob/master/ModeloPredictivo/Analisis%20Exploratorio.ipynb>

[2] Modelo Clasificador De Bosque Aleatorio

<https://github.com/Cristian-Blanco/investigacion-operaciones/blob/master/ModeloPredictivo/Modelo%20Predictivo.ipynb>