Valledupar 24 de septiembre del 2018

Ingeniero
Braulio Barrios
Miembros Comité de Proyecto de Grado

Cordial saludo

la presente es con el fin de entregar un informe con relación al proyecto de grado **“ MODELO DE MINERÍA DE DATOS PARA LA IDENTIFICACIÓN DE FACTORES ASOCIADOS A LA DESERCIÓN ESCOLAR EN UNA INSTITUCIÓN EDUCATIVA OFICIAL**” que fui asignado como jurado que a continuación enumero.

El documento fue revisado a atreves el **SOFTWARE TURNITIN,** es un servicio de prevención de plagio en Internet resalta las similitudes encontradas en el documento, reportando un 33% donde, no debería ser mayor del 10%. Al revisar se observa, que hay muchos párrafos parafraseados, es decir le cambiaron palabras, sugiero que estos párrafos deberían redactarse mejor nuevamente.

Con relación a las estructura del Documento en cada capitulo

1. SECCIÓN DEL MARCO TEÓRICO

- En la sección 3.1.1 los autores resalta la deserción escolar enmarcado en un enfoque sociológico y psicológico debe fundamentarse en el caso de estudio.

- En la sección 3.2.2 el autor describe el concepto de Minería de Datos y define algunas técnicas que no fueron analizadas en la Construcción del modelo y a demás no se evidencia en el marco teórico las medidas de evaluación del modelo.

- En la sección 3.1.2.3 los autores explican la Minería de Datos en la Educación hacen referencia a la aplicación de unos métodos o metodología y definen unas técnicas de EDM ¿Qué diferencia hay entre Minería de Datos y Minería de Datos para la Educación?

- En la sección 3.2 se le sugiere a los autores presentar un cuadro donde se relaciones las tareas y técnicas de minería de datos utilizadas para analizar la deserción escolar.

2. SECCIÓN DESARROLLO DE LAS FASES DE LA METODOLOGIA

La metodología planteada por los autores CRISP-DM consiste en desarrollar unas Fases con unas Tareas asociadas a la realización de unas actividades, se observa en el documento que algunas tareas las relaciona con otras fase se evidencia que no siguen con la estructura de la metodología por ejemplo página 89. La Selección de los Datos no hace referencia a la Fase de Comprensión de los datos, Revisar el orden de la metodología.

Se le recomienda a los autores realizar una Introducción en cada Fase explicando un resumen sobre las tareas y actividades a realizar y finalmente terminar con las conclusiones realizadas en la fase.

FASE COMPRENSIÓN DE LOS DATOS

1. Recolección de los Datos Iniciales

el autor debe abordar las siguientes actividades de la metodología :

- ¿qué atributos (columnas) de la base de datos parecen más prometedores?
- ¿qué atributos no parecen relevantes y se pueden excluir?
- ¿existen datos suficientes para obtener conclusiones generales o realizar predicciones precisas?
- ¿dispone de atributos suficientes para su método de modelado?
- ¿está fusionando varios orígenes de datos? en caso afirmativo, ¿existen áreas que puedan plantear problemas al fusionar?
- ¿ha considerado cómo se gestionan los valores perdidos en cada origen de datos?

2. Descripción de los datos

En la descripción de los datos se debe describir la cantidad de los datos, cuál es el formato de los datos, identifique el método utilizado para capturar los datos, qué dimensiones tiene la base de datos, y los tipos de datos y esquema de codificación. ademas un resumen estadístico delos datos numéricos (medidas de tendencias, de asociación y correlacion)

3. Exploracion de los datos

se debe utilizar herramientas de visualización para formular hipótesis o preguntas de investigación acerca de cómo los datos pueden adaptarse a los objetivos planteados. Estos análisis pueden ayudarle a describir los objetivos de minería de datos. Se recominda a los autores realizar las siguientes actividades :

- Utilizar gráficos de histograma de densidad, Tabla de frecuencia y bloxplot para comparar grupos de variables numéricas

- Características de centralidad y dispersión de los Datos
- Las distribuciones de los datos por grupo
- Las relaciones entre variables numéricas (correlacion)

Para comparar Dos grupos de variables se recomienda los siguientes pasos

1. Formular las dos una de hipótesis Una de investigación (lo que estoy buscado) y la otra hipótesis nula (H1 y H0)
2. Establecer el nivel de significancia o el margen de error 5% Eligir una prueba o test estadístico (t-test) para datos Numéricos y Categóricos (chi cuadrado con corrección de yates) , para correlaciones entre dos variables realizar el test de correlación de pearson
3. Calcula el P-Valor es una medida de incertidumbre entre más pequeño menos incertidumbre tengo por lo tanto se toma la hipótesis de investigación

4. Verificación de calidad de los datos

En esta tarea el autor solo relaciona en un tabla donde muestra relaciona los nombres de los atributos y valores faltantes y no identificados. No especifican las técnicas utilizadas para verificar la calidad de los datos, los autores expresan en el documento que le dataset presenta datos vacíos se debe que los estudiantes son retirados de la institución y eliminados del Simat por lo que se pierden los datos socioeconómicos surge el siguiente interrogante como se recuperan los datos socioecoomicos perdidos.

Se le recomienda a los autores responder los siguientes interrogantes.

- ¿ha identificado atributos perdidos y campos vacíos? si es así, ¿los valores perdidos tienen significado?
- ¿ha detectado desviaciones para determinar si son “ruido” o fenómenos que merecen un análisis en profundidad?
- ¿ha considerado excluir los datos que no tengan ninguna influencia en sus hipótesis?

FASE PREPARACIÓN DE LOS DATOS

1. Selección de datos

Los autores expresan en el libro que seleccionaron o elgieron los mejores atributos para ser utilizados en el modelo, por lo tanto no justifican por que algunos tributos no se tienen en cuenta para la construcción del modelo; de 50 atributos solo se tomaron 19 atributos se recomienda revisar la tarea de exploración de datos con el fin de verificar que relación tiene estos atributos que no fueron tenidos en cuenta para el modelamiento.

Se le recomienda a los autores realizar una tabla que relacione la cantidad de registros y atributos iniciales y los seleccionados para el modelo.

2. Limpieza de datos

En la limpieza de los datos los autores explican que se encontraron 29 registros sin datos académicos, no explica si estos registro fueron eliminados del dataset o que métodos utilizo para eliminar el ruido?, ademas se observa que los atributos no seleccionados no presentaron valores faltantes o ruido por lo tanto por que no fueron utilizdos en el moelamiento.

FASE DE MODELADO

Modelamiento Descriptivo de los Datos

Para esta fase se le recomienda al autor realizar ls siguientes tareas.

1. Selección de la técnica del modelo descriptivo

Se observa en el documento que el autor rotula como descripción de la técnica y explica la técnica seleccionada para el modelamiento, pero no relaciona que medida de similitud va utilizar para determinar la distancia entre grupos como se referencia en las base teóricas (técnicas descriptivas (clustering) La figura 7 Esta figura no se referencia en el contexto teórico del libro .

2. Construcción del modelo

En esta tarea el autor selección el numero de grupos donde se evidencia en la gráfica 2 utilizó el método del codo, que consiste en seleccionar el valor a partir del cual las variaciones no disminuyen de manera significativa, se evidencia el valor de k con 6 grupos, por lo tanto el autor expresa que el algoritmo genero un modelo con 20 clusters revisar.

3. Análisis del modelo

Revisando el documento se evidencia un análisis con respecto al desempeño académico de los estudiantes en el titulo 6.8.3 le falta adicionar la información socioeeconomica. En el análisis de la representación gráfica de la distribución de los valores de los diferentes atributos para cada uno de los grupos, se le recomienda al autor visualizar la distribución de las variables socioeconómicas por Clusters.

4. Evaluacion del modelo

- Se le recomienda a los autores examinar si hubo alguna relación entre los clusters con la perdida del estado academico del estudiante.
- El autor expresa que la evaluación del modelo fue realizado por error cuadrático pero no evidencia los resultados
- Presentar un resumen o conclusiones sobre los resultados obtenido

Modelamiento Predictivo de los Datos

Corregir la redacción en el párrafo en la sección 6.9

1. Selección de la técnica de modelado

En estas sección los autores mencionan tres métodos arboles de deciccion, naive bayes y bayes net; en el estado del arte los autores relaccionan otras técnicas aplicado en la minería de datos para la educación por que no fueron tomadas en cuenta para este caso de estudio en el documento no se evidencia la justificación.

La medida de desempeño del modelo el autor presenta las métricas obtenidas en la fase de experimentación con cada uno de los algoritmos no se evidencias en el documento en la sección del marco teorico las medidas de desempeño para la validación y diseño experimental.

Los autores no presentan la forma cómo se divide el conjunto de datos disponibles, es decir el conjunto de los datos de prueba y el segundo conjunto los datos de entrenamiento y validación.

Por lo tanto es necesario una reunión con los autores para revisar los resultados obtenidos y los observaciones realizadas .

Saludos

ALVARO OÑATE BOWEN