Universidad Tecnológica de Panamá

Ingeniería en Sistemas y Computación

Licenciatura en ingeniería de sistemas computacionales

Proyecto Semestral de Probabilidad Aplicada a TIC

Docente: Juan Marcos Castillo, PhD

Integrantes: Mercado, José 8-1045-2452

Ayala, Ariel 8-1037-61

Grupo: 1IL-124

Introducción:

En el contexto del análisis de datos y la estadística, las bases de datos juegan un papel fundamental al almacenar grandes volúmenes de información que pueden ser estudiados mediante herramientas probabilísticas. La probabilidad permite analizar la incertidumbre inherente a muchos fenómenos del mundo real, y cuando se combina con bases de datos, se convierte en una herramienta poderosa para la toma de decisiones informadas.

El uso de bases de datos en probabilidad facilita la recopilación, organización y procesamiento de datos que se utilizan para calcular medidas estadísticas como la media, la varianza o la probabilidad de eventos. Por ejemplo, en un sistema de gestión de pacientes en un hospital, una base de datos puede almacenar información sobre enfermedades, tratamientos y resultados, lo que permite aplicar modelos probabilísticos para predecir riesgos, necesidades o comportamientos futuros.

Además, el análisis probabilístico en bases de datos también es esencial en campos como la inteligencia artificial, la minería de datos, y la ingeniería de software, donde es crucial estimar la probabilidad de eventos basados en patrones históricos. Con el crecimiento de los datos digitales, se ha vuelto cada vez más importante comprender cómo aplicar técnicas probabilísticas sobre grandes volúmenes de información de forma eficiente y confiable.

Justificación:

Nosotros escogimos esta base de datos ya que los estudiantes están presentes en dicha base, el sentido de representación es notorio ya que hay bastantes estudiantes que trabajan y /o están presentes en actividades extracurriculares dentro o fuera del complejo universitario, lo que dificulta o reduce la organización o distribución correcta y adecuada del tiempo, aunque todos las personas son diferentes, hay casos similares de la misma organización del tiempo que tienen promedios completamente diferentes, gracias a esas diferencias, surgió el interés en escoger esta base de datos , con el fin de poder analizarla y representar los distintos resultados reflejados según sus respectivas variables.

Además este proyecto nos puede guiar o recomendar a organizar mejor los presentes tiempos de estudio, esto con el fin de considerar si es lo mas adecuado tener actividades extracurriculares o un trabajo de medio tiempo, sin afectar los estudios universitarios; aunque según la base de datos , la carrera influye en un pequeño porcentaje al índice , ya que según los datos analizados , los estudiantes con la mismas variables solo se les vio afectado el índice por la carrera que escojieron; esto quiere decir que como tal la carrera no es que sea fácil, sino que presenta un grado de complejidad menor, lo que otorga mayor tiempo disponible sin tener que afectar el desempeño en GPA.

Antecedentes

La facultad detectó un aumento en la preocupación por el bajo rendimiento y la deserción temprana.

Se planteó identificar factores académicos y personales que expliquen el GPA y ayuden a predecir quién podría necesitar apoyo.

Fuente de datos

Encuesta institucional aplicada al cierre del semestre 2024-1 (n ≈ 300 estudiantes).

Variables recogidas:

· Demográficas: Gender, Age

· Académicas: StudyHoursPerWeek, AttendanceRate, GPA, Major

· Situacionales: PartTimeJob (Sí/No), ExtraCurricularActivities (Sí/No)

Definición del problema

La incógnita que formulamos con los datos obtenidos fue: Las horas de estudio, asistencias, trabajos y/o actividades extracurriculares afectan en nuestro desempeño académico?

Esta incógnita la formulamos ya que eran las variables presentes en la base de datos, aunque también tenemos la edad; esta variable representa un cambio mínimo entre los diferentes índices, así que la interrogante principal es la anterior mencionada.

Al analizar las variables en nuestra base de datos; nos dimos cuenta de que, a mayor tiempo de estudio disponible, asistencia casi perfecta y no tener responsabilidades laborales; influyen positivamente en el índice mostrando que la mayoría de estos estudiantes presentan un índice arriba de 2.75.

Por el contrario, si se presenta muy pocas veces y no le dedica mucho tiempo de estudio, ya sea por trabajo o actividades fuera del periodo universitario, el índice decrece significativamente.

Aunque quiero resaltar que hay estudiantes con las horas de asistencia y de estudio al mínimo, presentan un índice elevado y viceversa; esto lo conocemos como puntos atípicos, es decir valores que no coinciden con el orden o patrones de los demás.

Análisis con diferentes modelos de estocásticos

- **Determinación de la base de datos:**

La base de datos proporcionados por el docente y elegida de manera colectiva por los compañeros de equipo es de la plataforma ¨Kaggle¨más detalle Performance Data Set , que presenta una base de datos no depurados, pero sin datos basura , es decir que las pocas variables que posee a diferencia de las demás variables de las otras bases de datos, son de mucha utilidad e indispensables para el correcto análisis de la misma.

- **Pre-procesamiento y limpieza:**

Para este paso nosotros lo que tuvimos que realizar fue primordialmente eliminar registros con datos imposibles (ej.: GPA negativo o > 4), después optamos por convertir columnas numéricas con texto (ej.: “75 %”) a formato decimal y por último estandaricé las etiquetas de las variables categóricas (sexo, carrera, etc.).

- **Análisis descriptivo:**

Al analizar la base de datos, luego de organizar las respectivas variables utilizamos histogramas para poder tener una mejor representación visual de los datos, con el fin de un mejor entendimiento de los datos, además utilizamos los métodos estadísticos básicos como la moda, la media, la varianza; con el fin de reconocer patrones entre estudiantes con datos similares.

- **Selección de variables:**

Las variables presentes en la base de datos que analizamos fueron las siguientes:

- La edad que era un valor numérico entero
- Las horas de estudio que también representaban un valor numérico entero
- El porcentaje de asistencias que se representaba como valor en decimal
- El GPA o índice que se representaba como valor en decimal

- Los trabajos de medio tiempo y las actividades extracurriculares se representaban como booleanos, es decir verdadero o falso

- **Selección de Modelos:**

El modelo que utilizamos fue acerca de los Métodos Estadísticos Básicos, en concreto el modelo ARIMA (AutoRegressive Integrated Moving Average): ya que se encarga de predecir valores futuros basados en patrones históricos. También utilizamos la regresión logística y lineal múltiple.

Conclusiones

Ariel Ayala

El panorama descriptivo muestra un alumnado bastante homogéneo: la mayoría tiene entre 18 y 24 años, estudia unas 15 – 40 horas por semana, mantiene asistencias superiores al 70 % y sitúa su GPA alrededor de 2.8 – 3.4.

Las variables clave (edad, horas de estudio, asistencia) exhiben correlaciones moderadas entre sí, pero muy débiles con el GPA.

En el bloque predictivo, la regresión logística logró detectar con cierta utilidad a los estudiantes en riesgo (≈ 73 % F1), pero la regresión lineal múltiple apenas explicó el GPA ($R^2 \approx 0.01$). Esto confirma que:

El rendimiento académico no depende linealmente de los factores disponibles.

Probablemente influyen variables omitidas (motivación, calidad docente, estrés, hábitos de sueño, etc.).

José Mercado

Luego de haber desglosado y ordenado el presente archivo de base de datos; pudimos observar que la edad es un valor que representa un cambio mínimo en el GPA, seguido de las carreras que tienen una leve influencia en el índice; sin embargo, las que más presentan influencia son las horas de estudio y porcentaje de asistencias.

Los factores presentes en las variables no representan con exactitud su correcto desempeño reflejado en su índice, esto quiere decir que la persona puede experimentar factores alternos que afectan su correcto desempeño

Recomendaciones prácticas

Ariel Ayala

- Emplear la clasificación “bajo rendimiento” (modelo logístico/RandomForest) para disparar tutorías y seguimiento antes del examen final.
- Priorizar a quienes combinan: asistencia < 70 %, < 15 h de estudio, y trabajan medio tiempo.

Futuros estudios

- Modelos no lineales y conjuntos (Gradient Boosting, XGBoost) para capturar posibles umbrales y relaciones complejas.
- Análisis de supervivencia académica: ¿en qué semestre abandonan los estudiantes y por qué?
- Estudios longitudinales para medir cómo cambios en hábitos (más horas de estudio o asistencia) se traducen en incrementos de GPA.

José Mercado

Tratar de eliminar los datos que tengan menos influencia, como en nuestro caso la edad ya que los estudiantes a los que fueron encuestados tienen un promedio de entre 18 a 24 años, y según estudios, los jóvenes procesan el estrés académico de manera mas deficiente que un adulto, pero la comparativa de este estudio es de jóvenes de 18 años y adultos de 30 a 35 años.

Futuros estudios:

Implementarle a la base de datos el nivel de estrés promedio que presenta un estudiante, si su trabajo requiere un gasto considerable de energía o si trabaja de manera sedentaria su medio tiempo; además de agregar la variable de

facilitaciones, como por ejemplo el acceso al internet o a la biblioteca con el fin de poder adquirir conocimientos adicionales que le puedan servir a futuro.

Bibliografía

Influencia de la edad en los estudios

Anexos

1. https://github.com/ArieUr25/Proyecto-final/blob/262cbbfe4acf4bcbf9e993285764b3c12abc9b94/README.md
2. https://github.com/JoseJulian1901/Jose-Mercado