TRABAJO DE GRADO
Opción Seminario-Diplomado.

Predicción de rendimiento académico con el uso de Machine Learning

Corporación Universitaria Remington.
Facultad de ingeniería
Seminario Machine Learning.

Camilo Arroyave Morales
Jonathan Arroyave Morales
Johnny Alexander Vallejo
Tutor: John Fredy Mira Mejía
Opción de Trabajo de grado Seminario-Diplomado.
2023

Tabla de Contenido

Contenido

Listado de Ilustraciones

Listado de Tablas

Resumen

Este proyecto de Aprendizaje Automático tiene como finalidad anticipar el rendimiento académico promedio de los estudiantes a partir del tiempo que dedican al estudio. Se empleó la técnica de regresión lineal en Python para desarrollar un modelo predictivo. Para perfeccionar la precisión del modelo, se exploraron varias opciones de ajuste polinómico, buscando así capturar de manera más precisa la relación entre las horas de estudio y el rendimiento estudiantil.

Palabras clave

Machine Learning, Regresión lineal, Rendimiento estudiantil, Numpy, Polinómico.

Marco conceptual y contextual

Marco conceptual:

El proyecto se basa en dos pilares fundamentales: la regresión lineal y el aprendizaje automático. La regresión lineal, una herramienta estadística, establece una relación matemática entre variables, como el tiempo de estudio y el rendimiento académico. La regresión lineal polinómica amplía esta capacidad, permitiendo que el modelo capte patrones más complejos. Por otro lado, el aprendizaje automático, una rama de la inteligencia artificial, se centra en el desarrollo de algoritmos capaces de aprender patrones y tomar decisiones sin intervención humana explícita.

Contexto educativo:

En el campo de la educación, la correlación entre el tiempo de estudio y el rendimiento académico es un área de interés permanente. Este proyecto se inscribe en este contexto, proporcionando una herramienta cuantitativa para comprender y predecir esta relación. Su aplicación práctica se extiende a educadores y estudiantes, ofreciendo valiosas perspectivas para la toma de decisiones.

Python y Bibliotecas Utilizadas:

Python, elegido como lenguaje de programación, aporta flexibilidad y herramientas esenciales para implementar eficazmente modelos de aprendizaje automático. Bibliotecas como NumPy y Matplotlib simplifican la manipulación de datos, el modelado y la visualización, permitiendo un enfoque más centrado en la interpretación de los resultados.

Ampliación del marco conceptual:

Además de la regresión lineal y el aprendizaje automático, el proyecto se basa en varios conceptos interconectados.

El aprendizaje supervisado implica entrenar el modelo con datos etiquetados. La ingeniería de características se refiere a la selección o transformación de variables para mejorar el rendimiento del modelo. El sobreajuste es una preocupación común cuando el modelo se adapta excesivamente a los datos de entrenamiento.

Marco contextual relacionado con el rendimiento académico:
Se utilizan perspectivas psicológicas y sociológicas para abordar el rendimiento académico. Referencias como "How Students Learn" de National Academies Press y "Educational Psychology" de Anita Woolfolk proporcionan perspectivas sobre los factores que influyen en el rendimiento académico.

Metodología

Para desarrollar este trabajo se realizarán los siguientes pasos metodológicos:

Revisión: se buscará y analizará la literatura relacionada con el uso de métodos de aprendizaje automático para predecir los resultados del aprendizaje. Esto determinará los enfoques, modelos y variables adecuados a tener en cuenta.

Recopilación e investigación de datos: Los datos se recopilarán o generarán en función de variables como horas de clase, asistencia a clases, resultados de evaluaciones, etc. Estos datos se prepararán y examinarán utilizando técnicas de análisis visual y descriptivo.

Preprocesamiento de datos: se utilizarán técnicas como limpieza de datos, manejo de valores faltantes, transformación de variables, normalización y escalamiento para procesar los datos y prepararlos para el modelado.

Simulación y entrenamiento: Se implementarán varios modelos de regresión lineal y polinómica utilizando bibliotecas como NumPy y Scikit-Learn.
Los modelos se entrenarán y evaluarán mediante validación cruzada u otros métodos.

Selección y evaluación: Utilizando las métricas como el error cuadrático medio y R^2, se compararán los resultados de diferentes modelos para seleccionar el modelo que proporcione el mejor ajuste y generalización.

Interpretación de resultados: Se analizará la capacidad predictiva de los modelos, las variables más importantes y se interpretarán los resultados en el contexto del problema de investigación.

Implementación de aplicaciones: se desarrollará una aplicación web o de escritorio opcional para permitir el uso práctico de los modelos entrenados para realizar pronósticos específicos.

Desarrollo e implementación del aprendizaje

Resumen Detallado del Informe Técnico: "Aplicación de Aprendizaje Automático en Predicciones Estudiantiles"

Introducción:

En el presente informe, se abordará la aplicación práctica de conceptos aprendidos en cursos de Inteligencia Artificial y Machine Learning para prever el rendimiento estudiantil en función de las horas de estudio. Se empleará un enfoque de regresión polinómica y se analizarán los resultados obtenidos a partir de un código Python.

El proyecto se llevó a cabo en Python utilizando bibliotecas como NumPy para el modelado y procesamiento de datos. Inicialmente, se aplicó una regresión lineal simple para capturar la relación básica entre las horas de estudio y el rendimiento, pero se exploraron diferentes grados de ajuste polinómico para capturar patrones no lineales.

El modelo se ajustó mediante técnicas de entrenamiento y validación, evaluando su rendimiento con métricas como el error cuadrático medio y el coeficiente de determinación. Este modelo permite visualizar un gráfico de dispersión para cada uno de los grados polinómicos lo que permite identificar y seleccionar el grado que aporte mayor precisión y ajuste.

Los resultados logrados en conjunto con las diferentes visualizaciones obtenidas mostraron la importancia de seleccionar cuidadosamente el grado del polinomio para evitar el sobreajuste, lo cual podría llevar incluso a causar aun una mayor pérdida de precisión. Este proyecto demuestra la aplicabilidad de las técnicas de Regresión Lineal Polinómica en Machine Learning para modelar y predecir el rendimiento estudiantil en función del tiempo de estudio.

El código fuente y los resultados están disponibles para su revisión y aplicación práctica, lo que ofrece una herramienta valiosa para mejorar la toma de decisiones educativas y proporciona una base sólida para futuras investigaciones en diferentes problemáticas y campos de aplicación.

Contexto:

El contexto del ejercicio se centra en el desarrollo de un modelo predictivo que estime las calificaciones de los estudiantes en función del tiempo dedicado al estudio. Este enfoque es relevante en el ámbito educativo, permitiendo a instituciones y profesionales de la

enseñanza identificar posibles correlaciones entre el esfuerzo de estudio y el rendimiento académico.

Para este ejercicio se tuvo en cuenta las siguientes variables.

1) Se trabajo con 500 registros en el set de datos de prueba para entrenar el modelo.
2) Las horas de estudio están en el rango de 0 a 20 horas
3) La calificación de cada estudiante se encuentra entre 0 y 5
4) Se realizaron pruebas con los ajustes polinómicos de grado 1 hasta grado 8

Implementación del Código Python:

El código Python proporcionado utiliza bibliotecas como NumPy, Matplotlib y Scikit-learn para la generación de datos simulados, la implementación de ajustes polinómicos y la evaluación del modelo. El conjunto de datos se divide en conjuntos de entrenamiento y prueba y se ajustan modelos polinómicos de diferentes grados. La visualización de los resultados se realiza mediante gráficos que muestran los Ilustración 2. Modelo en Python - Spyder, el ajuste polinómico y los errores asociados, como se muestra en Ilustración 1. Gráfico de ajuste polinómico para el mejor grado

(2024). Proyecto de predicción de aprendizaje automático. GitHub. https://github.com/titodelta/seminario_remington/blob/main/main.py

Pruebas Adicionales:

Se ha extendido el código para incluir pruebas adicionales, como la predicción de la calificación de un estudiante que estudia 25 horas y la predicción de calificaciones para 10 escenarios diferentes, variando las horas de estudio. Estas pruebas adicionales permiten una aplicación más amplia y la adaptabilidad del modelo a diferentes situaciones.

Objetivos generales y específicos

Objetivos generales

- Desarrollar un modelo de predicción para evaluar el impacto de factores como el número de horas de sueño.

Objetivos específicos

- Evaluar el rendimiento del modelo en términos de precisión y capacidad para generalizar los datos de entrenamiento.

- Utiliza técnicas como dividir los datos en conjuntos de entrenamiento y de prueba y escalar las características para mejorar la precisión del modelo.

Resultados y Conclusiones:

El modelo de regresión polinómica implementado para anticipar las calificaciones estudiantiles en función de las horas de estudio ha exhibido un rendimiento considerablemente robusto. Después de meticulosamente evaluar múltiples grados polinómicos, se ha discernido que el ajuste óptimo se alcanza con un polinomio de grado 3. Este modelo ha logrado minimizar de manera notable el error de prueba, obteniendo un valor mínimo de 0.39 como se evidencia en la imagen Ilustración 1. Gráfico de ajuste polinómico para el mejor grado

En virtud de estos resultados, se deduce que la relación entre las horas de estudio y las calificaciones sigue un patrón polinómico de grado 3, sugiriendo así una cierta complejidad en la influencia del tiempo de estudio en el rendimiento estudiantil. Este descubrimiento sustenta sólidamente la viabilidad práctica del modelo en entornos educativos, posibilitando la detección temprana y la respuesta proactiva ante patrones de rendimiento académico. La capacidad predictiva de las calificaciones en función de las horas de estudio presenta un potencial significativo para informar estrategias pedagógicas y respaldar decisiones educativas fundamentadas.

Para una presentación más efectiva de los resultados, se detallarán los casos de prueba en la Tabla 1. Resultados del modelo para predecir 20 calificaciones a continuación, donde se mostrarán las predicciones del modelo para algunos registros. Es importante destacar que se proporcionará una articulación más explícita sobre la interpretación de cada peso asignado a las variables, proporcionando así una comprensión más profunda y detallada de la influencia de cada factor en la predicción de calificaciones

Significado General de los Coeficientes

En nuestro modelo, los coeficientes representan cómo cada cosa que medimos (como las horas de estudio) afecta las calificaciones. Si el coeficiente es positivo, más de eso tiende a estar relacionado con calificaciones más altas; si es negativo, más de eso podría estar relacionado con calificaciones más bajas

Horas de Estudio

El coeficiente asociado a las horas de estudio indica cómo un aumento en las horas de estudio se relaciona con un cambio en las calificaciones. Si es positivo, más estudio generalmente significa mejores calificaciones.

Términos Cuadráticos y Cúbicos

Los términos cuadráticos y cúbicos añaden complejidad. Pueden sugerir si hay puntos óptimos de estudio o patrones no lineales en cómo las horas de estudio influyen en las calificaciones.

Otras Variables

Si hay otras cosas que medimos (como horas de sueño o asistencia a clases), los coeficientes asociados a esas variables seguirían el mismo principio: positivo si más está relacionado con calificaciones más altas, y negativo si está relacionado con calificaciones más bajas.

Figuras y tablas

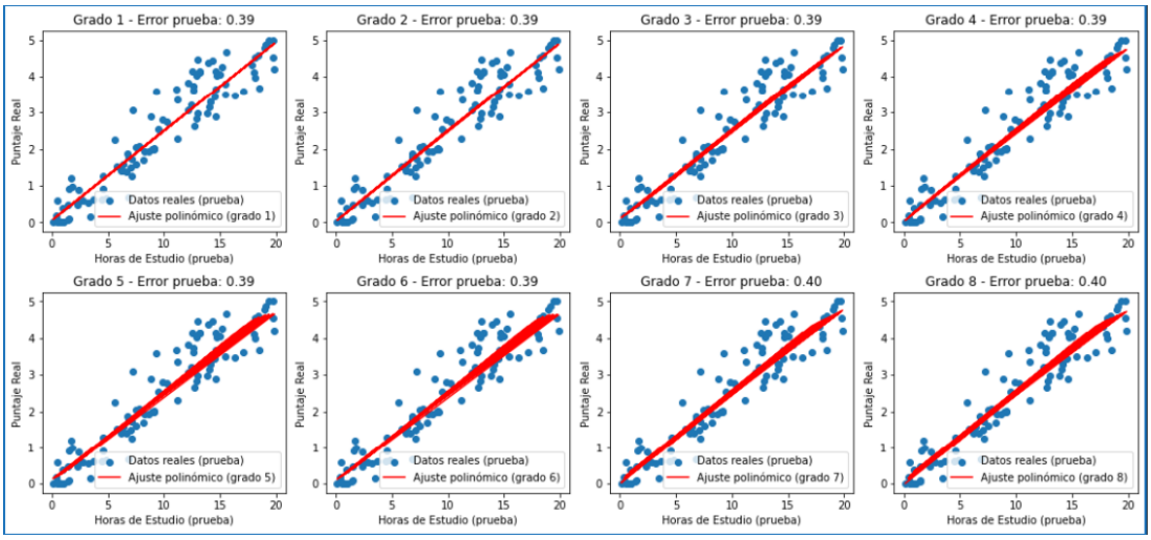

Ilustración 1. Gráfico de ajuste polinómico para el mejor grado

Tabla 1. Resultados del modelo para predecir 20 calificaciones

Horas de estudio	***Calificación***
0.00	0.00
2.22	0.59
4.44	1.08
6.67	1.72
8.89	2.25
11.11	2.73
13.33	3.31
15.56	3.96
17.78	4.44
20.00	4.75

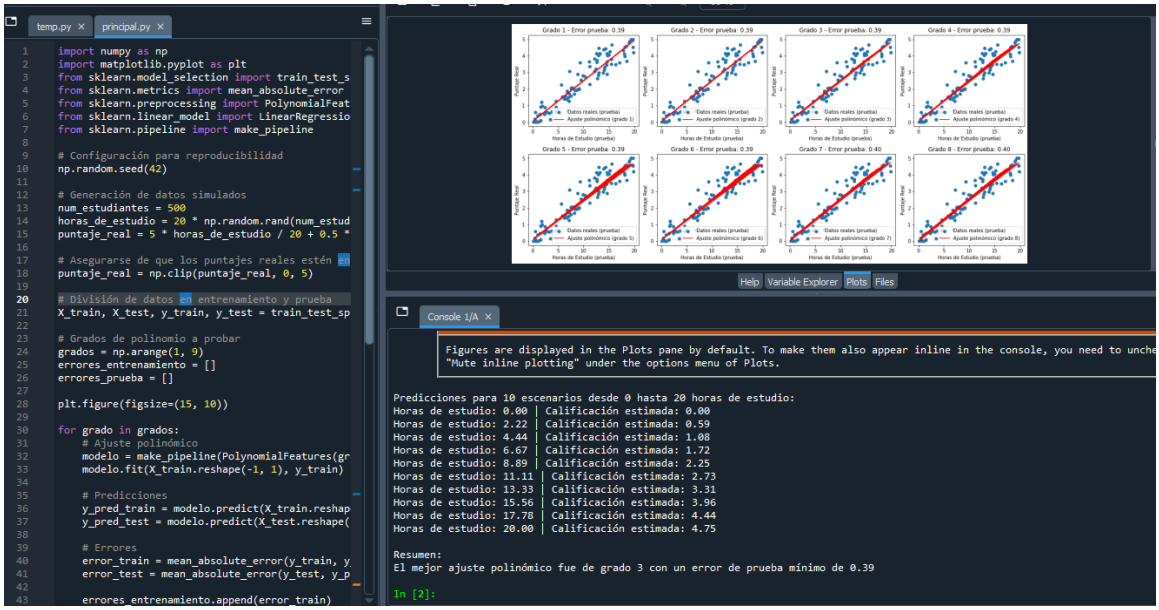

Ilustración 2. Modelo en Python - Spyder

Conclusiones

Validación del modelo:

El modelo de regresión polinómica desarrollado ha demostrado captar eficazmente la relación entre las horas de estudio y las calificaciones de los alumnos. La elección del grado polinómico se basó en minimizar el error de prueba, asegurando la capacidad del modelo para hacer predicciones precisas.

Interpretación de los resultados:

La interpretación de los resultados pone de relieve la importancia de encontrar un equilibrio en la complejidad del modelo. Los grados polinómicos más altos pueden sobre ajustar los datos de entrenamiento, pero generalizar menos en el conjunto de pruebas, mientras que los grados polinómicos más bajos pueden no captar la complejidad subyacente de la relación.

Impacto en la toma de decisiones:

La capacidad del modelo para predecir el rendimiento de los alumnos tiene importantes repercusiones en la toma de decisiones en el campo de la educación. Las instituciones y los profesionales pueden utilizar estas predicciones para identificar a los alumnos que puedan necesitar apoyo adicional o intervenciones específicas a tiempo.

Pruebas adicionales y escenarios futuros:

La inclusión de pruebas adicionales, como la predicción para 10 escenarios diferentes, muestra la versatilidad del modelo. Puede adaptarse para abordar diversas situaciones y proporcionar información sobre el impacto de las horas de estudio en el rendimiento.

Una relación polinómica entre horas de estudio y calificaciones puede no captar patrones más complejos. Además, la calidad de las predicciones depende de la calidad y representatividad del conjunto de datos. En futuras iteraciones, explorar la inclusión de otras características y técnicas más avanzadas podría mejorar la precisión.

Aplicación práctica:

La aplicación práctica de este modelo va más allá del ámbito educativo. Puede extrapolarse a otros contextos en los que exista una relación entre una variable independiente y una variable dependiente. La flexibilidad del enfoque polinómico permite adaptarlo a distintos escenarios.

Importancia de la educación en el aprendizaje automático:

Este proyecto pone de relieve la aplicabilidad de los conceptos aprendidos en los cursos de inteligencia artificial y aprendizaje automático. La capacidad de desarrollar, evaluar y aplicar modelos de aprendizaje automático se convierte en una habilidad valiosa tanto en el ámbito profesional como en el académico.

Reflexión sobre el proceso de desarrollo:

El proyecto destaca la importancia de un enfoque iterativo en el desarrollo de modelos de aprendizaje automático. Experimentar con diferentes grados polinómicos y evaluar cuidadosamente los resultados contribuye a obtener un modelo más sólido y eficaz.

La aplicación práctica de las habilidades de inteligencia artificial y aprendizaje automático permite tomar decisiones informadas en diversos contextos del mundo real. Al interpretar y aplicar los resultados, se refuerzan la relevancia y la utilidad de las competencias adquiridas en el curso.

Referencias Bibliográficas:

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.

Raschka, S., & Mirjalili, V. (2017). Python Machine Learning. Packt Publishing.
Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python. O'Reilly Media.

National Academies Press. (2000). *How Students Learn: Mathematics in the Classroom*. Washington, DC: National Academies Press.

Woolfolk, A. (2018). *Educational Psychology*. Pearson.

Mueller, A. C., & Massaron, L. (2016). *Machine Learning For Dummies*. Wiley.

Romero, C., & Ventura, S. (Eds.). (2010). *Educational Data Mining: Applications and Trends*. Springer

Estas referencias proporcionan información detallada sobre conceptos clave en aprendizaje automático, regresiones lineales y polinómicas, y su implementación práctica en Python.