[[1]](#footnote-0)

Modelos de predicción de quiebra financiera

Juan Tovar\*, David Mautino\*, Rosmery Aguilar\*, Luis Enciso\*

***Resumen*—La evaluación de diferentes tipos de indicadores financieros para la predicción de quiebra en compañías es un tema de investigación que viene ganando importancia debido a la mayor disponibilidad de datos abiertos; más aún se ha demostrado en trabajos anteriores la alta eficacia de algunos métodos de aprendizaje máquina para la estimación de dicha condición. En el presente trabajo se busca identificar los modelos de aprendizaje máquina más óptimos para la realización de predicción de quiebra financiera basado en indicadores económicos que puedan ser en adelante adecuados a la realidad financiera peruana.**

***Index Terms*—quiebra financiera, aprendizaje máquina, indicadores financieros**

# INTRODUCCIÓN

L

A estimación efectiva de bancarrota en empresas es vital en muchos aspectos de la vida económica nacional, e.g. en muchas instituciones financieras se considera para decisiones de préstamo; además la bancarrota económica tienen un efecto no solo en la misma empresa, sino también en la economía global. Por los efectos e intereses mencionados existe una gran necesidad de generar modelos de predicción de bancarrota financiera; además la disponibilidad de abundante data financiera histórica permite que la generación de modelos predictivos pueda realizarse basado en modelos de aprendizaje máquina, lo que es reforzado por el buen desempeño mostrado por tales técnicas en tareas similares.

El objetivo de este estudio es por ello la generación de modelos de aprendizaje máquina adecuados para la predicción de quiebra financieras en empresas, bajo las siguientes condiciones: las empresas deben tener un mínimo de tres años de información financiera disponible, y debe existir un número suficiente de empresas en la misma área industrial para la comparación de casos de quiebra y no quiebra. Como objetivos específicos se consideran, la realización de una selección adecuada de indicadores económicos para el entrenamiento, la generación de modelos y el análisis comparativo de sus desempeños; y la la evaluación de modelos de ensamble.

El presente informe se compone de las siguientes secciones; en la sección 2 se describe el estado del arte en la generación de modelos de predicción de quiebra basado en indicadores financieros, y en la sección 3 se describe el diseño del experimento, mencionando las consideraciones de selección de indicadores financieros y se describe el data-set, el preprocesamiento e imputación del data-set, el procedimiento para el entrenamiento y los criterios de evaluación a utilizar.

# Estado del Arte

Las primeras aproximaciones a la predicción de quiebra financiera se dieron basadas en estudios empírico que permitían determinar el nivel de solvencia con el fin de predecir o evitar la posible quiebra a través de utilización de ratios, métodos estadísticos y el análisis financiero sobre los componentes de la solvencia, liquidez, rentabilidad, entre otros factores, numerosos estudios a partir de 1932 [1] han elaborado este aproximamiento. En los últimos años, la resurgencia de los métodos de aprendizaje máquina ha generado un enorme interés en la generación de modelos basados en aprendizaje supervisado. En [2] Liang presenta el data-set de indicadores financieros (FRs) basado en tres años de actividad empresarial en Taiwan, con el objetivo de utilizarlos en conjunto con indicadores de gobierno corporativo (CGI), no obstante, estos últimos son específicos a la realidad de ese país. Muchos trabajos en la materia han demostrado que la utilización exclusiva de FRs puede alcanzar una precisión alta, en [3] se presenta un resumen de los modelos aplicados a esta tarea de predicción usando indicadores financieros usando métodos estadísticos y aprendizaje máquina, e indica que estos últimos reportan mejor rendimiento; además se brinda una mirada comparativa a la aplicación de Redes Neuronales Artificiales (ANN), Regresión Logística (LR), Árboles de decisión (DT), y Matrices Vectoriales de Soporte (SVM), en la comparación SVM logra mejores resultados, además se puede realizar una optimización de hiper-parámetros mediante algoritmos evolutivos como Algoritmos Genéticos (GA) y Optimización por Enjambre de Partículas (PSO). En muchas ocasiones se realiza una mejora basada en modelos conjuntos (ensemble), usando técnicas tales como bagging, boosting, stacking, y sistemas de votaciones (e.g. mayoritaria, unánime). En [4] Liang propone un esquema de ensamble basado en voto unánime (UV) aplicado al data-sets como el de [2] , se reporta el mejor modelo combinado de SVM, ANN basado en bagging y CART basado en boosting, este enfoque es mejorado en [5] con un esquema de stacking ensemble para integrar un modelo basado en FRs con otro modelo basado en CGIs y score-Z, definido en [6], en este último trabajo [5] se usa Discrimante Análitico de Pasos (SDA) para la selección de FRs y CGIs y SVM para los modelos. Otros trabajos recientes tales como [7], donde Lin presenta un análisis para tres data-sets, muestra que ningún modelo es superior de manera absoluta en los tres casos, no obstante se reporta que los mejores modelos obtenidos son Naive Bayes (NB) + GA y SVM + GA, además ellos pueden complementarse con bagging; mientras que en [8] Tang presenta una comparativa de modelos de predicción basado en FRs y factores textuales y de manejo empresarial, estos factores son separados en sub-grupos dependiendo de su naturaleza y los modelos obtenidos son comparados, se puede observar una ligera mejora al incluir estos factores, sobretodo en modelos de emsemble por stacking y de Redes Neuronales Recurrentes (RNN).

# Diseño del Experimento

El diseño del experimento se realizó tomando en consideración variables financieras de la empresas de Taiwán obtenidas del Taiwan Economic Journal[[2]](#footnote-1) en los años 1999 a 2009 que nos indicará cual ha sido el comportamiento de las empresas y además nos indica cuando una empresa quebró. La quiebra de la empresa está definida en base a las regulaciones comerciales de la Bolsa de Valores de Taiwán.

* 1. ***Descripción del conjunto de Datos***

El dataset consta de 6819 casos y 96 variables, se valida que no se tienen datos faltantes en cada variable.

El dataset del experimento será normalizado en el rango de 0 a 1, en base a la siguiente fórmula:



donde F es un conjunto de una característica específica (es decir, variable), x es el valor de la característica y max (F) y min (F) son los valores máximo y mínimo del conjunto de características específicas, luego se observa que la distribución de tipo de característica en las variables financieras del dataset es el siguiente:

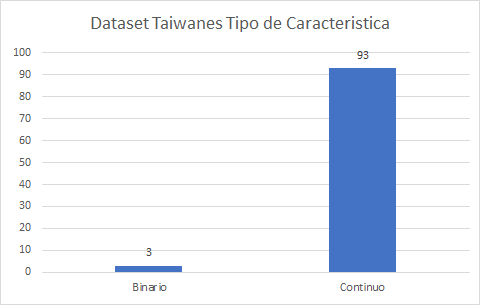
* 3 del tipo binario, que incluye al dato indicador de quiebra (flag)
* 93 del tipo continuo y
* Ninguno de tipo discreto.

Figura 1. Dataset Taiwanes Tipo de Característica en Variables

Debido a que el dato a evaluar (flag) es binario, los modelos serán de clasificación, considerando ello se utilizara una muestra aleatoria del 10% para el conjunto de prueba y el 90% restante para el conjunto de entrenamiento. Se debe resaltar al igual que el estudio realizado en [2] que existe una diferencia significativa entre el número de casos en quiebra y no en quiebra:

Tabla 1. Diferencia casos en quiebra y no en quiebra



Esto da como resultado un problema de desequilibrio de clases, que probablemente lleve a una degradación en el desempeño de la predicción final. Por lo tanto, utilizaremos el método de muestreo estratificado [6] para recopilar el mismo número de casos en quiebra y no en quiebra.

* 1. ***Metodología*** 
     1. **Selección de características**

El objetivo de la selección de características es determinar las características más representativas, reduciendo las características redundantes o irrelevantes para el apoyo en la optimización del modelo predictivo. Se utilizó el siguiente método para la selección de características:

* Random Forest Classifier
  + 1. **Modelos predictivos**

Al tratarse de la predicción de quiebra o no quiebra, se emplearon, en la etapa de experimentación, los siguientes algoritmos de clasificación:

* Decision Tree
* Logistic Regression
* Random Forest Classifier
* XGBoost Classifier
* Genetic Algorithm + Support Vector Machine
  + 1. **Métricas de evaluación**

Se utilizaron las siguientes métricas para la evaluación:

* Accuracy
* F1-Score
* Área bajo de la curva ROC
* Type I - error
  + 1. **Optimización de modelos**

Se utilizaron los siguientes métodos de optimización:

* Grid Search Cross Validation
* K-Fold Validation
  + 1. **Marco de la metodología**

Este estudio se desarrolló en el marco que se muestra en la

Figura 2, que implica seis fases claves [11], de las cuales 5 fueron realizadas en el presente trabajo: comprensión de los datos, preparación de los datos, modelamiento, evaluación y despliegue.

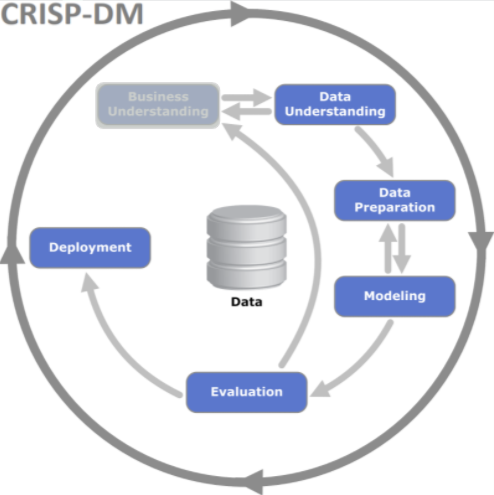


Figura 2. Metodología del desarrollo del modelo a utilizar

# Experimentación y Resultados

Las pruebas se realizaron con cada uno de los modelos base, además se realizó el ensamble de los mejores y por último, la búsqueda de un modelo en base a la reducción de las variables más importantes seleccionadas mediante ‘feature importance’ basado en Random Forest.

En cada caso se usó un división del dataset en sets de entrenamiento y prueba, con 80% y 20% respectivamente, usando el comando train\_test\_split de scikit-learn, en modo estratificado.

***4.1 Modelos Base***

Se ensayaron los siguientes modelos base:

**4.1.1 Logistic Regression (Figura 3)**

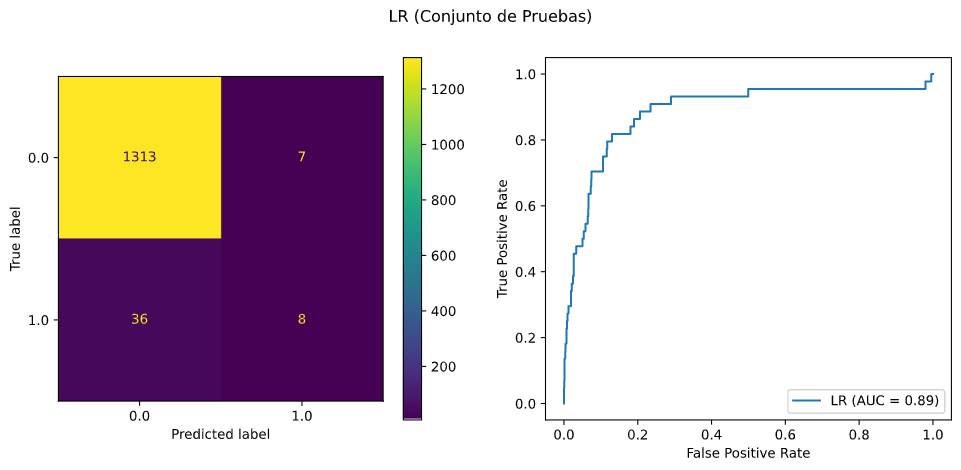


Figura 3. Resultados del modelo de Regresión Logística

**4.1.2 Random Forest (Figura 4)**

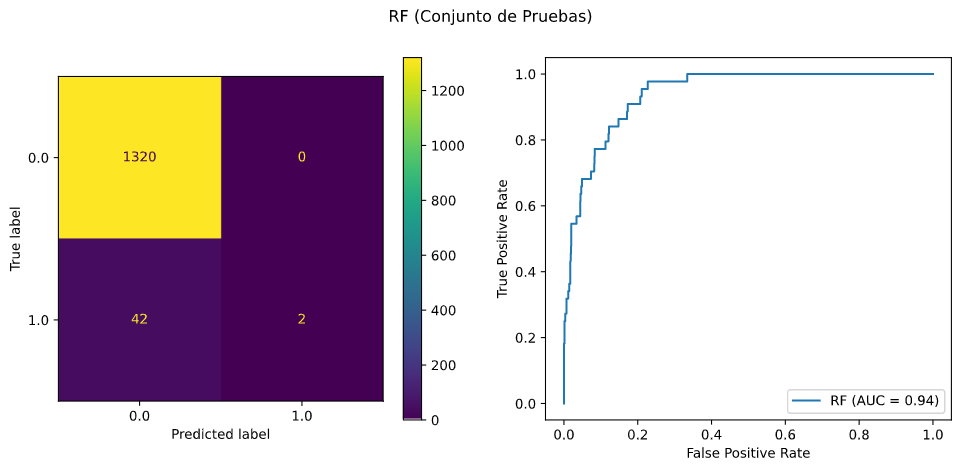


Figura 4. Resultados del modelo de Random Forest

**4.1.3 XG Boost (Figura 5)**

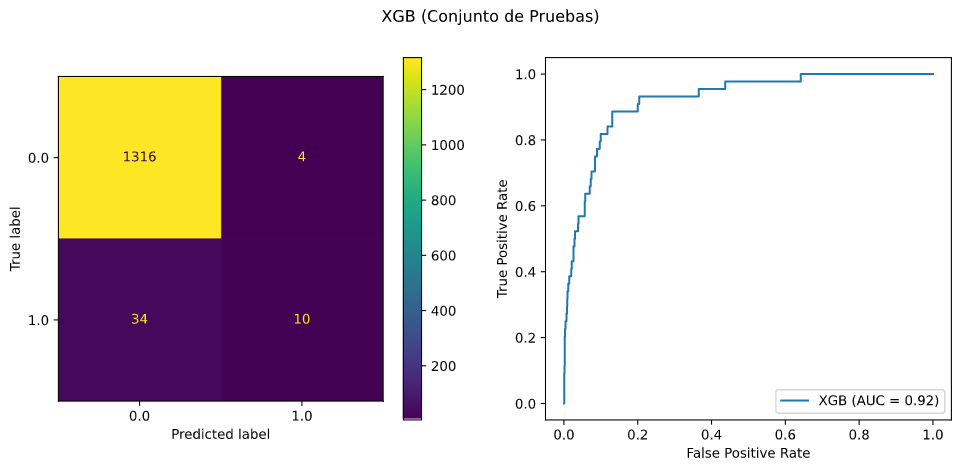


Figura 5. Resultados del modelo XG Boost

**4.1.4 Support Vector Machine (Figura 6)**

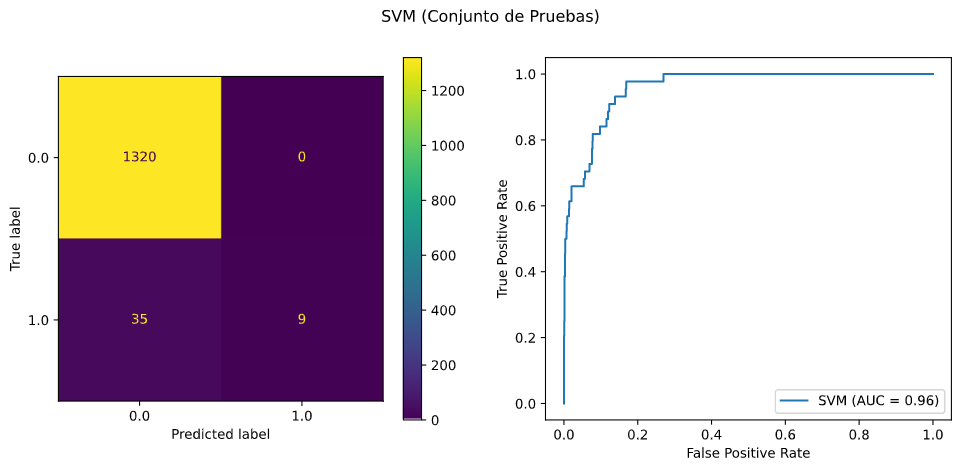


Figura 6. Resultados del modelo SVM

**4.1.5 Support Vector Machine + Genetic Algorithm(Figura 7)**

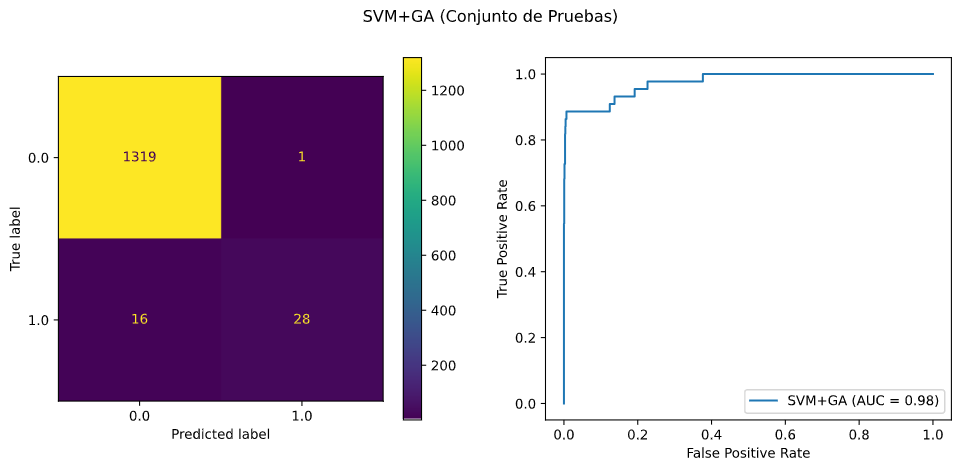


Figura 7. Resultados del modelo SVM + GA

***4.2 Modelo Ensamble* (Figura 8)**

Para el caso de ensamble se ensayó la obtención de un modelo mediante Stacking con optimización por medio de Logistic Regression.

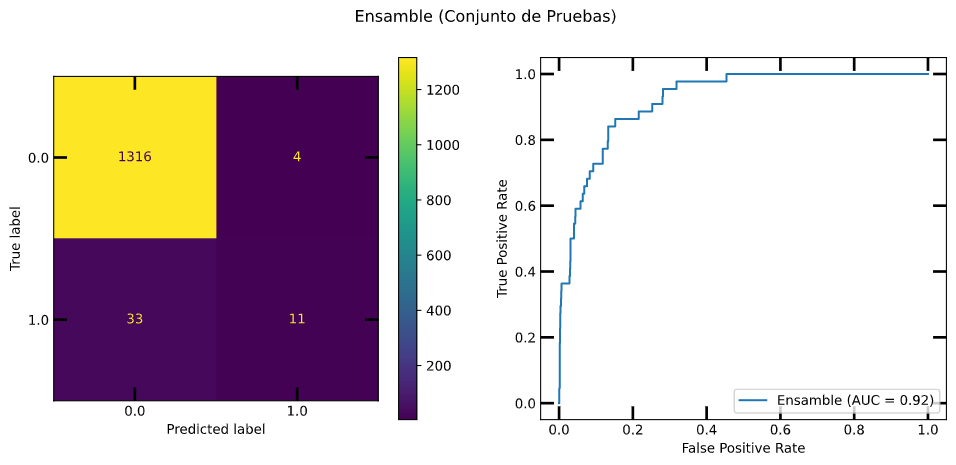


Figura 8. Resultados del modelo ensamble

***4.3 Modelos de Implementación***

Finalmente se entrenaron modelos con un subgrupo de las características, se encontró que con 70 características se pueden obtener resultados casi tan buenos como el mejor de los modelos previamente entrenados. Para este caso se ensayó con el algoritmo SVM+GA y auto-sklearn [10].

**4.1.1 SVM+GA (Figura 9)**

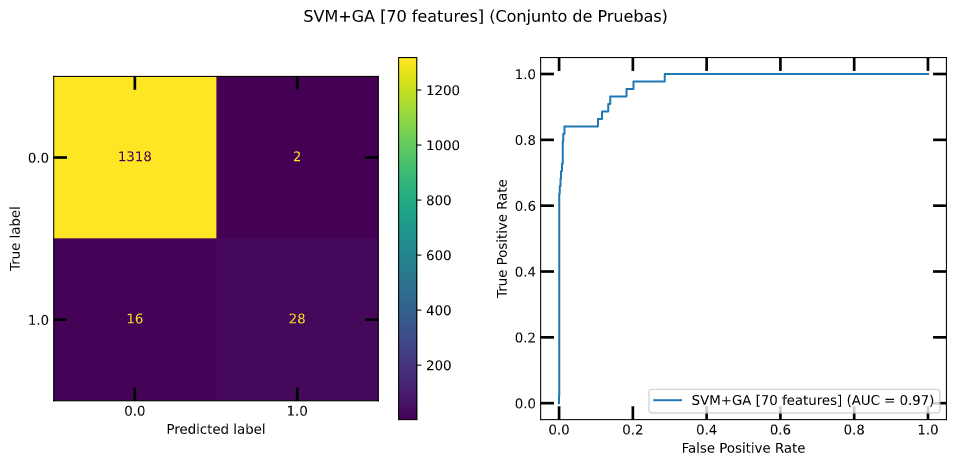
****

Figura 9. Resultados del modelo SVM+GA

**4.1.1 Auto-sklearn (Figura 10)**

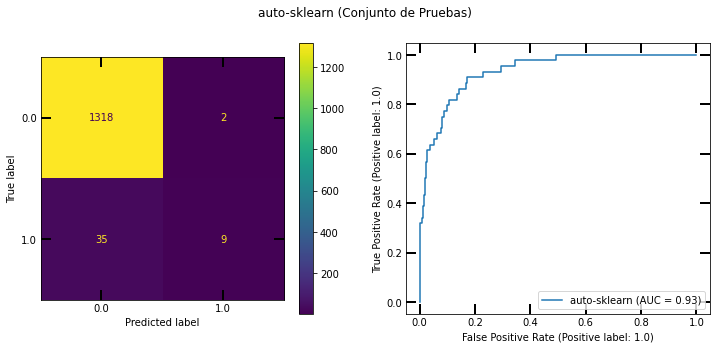


Figura 10. Resultados del modelo Auto-sklearn

# Discusión

# Se observa que los mejores resultados se obtuvieron con el modelo de SVM+GA, con una exactitud del 98.75% y con un F1-score del 76.71% (Tabla 2).

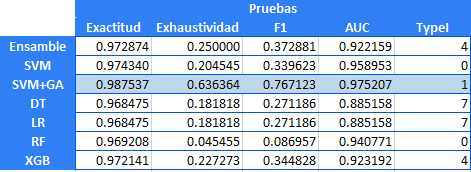


Tabla 2. Tabla comparativa de métricas usando las 94 características

* El experimento proporciona una diferencia del modelo auto-sklearn con un F1- score de 32.72% a comparación del modelo SVM+GA con un F1- score de 75.67% . Además se evidencia del despliegue del modelo auto-sklearn con un AUC de 93.35% a comparación del modelo SVM+GA con un AUC de 97.14% (Tabla 3).

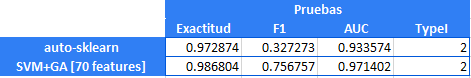


Tabla 3. Tabla comparativa de métricas usando las 70 características

* Estos resultados deben tenerse en cuenta al considerar 70 features en los modelos SVM-GA y auto-sklearn.

# Conclusiones y Trabajos Futuros

* Se obtuvieron mejores resultados con el modelo SVM-GA, con una exactitud de 98.7% y una exhaustividad de 63.6%.
* Se generó un modelo ensamble (stacking) con los 4 mejores modelos, sin embargo las métricas no mejoraron.
* Se realizó el feature importance con el algoritmo de Random Forest Classifier y se seleccionaron las 70 más relevantes en el modelamiento de SVM-GA y auto-sklearn.
* El modelo auto-sklearn no mostró mejoras en las métricas, por lo que se sugiere aumentar el tiempo de optimización de la selección del mejor modelo.
* Aplicar métodos más avanzados para solucionar el tema del balanceo del target u obtener mayor cantidad de casos en quiebra.

REFERENCIAS

1. FitzPatrick, Paul J., Ph.D. 1932. "A Comparison of the Ratios of Successful Industrial Enterprises With Those of Failed Companies". The Certified Public Accountant Beaver 1968. *Journal of Accounting Research*. (In three issues: October, 1932, p. 598-605; November, 1932, p. 656-662; December, 1932, p. 727-731.
2. D. Liang, C. C. Lu, C. F. Tsai y G. A. Shih, «Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study,» European Journal of Operational Research, vol. 252, p. 561–572, 2016.
3. S. S. Devi y Y. Radhika, «A survey on machine learning and statistical techniques in bankruptcy prediction,» International Journal of Machine Learning and Computing, vol. 8, p. 133–139, 2018.
4. Liang, D., Tsai, CF., Dai, AJ. et al. A novel classifier ensemble approach for financial distress prediction. *Knowl Inf Syst 54*, 437–462 (2018). https://doi.org/10.1007/s10115-017-1061-1
5. D. Liang, C. F. Tsai, H. Y. (. Lu y L. S. Chang, «Combining corporate governance indicators with stacking ensembles for financial distress prediction,» Journal of Business Research, vol. 120, p. 137–146, 2020.
6. Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23, 589–609.
7. W. C. Lin, Y. H. Lu y C. F. Tsai, «Feature selection in single and ensemble learning-based bankruptcy prediction models,» Expert Systems, vol. 36, p. 1–8, 2019.
8. Tang X, Li S, Tan M, Shi W. Incorporating textual and management
9. factors into financial distress prediction: A comparative study of machine learning methods. *Journal of Forecasting*. 2020;39:769–787. <https://doi.org/10.1002/for.2661>
10. Efficient and Robust Automated Machine Learning, Feurer et al., Advances in Neural Information Processing Systems 28 (NIPS 2015).
11. R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining", Proc. 4th Intl. Conference on Practical Applications of Knowledge Discovery and Data mining, pp. 29-39, 2000

ANEXO

**Dataset**

**ID Variable**

**y flag**

**x1 Cost\_of\_Interest-bearing\_Debt**

**x2 Cash\_Reinvestment\_Ratio**

**x3 Current\_Ratio**

**x4 Acid\_Test**

**x5 Interest\_Expenses/Total\_Revenue**

**x6 Total\_Liability/Equity\_Ratio**

**x7 Liability/Total\_Assets**

**x8 Interest-bearing\_Debt/Equity**

**x9 Contingent\_Liability/Equity**

**x10 Operating\_Income/Capital**

**x11 Pretax\_Income/Capital**

**x12 Working\_Capital\_to\_Total\_Assets**

**x13 Quick\_Assets/Total\_assets**

**x14 Current\_Assets/Total\_Assets**

**x15 Cash/Total\_Assets**

**x16 Quick\_Assets/Current\_Liability**

**x17 Cash/Current\_Liability**

**x18 Current\_Liability\_to\_Assets**

**x19 Operating\_Funds\_to\_Liability**

**x20 Inventory/Working\_Capital**

**x21 Inventory/Current\_Liability**

**x22 Current\_Liabilities/Liability**

**x23 Working\_Capital/Equity**

**x24 Current\_Liabilities/Equity**

**x25 Long-term\_Liability\_to\_Current\_Assets**

**x26 Current\_Liability\_to\_Current\_Assets**

**x27 One\_if\_Total\_Liability\_exceeds\_Total\_Assets;**

**x28 Equity\_to\_Liability**

**x29 Equity/Total\_Assets**

**x30 (Long-term\_Liability+Equity)/Fixed\_Assets**

**x31 Fixed\_Assets\_to\_Assets**

**x32 Current\_Liability\_to\_Liability**

**x33 Current\_Liability\_to\_Equity**

**x34 Equity\_to\_Long-term\_Liability**

**x35 Liability\_to\_Equity**

**x36 Degree\_of\_Financial\_Leverage**

**x37 Interest\_Coverage\_Ratio**

**x38 Operating\_Expenses/Net\_Sales**

**x39 (Research\_and\_Development\_Expenses)/Net\_Sales**

**x40 Effective\_Tax\_Rate**

**x41 Book\_Value\_Per\_Share(B)**

**x42 Book\_Value\_Per\_Share(A)**

**x43 Book\_Value\_Per\_Share(C)**

**x44 Cash\_Flow\_Per\_Share**

**x45 Sales\_Per\_Share**

**x46 Operating\_Income\_Per\_Share**

**x47 Sales\_Per\_Employee**

**x48 Operation\_Income\_Per\_Employee**

**x49 Fixed\_Assets\_Per\_Employee**

**x50 total\_assets\_to\_GNP\_price**

**x51 Return\_On\_Total\_Assets(C)**

**x52 Return\_On\_Total\_Assets(A)**

**x53 Return\_On\_Total\_Assets(B)**

**x54 Gross\_Profit\_/Net\_Sales**

**x55 Realized\_Gross\_Profit/Net\_Sales**

**x56 Operating\_Income\_/Net\_Sales**

**x57 Pre-Tax\_Income/Net\_Sales**

**x58 Net\_Income/Net\_Sales**

**x59 Net\_Non-operating\_Income\_Ratio**

**x60 Net\_Income-Exclude\_Disposal\_Gain\_or\_Loss/Net\_Sales**

**x61 EPS-Net\_Income**

**x62 Pretax\_Income\_Per\_Share**

**x63 Retained\_Earnings\_to\_Total\_Assets**

**x64 Total\_Income\_to\_Total\_Expenses**

**x65 Total\_Expenses\_to\_Assets**

**x66 Net\_Income\_to\_Total\_Assets**

**x67 Gross\_Profit\_to\_Sales**

**x68 Net\_Income\_to\_Stockholder's\_Equity**

**x69 One\_if\_Net\_Income\_is\_Negative\_for\_the\_Last\_Two\_Years;\_Zero\_Otherwise**

**x70 (Inventory\_+Accounts\_Receivables)\_/Equity**

**x71 Total\_Asset\_Turnover**

**x72 Accounts\_Receivable\_Turnover**

**x73 Days\_Receivable\_Outstanding**

**x74 Inventory\_Turnover**

**x75 Fixed\_Asset\_Turnover**

**x76 Equity\_Turnover**

**x77 Current\_Assets\_to\_Sales**

**x78 Quick\_Assets\_to\_Sales**

**x79 Working\_Capital\_to\_Sales**

**x80 Cash\_to\_Sales**

**x81 Cash\_Flow\_to\_Sales**

**x82 No-credit\_Interval**

**x83 Cash\_Flow\_from\_Operating/Current\_Liabilities**

**x84 Cash\_Flow\_to\_Total\_Assets**

**x85 Cash\_Flow\_to\_Liability**

**x86 CFO\_to\_Assets**

**x87 Cash\_Flow\_to\_Equity**

**x88 Realized\_Gross\_Profit\_Growth\_Rate**

**x89 Operating\_Income\_Growth**

**x90 Net\_Income\_Growth**

**x91 Continuing\_Operating\_Income\_after\_Tax\_Growth**

**x92 Net\_Income-Excluding\_Disposal\_Gain\_or\_Loss\_Growth**

**x93 Total\_Asset\_Growth**

**x94 Total\_Equity\_Growth**

**x95 Return\_on\_Total\_Asset\_Growth**

1. INF648

   \*Pontificia Universidad Católica del Perú, Escuela de Posgrado, Maestría de Informática (e-mail: author@pucp.edu.pe). [↑](#footnote-ref-0)
2. <https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction> [↑](#footnote-ref-1)