INFORME DEL MÉTODO DE LA INGENIERÍA

IDENTIFICACIÓN DEL PROBLEMA

Identificación de necesidades

- Mostrar una tabla en la cual se pueda apreciar la información de cada uno de los estudiantes.
- Mostrar los 5 factores más representativos de la información en un gráfico cada uno.
- Predecir las notas de los exámenes finales de futuros estudiantes, clasificándolos en base a la información que se tiene del dataset.

Definición del problema

El departamento de educación de Lisboa, Portugal ha decidido realizar una investigación frente al desempeño académico de los estudiantes de secundaria de la ciudad. Para esto ha tomado una muestra de dos escuelas: Gabriel Pereira y Mousinho da Silveira.

El objetivo del departamento con este estudio es comprender cómo afectan los distintos factores tales como: Tipo de asentamiento (Urbano o rural), nivel de educación de los padres, tiempo semanal dedicado a estudiar, etc. Con esta información se busca poder predecir la nota que posteriores estudiantes teniendo situaciones similares podrían obtener en su examen final.

Por último, han manifestado que les gustaría ver gráficamente la información que han recolectado utilizando tablas y diagramas.

RECOPILACIÓN DE INFORMACIÓN

Definiciones

Clasificación:

Relación de los clasificados en una determinada prueba.[1]

Clasificar:

Ordenar o disponer por clases algo.[2]

Clasificación (ML): Proceso de categorizar un set de datos en clases, puede ser realizado en datos estructurados o no estructurados. El proceso empieza prediciendo la clase de puntos de datos dados. Las clases son comúnmente referidas como objetivo, marcador o categoría.[3]

Base de datos:

Una base de datos es una colección organizada de información estructurada, o datos, típicamente almacenados electrónicamente en un sistema de computadora.[4]

Diagrama de barras:

Es un gráfico que se utiliza para representar datos de variables cualitativas o discretas. Está formado por **barras** rectangulares cuya altura es proporcional a la frecuencia de cada uno de los valores de la variable.[5]

Machine Learning:

El aprendizaje automático(Machine learning) es una aplicación de inteligencia artificial (IA) que proporciona a los sistemas la capacidad de aprender y mejorar automáticamente a partir de la experiencia sin estar programados explícitamente. El

---

[1] "clasificación | Definición | Diccionario de la lengua española | RAE ...." https://dle.rae.es/clasificaci%C3%B3n.
[2] "clasificar | Definición | Diccionario de la lengua española | RAE ...." https://dle.rae.es/clasificar.
[3] "Classification In Machine Learning | Classification ... - Edureka." 21 jul. 2020, https://www.edureka.co/blog/classification-in-machine-learning/.
[4] "¿Qué es una base de datos? | Oracle Colombia." https://www.oracle.com/co/database/what-is-database/.
[5] "Diagrama de barras - Universo Formulas." 14 abr. 2014, https://www.universoformulas.com/estadistica/descriptiva/diagrama-barras/. Se consultó el 9 abr. 2021.

aprendizaje automático se centra en el desarrollo de programas informáticos que pueden acceder a los datos y utilizarlos para aprender por sí mismos.[6]

Diagrama de pastel:

Un diagrama de pastel es un círculo dividido en partes, donde el área de cada parte es proporcional al número de datos de cada categoría.[7]

Diagrama de puntos:

Un diagrama de puntos es una gráfica utilizada para ilustrar un número reducido de datos, la cual permite identificar con facilidad dos características:

1. La localización de los datos.

2. La dispersión o variabilidad de los datos.

Este diagrama muestra cada uno de los elementos de un conjunto de datos numéricos por encima de una recta numérica.[8]

---

[6] "What is Machine Learning? A definition - Expert System | Expert.ai." https://www.expert.ai/blog/machine-learning-definition/. Se consultó el 9 abr. 2021.
[7] "GRÁFICA DE PASTEL - UNAM." http://asesorias.cuautitlan2.unam.mx/Laboratoriovirtualdeestadistica/DOCUMENTOS/TEMA%201/5.%20GRAFICA%20DE%20%20PASTEL.pdf. Se consultó el 9 abr. 2021.
[8] "Diagrama de Puntos - Temas de Estadística." 16 ene. 2013, http://tsu-estadistica.blogspot.com/2013/01/diagrama-de-puntos.html. Se consultó el 9 abr. 2021.

Requerimientos funcionales:

1. Leer la base de datos(obtenida de http://archive.ics.uci.edu/ml/datasets/Student+Performance#) y guardar la información dentro de una lista. Donde cada columna es un factor y cada fila es un estudiante.
2. Mostrar 5 gráficos que representen: Nombre de la escuela a la que pertenece, tiempo libre, tiempo de estudio, tipo de asentamiento y acceso al internet.
3. Clasificar nuevos estudiantes entregando la nota final que estos obtendrán con base en sus factores, utilizando como herramienta el dataset previamente obtenido.
4. Mostrar a todos los estudiantes del dataset por medio de una tabla obtenida de la base de datos con los siguientes campos:
    - Escuela
    - Sexo
    - Edad
    - Dirección
    - Número de integrantes en la familia
    - Estado de convivencia de los padres
    - Nivel de educación de la madre
    - Nivel de educación del padre
    - Trabajo de la madre
    - Trabajo del padre
    - Razón para estudiar en esta escuela
    - Tutor del estudiante
    - Tiempo de viaje hacia la escuela
    - Número de clases perdidas en el pasado
    - Soporte educacional por parte de la escuela
    - Suporte educacional por parte de la familia
    - Clases extracurriculares relacionadas con el curso
    - Actividades extracurriculares
    - Estuvo en guardería
    - Desea adquirir educación superior
    - Acceso a internet
    - Está en una relación romántica
    - calidad de la relación con la familia
    - Tiempo libre después de la escuela
    - Frecuencia de salidas con amigos
    - Consumo de alcohol en días laborales
    - Consumo de alcohol en días no laborales
    - Estado de salud
    - Número de ausencias en la escuela
    - Nota del primer corte
    - Nota del segundo corte
    - Nota del tercer corte

    Para más información acerca de estos campos remítase al siguiente link: https://github.com/MarcoFidelVasquezRivera/portuguese-grade-classification/blob/master/data/attribute%20definitions%20(español).txt

Requerimientos no funcionales:

1. Realizar la clasificación utilizando un decision trees, el cual es un algoritmo de machine learning.
2. Realizar dos implementaciones del decision tree, una completamente propia y otra utilizando librerías.

BÚSQUEDA DE SOLUCIONES CREATIVAS

Posibles soluciones para cargar, guardar y filtrar los datos

Solución 1: Guardar los datos del CSV en una lista de estudiantes. Los datos serán separados en 2 categorías (CATEGORIC, NUMERIC). El csv será cargado por medio del método StreamReader.

Se utilizarán los headers que provienen del archivo CSV para organizar los datos y añadirlos al estudiante. Estos ayudarán al momento de la selección del filtrado de datos que el usuario utilice.

Solución 2: Guardar los datos del CSV en una LinkedList de estudiantes. Los datos serán separados en 2 categorías (NUMERIC, CHAIN). El csv será cargado por medio del paquete nuget Lumen csv reader.

Se utilizará el orden predefinido del CSV para organizar los datos y añadirlos a un objeto estudiante. El usuario tendrá la opción de filtrar por medio de los campos definidos dentro del objeto estudiante.

Solución 3: Guardar la primera fila del CSV en un Array de estudiantes. El csv será cargado utilizando su path absoluto en el ordenador y el método de System.Io.File.ReadAllLines().

Se separará cada línea en una matriz donde las filas son un estudiante y las columnas son la información de dicha este.

Posibles soluciones para clasificar a los estudiantes y realizar la implementación del decision tree.

Solución 1: Realizar tanto la implementación propia como la implementación utilizando librerías en c#.

Solución 2: Realizar la implementación propia utilizando c# y realizar otra implementación utilizando librerías de python y conectarlo al programa mediante un API rest.

Solución 3: Realizar tanto la implementación propia como la implementación utilizando librerías en java.

UNIVERSIDAD ICESI

TRANSICIÓN DE FORMULACIÓN DE IDEAS A DISEÑOS PRELIMINARES

Posibles soluciones para cargar, guardar y filtrar los datos

Descripción solución 1: Crear una Lista para el guardado de los estudiantes. Cargar los atributos de los estudiantes desde el CSV objetivo haciendo uso del método StreamReader para guardarlos en la lista. En un window form, usamos el objeto DataGridView para presentar a los estudiantes de manera ordenada en una tabla. El filtrado se realiza mediante el uso de los objetos TextBox y ComboBox, dependiendo de la categoría del atributo a filtrar. En caso de un atributo categórico, un comboBox con los posibles valores es usado. En caso de un atributo numérico, dos TextBox son usados para representar el rango mínimo y máximo del valor deseado.

Descripción solución 2: Crear un Array para el guardado de los estudiantes. Cargar los atributos de los estudiantes desde el CSV objetivo haciendo uso del paquete nuget Lumen csv reader. En un window form, usamos el objeto DataGridView para presentar a los estudiantes de manera ordenada en una tabla. El filtrado se realiza mediante el uso de los objetos TextBox, dependiendo de la categoría del atributo a filtrar. En caso de un atributo de cadena, un TextBox es usado para ingresar el valor a buscar. En caso de un atributo numérico, dos TextBox son usados para representar el rango mínimo y máximo del valor deseado.

Descripción solución 3: Crear un Array para el guardado de los estudiantes. Cargar los atributos de los estudiantes desde el CSV objetivo haciendo uso del método System.Io.File.ReadAllLines() con el path absoluto del CSV en el ordenador. En un window form, usamos el objeto DataGridView para presentar a los estudiantes de manera ordenada en una tabla. El filtrado se realiza mediante el uso de los objetos TextBox, ingresando el valor deseado de un atributo.

Posibles soluciones para clasificar a los estudiantes.

Solución 1: Implementar un algoritmo de árboles de decisión que permita clasificar a los estudiantes, otorgándoles una estimación de la nota que estos obtendrán en el examen final.

Descripción solución 1: Como nodos de decisión tenemos los campos que pertenecen a cada estudiante. Cada nodo tiene la siguiente información: una condición, medida de impureza, número de muestras, valor de cada clase, clase que se le asigna a las muestras que llegan al nodo. El árbol es capaz de predecir la nota final de un estudiante, en base a los nodos de decisión establecidos.

Solución 2: Implementar un algoritmo K-NN que permita clasificar a los estudiantes, otorgándoles una estimación de la nota que estos obtendrán en el examen final.

Descripción solución 2: Se toma una parte del dataset para entrenar el algoritmo, al agregar un nuevo elemento el algoritmo calcula la distancia de este a los k elementos más cercano, después se realiza una “votación” entre los k elementos para decidir su clasificación.

Solución 3: Implementar un algoritmo de validación cruzada que permita clasificar a los estudiantes, otorgándoles una estimación de la nota que estos obtendrán en el examen final.

Descripción solución 3: Se divide el dataset en k grupos de similar tamaño, se toman k-1 grupos para entrenar el algoritmo y se deja uno como validación, se realiza el proceso k veces cambiando el grupo de validación en cada iteración. El algoritmo arroja k estimaciones de error y se promedia para obtener la estimación final.

EVALUACIÓN Y SELECCIÓN DE LA POSIBLE SOLUCIÓN

solución	Conocimiento sobre el tema	Necesidad o valor para el cliente	Facilidad de desarrollo	Flexibilidad	Total	Aprobado/No aprobado
Posibles soluciones para cargar, guardar y filtrar los datos						
alternativa 1	4	5	4	4	17	Aprobado
alternativa 2	4	4	3	4	15	No Aprobado
alternativa 3	4	4	3	4	15	No Aprobado
Posibles soluciones para clasificar a los estudiates						
alternativa 1	3	5	4	4	16	Aprobado
alternativa 2	3	5	4	4	16	No Aprobado
alternativa 3	1	2	2	2	7	No Aprobado

Posibles soluciones para cargar, guardar y filtrar los datos:

tras evaluar las tres alternativas para la solución del problema se ha decidido utilizar la solución 1.

Se eligió esta alternativa debido a la flexibilidad que aporta la clase StreamReader de c#, además de que se tiene experiencia previa trabajando con esta clase. Además, esta también es la opción que más aportaba valor para el cliente ya que si bien todas eran bastante parecidas, las otras opciones utilizaban soluciones que no eran realmente necesarias, tales como el nuget Lumen csv reader o el método System.Io.File.ReadAllLines(), las cuales consideramos demasiado potentes para este caso.

Posibles soluciones para clasificar a los estudiantes:

tras evaluar las tres alternativas para la solución del problema se ha decidido utilizar la solución 1.

La razón por la que se escogió esta opción fue que si bien se tiene un poco de conocimiento previo tanto del decision tree tanto del KNN, se ha decidido utilizar el primero principalmente por razones académicas ya que se desea aprender cómo es el funcionamiento de este algoritmo y poder implementar uno.