Journal of Information Processing Vol.26 170–176 (febrero de 2018)

[DOI: 10.2197 / ipsjjip.26.170]

Papel invitado

Predecir el rendimiento académico de los estudiantes usando Regresión Lineal Múltiple y Principal Análisis de componentes

Stephen JH Yang1,†1,a) Owen HT Lu1 Anna YQ Huang1 Jeff CH Huang2 Hiroaki Ogata3 Albert JQ Lin1

Recibido: 5 de septiembre de 2017, Aceptado: 24 de noviembre de 2017

Abstracto: Con el auge del análisis de big data, el análisis de aprendizaje se ha convertido en una tendencia importante para mejorar la calidad. de Educación. El análisis de aprendizaje es una metodología para ayudar a los estudiantes a tener éxito en el aula; el principio es predecir el rendimiento académico de los estudiantes en una etapa temprana y así brindarles una asistencia oportuna. Respectivamente, este estudio usó regresión lineal múltiple (MLR), un método popular para predecir el rendimiento académico de los estudiantes, para establecer un modelo de predicción. Además, combinamos MLR con análisis de componentes principales (PCA) para mejorar la Precisión predictiva del modelo. El MLR tradicional tiene ciertos inconvenientes; en concreto, el coeficiente de determinación (R2) y el error cuadrático medio (MSE) y la técnica de gráfico cuantil-cuantil (gráfico QQ) no puede evaluar el rendimiento predictivo y precisión de MLR. Por lo tanto, proponemos MSE predictivo (pMSE) y promedio predictivo medidas de corrección de porcentaje absoluto (pMAPC) para determinar el rendimiento predictivo y la precisión del modelo de regresión, respectivamente. Los resultados del análisis revelaron que el modelo propuesto para predecir el desempeño académico de los estudiantes rendimiento podría obtener valores óptimos de pMSE y pMAPC utilizando seis componentes obtenidos de PCA.

Palabras clave: análisis de aprendizaje, regresión lineal múltiple, análisis de componentes principales

1. Introducción

En los últimos años, los educadores han aplicado análisis de aprendizaje para mejorar la calidad de la enseanza y el aprendizaje. En Europa y el Estados Unidos, el Informe Horizon ha investigado anualmente la beneficios y métodos de las analíticas de aprendizaje desde 2011. The Horizon Report: Edición 2011 propuso que el objetivo de las analíticas de aprendizaje es permitir la adaptación humana de las respuestas de los estudiantes a través de adaptando el contenido de aprendizaje y ayudando a los estudiantes en riesgo a la derecha tiempo [1]. Con la importancia del análisis de big data, Horizon Informe: Edición 2016 propuso que la analítica de aprendizaje se convirtiera en la tendencia futura en la educación [2]. La analítica del aprendizaje es un proceso de medir y analizar los datos de aprendizaje recopilados a partir del aprendizaje entornos [3], [4], [5]. Predecir el rendimiento de aprendizaje de los estudiantes es uno de los principales temas de investigación en el análisis del aprendizaje.

Por ejemplo, Hu et al. recolectó datos de 300 estudiantes y estableció un modelo de predicción de riesgo estudiantil. Resultados experimentales reveló una precisión del 95 % en la predicción de las tasas de aprobación o reprobación de los estudiantes en función de 1 a 4 semanas de datos [6]. Meier et al. diseñó un proceso de selección de vecindarios para predecir las calificaciones de los estudiantes. Ellos afirmó que el algoritmo propuesto logró una precisión del 76% [7].

1 Departamento de Ciencias de la Computación e Ingeniería de la Información, Nacional Universidad Central, Taoyuan, Taiwán
2 Departamento de Ciencias de la Computación e Ingeniería de la Información, Hwa Universidad Tecnológica de Hsia, Xinbei, Escuela de
3 Posgrado en Informática de Taiwán, Universidad de Kyoto, Kyoto 606–8501, Japón
†1 Actualmente con la Universidad de Asia, Taiwán
a) stephen.yang.ac@gmail.com

Después de predecir el desempeño final de los estudiantes, la Universidad de Purdue diseñó e implementó “Course Signals” [8], una de las primeras solución de advertencia para aumentar el éxito de los estudiantes mediante la identificación temprana de riesgos. Además, las analíticas de aprendizaje se combinaron con diversas estrategias, como el aprendizaje colaborativo asistido por computadora. (CSCL). Por ejemplo, Van Leeuwen et al. desarrolló una herramienta de chat donde el instructor puede decidir cuándo intervenir en el grupo discusión basada en los resultados del análisis de emociones de texto [9]. Lu et al. desarrolló una herramienta de programación en pareja, en la que los instructores pueden proporcionar una intervención oportuna de acuerdo con el compromiso resultados de la medición [10].

Varios investigadores han aplicado la regresión lineal múltiple (MLR) para predecir el rendimiento de aprendizaje de los estudiantes [11], [12], [13], [14], [15], [16] o para identificar a los estudiantes en riesgo prediciendo el aprobar o reprobar el curso [17], [18], [19]. Con el rápido crecimiento de la tecnología de la información, el número de variables de datos recopilados de Los entornos de aprendizaje combinados también han aumentado considerablemente. Sin embargo, el número de variables utilizadas afecta considerablemente la bondad de ajuste del modelo de predicción obtenido mediante MLR. Para predecir el desempeño de aprendizaje de los estudiantes usando MLR, varios los investigadores han reducido el número de variables mediante la selección de algunas variables con mayor poder predictivo [17], [18], [19].

Por lo tanto, el objetivo del presente estudio fue investigar si MLR es adecuado para predecir el rendimiento académico de los estudiantes mediante el uso de un perfil de aprendizaje multivariable recopilado del curso de cálculo combinado propuesto.

Las medidas tradicionales utilizadas en MLR incluyen el cuadrado medio

(MSE), coeficiente de determinación (R2) y gráfico de cuantil cuantil (gráfico QQ). Estas medidas sólo pueden medir la
bondad de ajuste de un modelo de regresión, pero no puede evaluar el rendimiento de predicción de MLR. En el campo de la educación, es difícil para los profesores determinar si los resultados de la predicción de
MLR son creíbles a través de estas medidas. Por lo tanto, es necesario definir medidas de desempeño cuando se usa MLR para construir un
modelo de predicción. En consecuencia, este estudio se centró en el diseño medidas de rendimiento para medir el rendimiento de predicción de un modelo de regresión.

Reducir la intervención de los profesores, proporcionando mayor valor predictivo la precisión a los maestros es necesaria. Para predecir el rendimiento de aprendizaje de los estudiantes, varios investigadores se han centrado en cómo mejorar la precisión predictiva. Por lo tanto, este estudio investigó
métodos para mejorar la precisión predictiva de los modelos de regresión, y trató de responder las siguientes preguntas de investigación:

- **RQ1:** ¿Es MLR adecuado para predecir el desempeño académico de los estudiantes? desempeño mediante el uso de un perfil de aprendizaje multivariable recopilado del curso de cálculo combinado propuesto?
- **RQ2:** ¿Es posible mejorar la precisión predictiva de la proceso MLR propuesto?

2. Revisión de la literatura

MLR es un método de análisis predictivo basado en el análisis multivariado técnica estadística, y ha sido ampliamente utilizada en la educación.
El número de variables tiene una influencia considerable en el desempeño del proceso MLR. Las medidas MSE y R2 y la
La técnica de gráfico QQ se ha utilizado para evaluar la bondad de ajuste de modelos de regresión [20]. Sin embargo, estas medidas no pueden evaluar el desempeño predictivo del modelo de regresión. Por lo tanto, Proporcionar medidas de desempeño más sólidas a los maestros en el campo de la educación es necesario.

Para medir el error de predicción, el porcentaje absoluto medio error (MAPE) se calcula para medir el porcentaje de error de predicción de un modelo de predicción [21], [22]. El MAPE es uno de
los métodos más utilizados para evaluar el error de predicción.
Cuanto más bajo es el valor de MAPE, más bajo es el error de predicción de el modelo de predicción Por lo tanto, este estudio propone la corrección porcentual absoluta media predictiva (pMAPC) basada en la
concepto de cálculo de MAPE.

Sobre la base de la matriz de covarianza de los datos, el análisis de componentes principales (PCA) se usa típicamente para determinar
vectores propios no correlacionados a través de la descomposición en valores singulares y establecer los vectores propios como los componentes principales de la datos [23], [24]. Los componentes determinados se pueden utilizar como un nuevo conjunto de variables con mayor poder discriminativo en una regresión lineal. Algunos investigadores han propuesto que la precisión predictiva
de MLR se puede mejorar usando PCA [25], [26], [27]. Por lo tanto, este estudio combinó PCA con MLR para mejorar el pronóstico exactitud. Además, este estudio aplicó MAPC para medir la Precisión predictiva del modelo de regresión.

3. Curso de Cálculo Combinado

3.1 Participantes

Cincuenta y ocho estudiantes universitarios de primer año del norte de Taiwán partici pated en este estudio, que se llevó a cabo a partir de septiembre de 2015 a febrero de 2016. Este experimento se realizó en un curso llamado Clases Unidas de Cálculo. Los participantes comprendían 33 estudiantes masculinos y 25 femeninos. Los estudiantes aprendieron cálculo en el propuesta de curso semipresencial.

3.2 Actividades de aprendizaje en el curso de cálculo mixto

Para mejorar la calidad de la enseñanza y el aprendizaje, la propuesta
El curso de cálculo mixto combinó un entorno de aprendizaje en línea y un entorno de práctica en línea con la enseñanza del cálculo en el aula. Las actividades de aprendizaje de la propuesta semipresencial
El curso de cálculo comprendía la vista previa de materiales de aprendizaje en línea, instruyendo cálculo, practicando ejercicios en línea, practicando trabajo en casa y cuestionarios. Los datos de aprendizaje en el curso propuesto fueron recopilados mediante el registro de secuencias de clics de los estudiantes en el curso en línea plataforma y entorno de práctica de cálculo en línea. Además de enriqueciendo el conjunto de datos, recopilamos las notas de aprendizaje obtenidas en los cuestionarios y tareas. La información detallada sobre el
Los datos de aprendizaje recopilados se describen en la siguiente sección.

El curso de cálculo combinado propuesto tenía como objetivo desarrollar la capacidad matemática de los estudiantes a través del aprendizaje propuesto.
ocupaciones. En este estudio, se creó una versión china de Open edX para [*1] era permitir a los estudiantes obtener una vista previa de los materiales de aprendizaje en línea. antes de la clase, después de lo cual el profesor instruyó el tema en el clase. Para continuar con el comportamiento de aprendizaje después de la clase, los estudiantes realizaron ejercicios y tareas en línea como parte de sus actividades de aprendizaje. Un entorno de aprendizaje de cálculo en línea, a saber
Maple TA *2, fue creado para que los estudiantes participen en actividades de aprendizaje de cálculo en línea. Además, la maestra le asignó tarea.
a los estudiantes a continuar con el comportamiento de aprendizaje después de la clase. Para Para medir el rendimiento de aprendizaje de los estudiantes, el maestro administraba un cuestionario a los estudiantes cada 2 semanas. Recolectamos aprendizaje datos de los entornos de aprendizaje en línea aplicados para analizar la conducta de aprendizaje. Además, construimos un modelo para los estudiantes predicción del rendimiento académico mediante la combinación de PCA y MLR.

4. Metodología

4.1 Recopilación de datos

En el curso de cálculo semipresencial propuesto, recogimos aprendizajes datos de Open edX y Maple TA Recopilamos los datos de los estudiantes comportamiento de visualización de videos y calificaciones de ejercicios de Open edX y Arce TA, respectivamente. Para predecir el desempeño académico de los estudiantes, construimos un modelo para predecir las calificaciones finales de los estudiantes.

4.2 Conjuntos de datos de actividad de aprendizaje y variables de aprendizaje

Para los conjuntos de datos de la actividad de aprendizaje, los datos de aprendizaje de los estudiantes recopilados en este estudio comprendieron el comportamiento de visualización de videos en Open edX, ejercicios en Maple TA, finalización de tareas y el calificaciones del examen. Para guiar a los estudiantes a continuar aprendiendo cálculo, el instructor asignó ejercicios de tarea en formato papel al estudiantes cada 2 semanas. Para medir el rendimiento del aprendizaje. para cada tema de aprendizaje, el instructor administró un cuestionario en el clase cada 2 semanas. El curso de cálculo semipresencial propuesto duró 18 semanas; es decir, hubo nueve asignaciones de tarea y

[*1] https://open.edx.org/
[*2] https://www.maplesoft.com/products/mapleta/

```
{
   "username":"■■■■■■",
   "event_type":"pause_video",
   "ip":"123.110.40.112",
   "agent":"Mozilla/5.0",
   "host":"courses.openedu.tw",
   "session":"4c0801d5ce19ce4e9485bcf5ad647a7e",
   "event":"{
    \"id\":\"i4x-NKUHTx-TC101-video-\
        437045b9661a4060?e4fff0a8ef0e24d\",
    \"currentTime\":923.549472,
    \"code\":\"html5\"}",
   "event_source":"browser",
   "context":{
      "user_id":14514,
      "org_id":"NKUHTx",
      "course_id":"NKUHTx/TC101/201511",
      "path":"/event"
   },
   "time":"2016-01-01T13:22:12.181487+00:00",
   "page":"https://courses.openedu.tw/courses/NKUHTx/TC101\
    /201511/courseware/4e5d487d59ac460890f71edbd37d7f1c\
    /b25a4f18519643a8b651a0c55af04ffa/"
}
```


Fig. 1 Ejemplo de registros de seguimiento para un video en pausa en formato JSON.

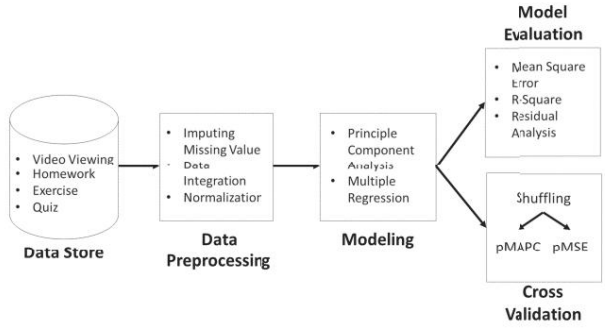


Fig. 2 Procedimientos involucrados en el modelo propuesto de predicción del rendimiento académico de los estudiantes.

nueve cuestionarios. Para recopilar datos de aprendizaje de la tarea y cuestionarios, registramos las calificaciones promedio obtenidas por los estudiantes en las tareas y cuestionarios. Para la recogida de los datos de la práctica de ejercicios se registró la nota media obtenida en los ejercicios online en Maple TA

Para recopilar datos de aprendizaje de Open edX, usamos Open edX Pipeline para recuperar las numerosas acciones de aprendizaje de los estudiantes de los registros de seguimiento en Open edX. Open edX Pipeline es un proyecto de código abierto que está completamente integrado con herramientas de análisis como Apache HDFS, Jenkins y MySQL. **La Figura 1** muestra un ejemplo de registros de seguimiento en formato JSON. Este estudio aplicó 27 variables extraídas de Open edX y Maple TA, tarea y prueba

4.3 Procedimientos Involucrados en el Rendimiento Académico de los Estudiantes

Modelo de predicción

Los procedimientos involucrados en el desarrollo del modelo propuesto de predicción del rendimiento académico de los estudiantes implicaron una fase de preprocesamiento de datos, una fase de modelado y una fase de evaluación. **(Figura** 2). La fase de preprocesamiento de datos involucró valor faltante imputación, integración de datos y normalización de datos. La fase de modelado implicó la ejecución de PCA y MLR. Finalmente, el la fase de evaluación comprendió la evaluación del modelo y la validación cruzada.

4.3.1 Fase de preprocesamiento de datos

La fase de preprocesamiento de datos implicó la extracción y trans formar datos no estructurados en datos estructurados para simplificar el análisis. El proceso de integración de datos se centró en integrar los datos de aprendizaje derivados de Open edX, Maple TA, tareas y cuestionarios para generar las 27 variables de aprendizaje propuestas. Los datos Se aplicó un proceso de normalización para redefinir el rango de datos. valores en un rango más pequeño y específico, porque el rango de varios valores de datos puede ser excesivamente amplio. Normalizamos la rango de las 27 variables propuestas de 1 a 10.

4.3.2 Fase de modelado

La fase de modelado implicó la construcción del estudiante académico modelo de predicción de rendimiento. En consecuencia, combinamos MLR y PCA. En primer lugar, se aplicó PCA para reducir el número de variables independientes mediante la extracción de un nuevo conjunto de variables del conjunto de variables original. Después de realizar PCA, MLR podría ejecutarse para construir el modelo de predicción del rendimiento académico de los estudiantes mediante utilizando las puntuaciones factoriales de los componentes extraídos a través de PCA.

4.3.3 Fase de Evaluación

La fase de evaluación consistió en medir el desempeño de el modelo propuesto de predicción del rendimiento académico de los estudiantes. Esta fase involucró la evaluación del modelo y la validación cruzada. El Los procesos de evaluación y validación cruzada del modelo se describen como sigue:

- Evaluación del modelo: al usar MLR para construir el modelo de predicción del rendimiento académico de los estudiantes, podríamos usar el MSE y medidas R2 y la técnica QQ plot para evaluar la bondad de ajuste del modelo de regresión. Un MSE más pequeño indica una mayor bondad de ajuste del modelo. Un valor de R2 más cercano a 1,0 indica una mayor bondad de ajuste del modelo.
- Evaluación cruzada: la validación cruzada es una validación de modelo tecnología que combina los valores medios medidos para derivar el valor estimado del rendimiento de la predicción y precisión del modelo de predicción. En validación cruzada, 10- se realiza una validación cruzada de pliegues con barajado para medir el rendimiento de la predicción y la precisión de la predicción modelo. En la validación cruzada de 10 veces con barajado, los datos originales primero se barajan, después de lo cual el conjunto de datos original se dividido en 10 subconjuntos de igual tamaño. Entre los 10 subconjuntos, 1 se selecciona como el conjunto de prueba y los 9 restantes son seleccionados como conjuntos de entrenamiento. El modelo de regresión de predicción se puede construir usando el conjunto de entrenamiento. El rendimiento de la predicción y la precisión del modelo de predicción se pueden calcular utilizando el conjunto de prueba. Cada uno de los 10 subconjuntos debe ser establecer exactamente una vez como el conjunto de prueba. La media de los 10 resultados. para el modelo de predicción se puede considerar como el estimado valor del rendimiento y la precisión de la predicción.
  Las medidas tradicionales de MSE y R2 y la técnica de gráfico QQ no pueden medir el rendimiento de la predicción y la precisión de los modelos de regresión. Por lo tanto, proponemos predictivo MSE (pMSE) y corrección porcentual absoluta media predictiva (pMAPC) para medir el rendimiento y la precisión de la predicción del modelo, respectivamente. Aplicamos 10 veces validación cruzada con barajado para calcular el pMSE y valores pMAPC. Modificamos el MSE y así obtuvimos pMSE para calcular el rendimiento de la predicción utilizando el prueba de datos en validación cruzada. Se usó MAPE para medir el porcentaje de error de predicción del modelo de predicción. Por

modificando el MAPE, derivamos la medida pMAPC a determinar la precisión del modelo de predicción. Las definiciones de pMSE y pMAPC se muestran en las ecuaciones. (1) y (2), respectivamente.

$$pMS\ E = \frac{1}{prueba} \sum_{prueba}^{yo=1} (pi - ai)^2 , pi \ddot{y} pprueba, ai \ddot{y} A \quad (1)$$

$$pMAPC = 1 \ddot{y} \frac{1}{prueba} \sum_{prueba}^{yo=1} \frac{pi - ai}{al} , pi \ddot{y} pprueba, ai \ddot{y} A \quad (2)$$

El conjunto $A = \{a1, a2, ..., an\}$ comprende la formación académica real calificaciones de los estudiantes. El símbolo *ntest* indica el número de elementos de datos en el conjunto de prueba. El conjunto *ptest* = p1, p2, ..., *pntest* comprende las calificaciones académicas pronosticadas en los datos de las pruebas. Podemos calcular los valores pMSE y pMAPC usando ecuaciones (1) y (2). Un valor de pMSE más bajo indica un rendimiento predictivo más alto. Además, un valor de pMAPC más alto indica una mayor precisión del modelo.

5. Resultados y Discusión

Para obtener el mejor poder explicativo del modelo de regresión, extrajimos los componentes principales de los datos originales después de la paso de preprocesamiento de datos en el modelo propuesto de predicción del rendimiento académico de los estudiantes. En el gráfico de pantalla que se muestra en la **Fig.** 3, cada barra del gráfico de barras y cada punto del gráfico de líneas representan la poder explicativo de cada componente y el poder explicativo acumulado, respectivamente. El poder explicativo de la primera componente para el modelo de regresión fue superior al 81%. Por Por el contrario, los niveles de poder explicativo acumulado de seis componentes fueron superiores al 96%, y el rendimiento predictivo del modelo de regresión para cada componente se discutirá en Sección 5.2.

5.1 Modelo MLR Evaluación de la bondad de ajuste

En general, MLR se usa para predecir el valor de las variables dependientes según la información histórica. Para seleccionar variables independientes, la relación causal entre variables independientes y Se deben considerar las variables dependientes. Para evaluar MLR, Las medidas tradicionales como R2 y MSE se utilizan para examinar la bondad de ajuste de los modelos de regresión. En este estudio, primero se examinó la bondad de ajuste del modelo de regresión usando el medidas tradicionales. Sin embargo, estas medidas tradicionales no pueden evaluar el rendimiento predictivo de los modelos de regresión. En consecuencia, los maestros no pueden obtener precisión predictiva usando medidas tradicionales en el entorno de enseñanza real. Por lo tanto, proponemos medidas adicionales para determinar el rendimiento predictivo de los modelos de regresión, lo que permite a los profesores evaluar la precisión predictiva.

La medida MSE se usa para evaluar qué tan cerca está una predicción la línea de regresión es un conjunto de valores reales de la variable dependiente. Esta medida se usa para calcular la varianza del error usando la suma residual de cuadrados dividida por el número de predichos datos. En un modelo de regresión, el residuo se define como el valor predictivo de los datos menos el valor real de los datos. Una baja El valor de MSE indica una mayor bondad de ajuste del modelo. En el modelo propuesto de predicción del rendimiento académico de los estudiantes, el aca

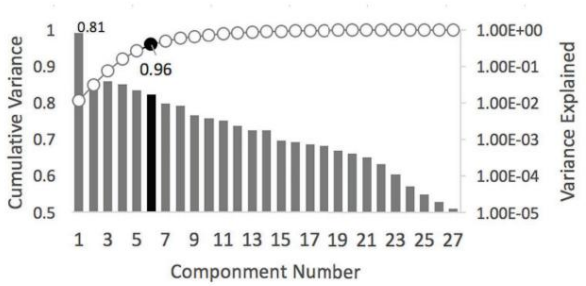


Fig. 3 El poder explicativo y los valores acumulados del poder explicativo para cada componente.

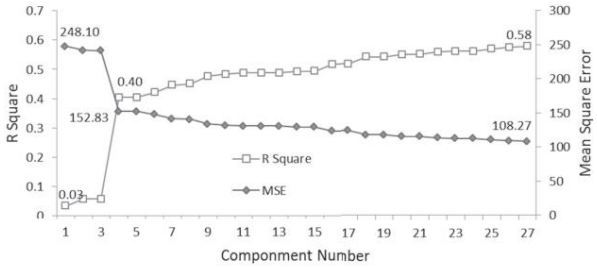
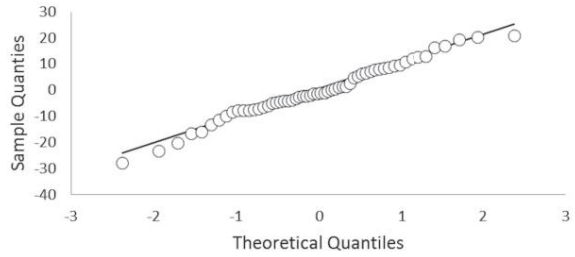


Fig. 4 Valor de MSE para cada componente en el modelo de regresión después de PCA.

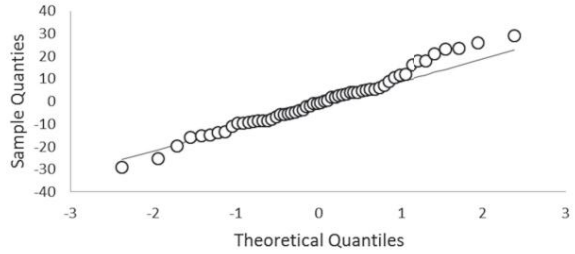
la puntuación démica en cálculo sirve como la variable dependiente predicha. Por lo tanto, en este estudio, primero se realizó PCA, seguida por MLR usando los componentes principales extraídos. los rangos de las variables dependientes pronosticadas y MSE fueron 0-100 y 0–10,000, respectivamente. Los valores de MSE obtenidos en el modelo de regresión después de PCA se presentan en la **Fig.** 4, indicando el MSE rango de 108,27 a 248,1. Esto implica que el rango de la el error predictivo de cada alumno fue de 10,4 a 15,8. La figura 4 muestra que cuando el número de componentes principales aplicados es 4, el valor de MSE se reduce drásticamente a 152,83, después de eso, el el valor de MSE continúa disminuyendo gradualmente.

R2 se utiliza para medir el poder explicativo de una regresión modelo utilizando varianzas de porcentaje entre los independientes y variables dependientes. Además, R2 es una de las medidas de bondad de ajuste para un modelo de regresión. Un valor R2 más alto indica un mayor poder explicativo para el modelo de regresión. Los valores R2 para MLR realizados en este estudio usando los componentes extraídos a través de PCA se presentan en la Fig. 4. De acuerdo con el primer componente principal en la Fig. 3, el valor de la explicativa la potencia para el primer componente del modelo de regresión propuesto fue 0.81, y esto se puede atribuir a alguna información faltante de los datos originales debido a PCA. Además, para el primer componente, el valor de R2 fue solo de 0,03 (Fig. 4). El poder explicativo del modelo de regresión aumentó cuando más componentes fueron aplicado. Posteriormente, el valor de R2 aumentó gradualmente, de 0,40 para 4 componentes a 0,58 para 27 componentes. De acuerdo a la tendencia creciente del valor R2 después de 4 componentes, el número de los componentes aplicados debe ser más de 4.

Además de examinar la bondad de ajuste de una regresión modelo utilizando R2 y MSE, la distribución de residuos para el Se debe examinar el modelo de regresión para determinar si se respalda la hipótesis de la distribución normal. En el análisis residual, el gráfico QQ es la técnica más utilizada para verificar esta hipótesis. Los resultados del análisis residual presentado usando el QQ gráfico se muestran en **la Fig.** 5. Como se indica en las Figs. 5 (a) y (b), el



(a) Result of residual analysis for the MLR without PCA



(b) Result of residual analysis for the MLR with the six components extracted from PCA

Fig. 5 Resultados del análisis de residuos utilizando el gráfico QQ.

las distribuciones de residuos para el modelo de regresión que involucra MLR sin y con PCA son similares a una línea recta. El valor de p de la prueba para el modelo de regresión que involucra MLR sin PCA fue 0,35 y para el modelo de regresión que involucra MLR con PCA fue 0,23. Por lo tanto, los dos resultados anteriores apoyan la hipótesis de la distribución normal.

Para abordar **RQ1,** de acuerdo con los resultados descritos de la prueba de bondad de ajuste y el análisis residual del modelo de regresión, los resultados de estas medidas obtenidas utilizando MLR con y sin PCA son satisfactorios. Sin embargo, el poder explicativo del modelo de regresión determinado usando MLR con PCA fue menor que el determinado usando MLR sin PCA. Mediante el uso de MLR con PCA, los valores de medidas como MSE y R2 pueden aceptarse en el campo de la educación. Por ejemplo, el valor de MSE determinado usando MLR sin PCA fue de 10,23 para cada estudiante, y el valor de MSE determinado usando MLR con PCA aumentó a 12,33 para cada estudiante, lo que indica que la brecha entre el método de MLR sin y con PCA es de 1,9.

Esta brecha es aceptable cuando se considera el rango de puntajes de los estudiantes de 0 a 100.

5.2 Mejora de la precisión predictiva de MLR mediante PCA

Las medidas R2 y MSE solo pueden evaluar la bondad del ajuste de un modelo de regresión, pero no pueden evaluar la precisión del modelo. Sin embargo, en la práctica, los docentes necesitan conocer la precisión del desempeño para reducir el riesgo de desperdiciar recursos a través de intervenciones incorrectas. Por lo tanto, este estudio introdujo una validación cruzada de 10 veces con barajado para dividir el conjunto de datos original en un conjunto de datos de entrenamiento y un conjunto de datos de prueba. El mecanismo de barajado permite superar el problema de errores residuales más altos influenciados por datos atípicos causados por una sola ronda de validación cruzada de 10 veces. Además, aplicamos la medida pMAPC para medir la precisión del modelo de regresión. De acuerdo con la **Fig.** 6, los valores de pMSE en las primeras 4 rondas cayeron de 503.4

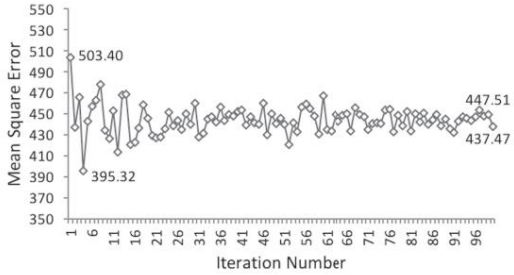


Fig. 6 Influencia de los tiempos de barajado en el valor de pMSE para MLR sin PCA.

Tabla 1 Comparación de pMSE y pMAPC entre MLR sin PCA y MLR con PCA (comp = 6).

	pMSE	pMAPC
MLR	455.87	0.71
MLR+PCA (comp=6)	198.62	0.81
p_{MSE}	<0.05	
p_{MAPC}	0.81 <0.05	

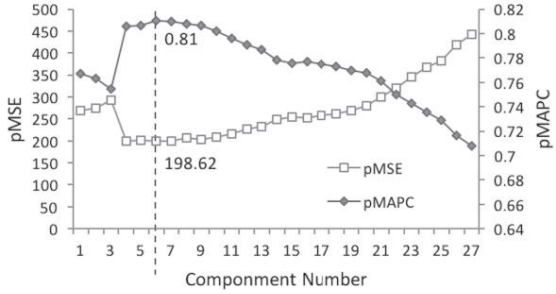


Fig. 7 Valor de pMSE y pMAPC para cada componente después de realizar PCA.

a 395.32. La diferencia máxima entre las primeras 4 rondas fue tan alta como 27%; esto fue generado por los valores atípicos en los conjuntos de datos de entrenamiento o prueba. Para reducir la diferencia entre las rondas, barajamos todos los elementos de datos después de cada ronda de valores cruzados de 10 veces. idación Después de 100 rondas de barajado, el valor de pMSE promedio podría considerarse como el valor de pMSE para el modelo de regresión.
El rango de diferencia de pMSE para las últimas rondas fue de 447,51 a 437,47, lo que significa que el rango de diferencia podría reducirse efectivamente al 2 %.

Los valores de pMSE para el modelo de regresión después de la validación cruzada se presentan en **la Tabla** 1. Los valores de pMSE y pMAPC para el modelo de regresión determinados usando MLR sin PCA fueron 455,87 y 0,71, respectivamente. En el campo de la educación, el error predictivo del rendimiento académico de los estudiantes fue cercano a 21, según el valor pMSE. Además, las puntuaciones académicas de 3 de cada 10 estudiantes se predijeron de forma incorrecta.

De acuerdo con la ecuación pMAPC de la Sección 4.3.3, podemos cuantificar fácilmente la tasa de predicción correcta, facilitando así las tareas de predicción en el campo real de la educación. La predicción incorrecta de los estudiantes en riesgo no solo aumenta los costos de enseñanza después de la escuela, sino que también influye considerablemente en la psicología de los estudiantes.

Para el modelo de regresión que involucra MLR con PCA, los valores de pMSE y pMAPC para cada componente se presentan en la **Fig.** 7, lo que revela que los valores óptimos de pMSE y pMAPC podrían obtenerse usando seis componentes. Los valores óptimos de pMSE y pMAPC fueron 198,62 y 0,81, respectivamente. Por lo tanto, el error predictivo del puntaje académico de cada estudiante fue cercano a 14;
además, los puntajes académicos de 8 de cada 10 estudiantes fueron

predicho con precisión.

Para abordar **RQ2,** se realizó una prueba t para examinar la diferencia en los valores de las medidas de desempeño predictivo entre MLR con y sin PCA. Los valores de *p* fueron menores que 0,05 para pMSE y pMAPC (Tabla 1). Por lo tanto, la predicción El rendimiento de MLR se puede mejorar considerablemente mediante el uso de seis componentes de PCA. Este resultado indica que el conjunto de datos original tenía dos propiedades: primero, existían fuertes correlaciones entre las variables independientes. Por lo tanto, el rendimiento predictivo podría mejorarse usando PCA, y seis componentes podrían usarse para obtener el mejor rendimiento predictivo. En segundo lugar, los valores atípicos tenían influyó en los componentes séptimo a vigésimo séptimo. Esto es por lo tanto, la razón principal de los valores óptimos de pMSE y pMAPC del modelo de regresión obtenido a partir de seis componentes.

5.3 Limitación

En este estudio, propusimos una metodología para establecer un modelo para predecir el rendimiento académico de los estudiantes. el modelo fue creado a partir de un conjunto de datos recopilados de Open edX y Maple TA Además, diseñamos una actividad de aprendizaje de 18 semanas que incluía tareas, cuestionarios y aprendizaje basado en videos, y fue integrado con el entorno de aprendizaje antes mencionado. En particular, el modelo de predicción está asociado con esta actividad de aprendizaje y estos atributos de datos particulares; por lo tanto, el modelo no es aplicable a otros cursos con diferentes actividades de aprendizaje y atributos de datos.

6. Conclusión

El objetivo de las analíticas de aprendizaje es mejorar el rendimiento del aprendizaje mediante la predicción de los estudiantes en riesgo y brindándoles la intervención necesaria. Con la creciente complejidad de la ambiente de aprendizaje y diversidad de herramientas de aprendizaje disponibles, Los métodos de predicción tradicionales tienen algunas limitaciones. En esto estudio, recopilamos datos de aprendizaje de la visualización de videos, ejercicios, cuestionarios y tareas en un curso de cálculo mixto para predecir rendimiento de los estudiantes. Primero, investigamos si MLR es adecuado para construir un modelo para predecir el desempeño académico de los estudiantes puntuaciones. Posteriormente, combinamos PCA y MLR para mejorar la precisión predictiva del modelo.

De acuerdo con los resultados del análisis de bondad de ajuste y residual para el modelo de regresión establecido, MLR es adecuado para construir un modelo de predicción del rendimiento académico de los estudiantes para el blended Curso de cálculo con muchas variables. Por proporcionar la predicción desempeño de los docentes, también proponemos el pMSE y el pMAPC medidas mediante la aplicación de validación cruzada. Según el análisis resultados, el rendimiento predictivo de MLR con PCA fue mayor que la de MLR sin PCA. En el futuro, el conjunto de datos original serán separados por medio examen para predecir los estudiantes en riesgo en un Etapa temprana. Para validar el rendimiento académico previsto de los estudiantes, también pretendemos proporcionar la información prevista del estudiantes en riesgo a la universidad.

Agradecimientos Este trabajo cuenta con el apoyo del Ministerio de Ciencia y Tecnología, Taiwán con subvenciones MOST 104-2511-S-008-006-MY2, MOST-105-2511-S-008-003-MY3, MOST-106-2511-S-008 -004 -MY3, MOST-105-2622-S-008 -002-CC2.

Referencias

[1] Consortium, NM et al.: El informe del horizonte de 2011 (2011).
[2] Becker, SA, Cummins, M., Davis, A., Freeman, A., Giesinger, CH, y Ananthanarayanan, V.: NMC Horizon Report: 2017 Higher Education Edition, *The New Media Consortium* (2017).
[3] Baker, RS e Inventado, PS: Aprendizaje y minería de datos educativos analítica, analítica de *aprendizaje,* págs. 61–75, Springer (2014).
[4] Papamitsiou, Z. y Economides, AA: Análisis de aprendizaje y minería de datos educativos en la práctica: una revisión sistemática de la literatura de evidencia empírica, *Journal of Educational Technology & Society,* Vol.17, No.4, p.49 (2014).
[5] Peña-Ayala, A.: *Learning Analytics: Fundamentos, Aplicaciones y Tendencias: una visión del estado actual del arte para mejorar el aprendizaje electrónico,* Vol.94, Springer (2017).
[6] Hu, Y.-H., Lo, C.-L. y Shih, S.-P.: Desarrollo de sistemas de alerta temprana para predecir el desempeño de aprendizaje en línea de los estudiantes, *Computers in Comportamiento Humano,* Vol.36, pp.469–478 (2014).
[7] Meier, Y., Xu, J., Atan, O. y van der Schaar, M.: Predicting grades, *Trans. IEEE. Procesamiento de Señales,* Vol.64, No.4, pp.959–972 (2016).
[8] Arnold, KE y Pistilli, MD: señales de curso en purdue: uso análisis de aprendizaje para aumentar el éxito de los estudiantes, *Proc. 2ª Internacional Conferencia sobre análisis de aprendizaje y conocimiento,* pp.267–270, ACM (2012).
[9] Van Leeuwen, A., Janssen, J., Erkens, G. y Brekelmans, M.: Profesores de apoyo para orientar a los estudiantes colaboradores: Efectos del aprendizaje análisis en cscl, *Computers & Education,* Vol.79, pp.28–39 (2014).
[10] Lu, OH, Huang, JC, Huang, AY y Yang, SJ: Aplicación del aprendizaje análisis para mejorar la participación de los estudiantes y los resultados de aprendizaje en un curso de programación colaborativa habilitado para moocs, *Interactivo Ambientes de aprendizaje,* Vol.25, No.2, pp.220–234 (2017).
[11] Huang, S. y Fang, N.: Predicción del rendimiento académico de los estudiantes en un curso de ingeniería dinámica: una comparación de cuatro tipos de modelos matemáticos predictivos, *Computers & Education,* Vol.61, págs. 133–145 (2013).
[12] Tempelaar, DT, Rienties, B. y Giesbers, B.: En busca de la datos más informativos para la generación de retroalimentación: análisis de aprendizaje en un contexto rico en datos, *Computers in Human Behavior,* Vol.47, pp.157–167 (2015).
[13] Zacharis, NZ: un enfoque multivariado para predecir los resultados de los estudiantes en cursos de aprendizaje combinado habilitados para la web, *Internet y Educación Superior,* Vol.27, pp.44–53 (2015).
[14] Morris, LV, Finnegan, C. y Wu, S.-S.: Seguimiento del comportamiento, persistencia y logros de los estudiantes en cursos en línea, *Internet y Educación Superior,* Vol.8, No.3, pp.221–231 (2005).
[15] Sorour, SE, Mine, T., Goda, K. y Hirokawa, S.: Un modelo predictivo para evaluar el desempeño de los estudiantes, *Journal of Information Processing,* Vol.23, No.2, pp.192–201 (2015).
[16] Yoo, J. y Kim, J.: Predecir el rendimiento del proyecto del alumno con funciones de diálogo en debates de preguntas y respuestas en línea, *Conferencia Internacional sobre Sistemas de Tutoría Inteligente,* pp.570–575, Springer (2012).
[17] Marbouti, F., Diefes-Dux, HA y Madhavan, K.: Models for early predicción de estudiantes en riesgo en un curso usando calificaciones basadas en estándares, *Computers & Education,* Vol.103, pp.1–15 (2016).
[18] Macfadyen, LP y Dawson, S.: Minería de datos de películas para desarrollar un "sistema de alerta temprana" para educadores: una prueba de concepto, Computadoras y *educación, Vol.54* , No.2, pp.588–599 (2010).
[19] Agudo-Peregrina, AF, Iglesias-Pradas, S., Conde-González, M. y Hernandez-Garc´ıa, A.: ¿Podemos predecir el éxito a partir de los datos de registro en vles? clasificación de interacciones para análisis de aprendizaje y su relación con el rendimiento en aprendizaje en línea y f2f con soporte de vle , Computers *in Human Behavior,* Vol.31, pp.542–550 (2014).
[20] Taneja, A. y Chauhan, R.: Un estudio de rendimiento de las técnicas de minería de datos: regresión lineal múltiple frente a análisis factorial, versión preliminar de arXiv arXiv: 1108.5592 (2011).
[21] O'Connell, RT y Koehler, AB: *Pronósticos, series temporales y regresión: un enfoque aplicado,* Vol.4, South-Western Pub (2005).
[22] Hyndman, RJ y Koehler, AB: Otra mirada a las medidas de precisión *de los pronósticos, International Journal of Forecasting,* Vol.22, No.4, págs. 679–688 (2006).
[23] Jolliffe, IT: Análisis de componentes principales y análisis factorial, *Prin análisis de componentes principales,* págs. 115–128, Springer (1986).
[24] Hira, ZM y Gillies, DF: Una revisión de la selección de funciones y la métodos de extracción aplicados a datos de micromatrices, *Avances en bioinformática,* Vol. 2015 (2015).
[25] Ul-Saufie, A., Yahya, A. y Ramli, N.: Mejora del modelo de regresión lineal múltiple mediante el análisis de componentes principales para predecir la concentración de PM10 en Seberang Prai, Pulau Pinang, *International Revista de Ciencias Ambientales,* Vol.2, No.2, p.403 (2011).
[26] Qiuhua, L., Lihai, S., Tingjing, G., Lei, Z., Teng, O., Guojia, H., Chuan, C. y Cunxiong, L.: Uso de puntajes de componentes principales en

modelos de regresión lineal múltiple para la simulación de clorofila-a y abundancia de fitoplancton en un reservorio kárstico profundo, al suroeste de china, *Acta Ecologica Sinica,* Vol.34, No.1, pp.72–78 (2014).
[27] Pires, J., Martins, F., Sousa, S., Alvim-Ferraz, M. y Pereira, M.: Selección y validación de parámetros en múltiples lineales y principales regresiones de componentes, *Environmental Modeling & Software,* Vol.23, N.° 1, págs. 50–55 (2008).
[28] Goossens, M., Mittelbach, F. y Samarin, A.: *The LaTeX Compan Ion,* Addison Wesley (1993).
[29] Lamport, L.: *Un sistema de preparación de documentos Guía del usuario de LATEX* y *Manual de referencia,* Addison Wesley (1986).

Stephen JH Yang es ahora el vicepresidente de la Universidad de Asia, Taiwán. Él es también asociado con la Central Nacional Universidad como Profesor Distinguido del Departamento de Ciencias de la Computación y Ingeniería de Información. El Dr. Yang fue el Director del Departamento de Información y Educación Tecnológica, Ministerio de Educación, Taiwán (2013–2014), durante los dos años de servicio en el gobierno de Taiwán, el Dr. Yang fue responsable de la educación en información y tecnología, también lanzó Taiwán iniciativa nacional de aprendizaje digital que incluye la construcción de 100 G Red Académica de Taiwán para infraestructura de red nacional estructura, la construcción de Education Cloud para datos nacionales infraestructura y programas de innovación como los MOOC de Taiwán y aprendizaje móvil. El Dr. Yang también se desempeñó como Coordinador de Disciplina de Educación en Información, Ministerio de Ciencia y Tecnología. El Dr. Yang recibió su Ph.D. Licenciatura en Ingeniería Eléctrica y Ciencias de la Computación de la Universidad de Illinois en Chicago en 1995. El Dr. Yang ha publicado más de 70 artículos en revistas SSCI/SCI, sus intereses de investigación incluyen Big Data, análisis de aprendizaje, inteligencia artificial en educación, minería de datos educativos y MOOC. Como se muestra en Google Scholar, los índices de citas de la publicación del Dr. Yang han sido de más de 8.800, especialmente en la página principal. temas de investigación, la minería de datos educativos ocupa el puesto n.° 3, MOOC ocupa el puesto n.º 3, la inteligencia artificial en la educación ocupa el puesto n.º 7, El análisis de aprendizaje ocupa el puesto número 8. El Dr. Yang recibió el Premio a la Investigación Sobresaliente del Ministerio de Ciencia y Tecnología (2010) y Medalla por Servicios Distinguidos del Ministerio de Educación (2015). El Dr. Yang es actualmente el coeditor en jefe de la Revista Internacional de Gestión del Conocimiento y e-Learning.

Owen HT Lu es estudiante de informática Ingeniería en Ciencias e Información, Universidad Nacional Central, y también la Gerente de Sección del Instituto de Sistemas Inteligentes, Instituto para la Industria de la Información, Taiwán. El Sr. Lu recibió su título de maestría en Departamento de Ingeniería Electrónica y Comunicaciones y Nacional Chung Hsing University en Taiwán en 2009. Sus intereses de investigación incluyen Cloud Computing, Big Data Technology, Data Security y Learning Analytics.

Anna Yu-Qing Huang es posdoctoral Investigador de Ciencias de la Computación e Ingeniería de la Información, Central Nacional Universidad, Taiwán. El Dr. Huang recibió su doctorado Licenciado en el Instituto de Ingeniería, Ciencia y Tecnología de la Primera Universidad Nacional de Ciencia y Tecnología de Kaohsiung en Taiwán en 2011. Sus intereses de investigación incluyen la tecnología Big Data, Learning Analytics, mobile learning, Massive Open Online Courses (MOOC), aprendizaje colaborativo asistido por computadora (CSCL).

Jeff Cheng-Hsu. Huang es asociado Profesor de Ciencias de la Computación e Ingeniería de la Información, Universidad de Tecnología Hwa Hsia, Taiwán. doctor huang recibió su Ph.D. grado en informatica Ingeniería en Ciencias e Información de Universidad Nacional Central de Taiwán en 2009. Sus intereses de investigación incluyen e aprendizaje, aprendizaje móvil, redes sociales, computación social, Mundo virtual 3D, diseño creativo, aprendizaje colaborativo asistido por computadora (CSCL), tecnología Big Data, análisis de aprendizaje.

Hiroaki Ogata es profesor en Learning e Investigación en Tecnologías Educativas Unidad, el Centro Académico de Estudios de Computación y Medios, y el Posgrado Facultad de Informática de la Universidad de Kyoto, Japón, y también profesor honorario de la Universidad de Educación de Hong Kong, y profesor visitante de la cátedra de Universidad de Asia en Taiwán también.

Albert JQ Lin es un ingeniero de Taiwán Empresa fabricante de semiconductores. El Sr. Lin recibió su maestría en Ciencias de la Computación e Ingeniería de la Información de la Universidad Nacional Central en Taiwán en 2017. Sus intereses de investigación en cluir Big Data Tecnología y Aprendizaje Analítica.