Recibido: julio 2017 Aceptado: agosto 2018

Factores que intervienen en el rendimiento académico en la Universidad

Factors that affect Academic Performance at the University

Wendoline Arteaga Sabja & Juan Pablo Sandoval

Universidad Católica Boliviana, Cochabamba, Bolivia

arteaga@ucbcba.edu.bo

Resumen: La minería de datos es ampliamente utilizada en el área de negocios, industrial o de servicio al consumidor. En este estudio se pretende darle una aplicación menos comercial y un poco más académica, que apoye en la toma de decisiones a los involucrados en el proceso de enseñanza-aprendizaje en la Universidad.

El objetivo de este estudio es identificar factores que afectan el rendimiento académico de los estudiantes, mediante técnicas del aprendizaje supervisado utilizando árboles de decisión, para lograrlo se analizan los datos de las materias cursadas desde el año 2012 al año 2015 en pre grado de la Universidad Católica Boliviana, regional Cochabamba.

Los resultados muestran que los factores que más afectaron el rendimiento académico fueron: la inscripción temprana, el mayor espacio libre en aula, repetir las materias, la hora de inicio de clases, el número de alumnos inscritos, la edad del estudiante y la experiencia del docente.

Palabras clave: Minería de datos, rendimiento académico, educación superior

Abstract: Data mining is widely used in business, industrial or consumer service areas. This study uses a data mining technique in academic scenarios, in order to support in decision-making to whom are involved in the teaching-learning process at the university.

The goal of this study is to identify factors that affect the academic performance of students, using supervised learning techniques with decision trees. For this purpose, this study analyzes the undergraduate student records from 2012 to 2015 of the Bolivian Catholic University, regional Cochabamba.

The study shows that the factors that most affect students' performance are: early registration, the largest free space in the classroom, repeating the subjects, the start time of classes, the number of students enrolled, the age of the student and the experience of the teacher.

Key words: Data mining, student performance, higher education

1 Introducción

Ciertamente no es fácil analizar todos los factores que intervienen en el rendimiento académico de los estudiantes en la Universidad. Algunos son bastante evidentes, como el nivel académico al ingresar a una carrera y el ambiente universitario, mientras que otros resultan más personales como la vocación por la profesión elegida, la motivación y la responsabilidad.

Si bien hay varios estudios interesantes al respecto, como el de K. P. Shaleena et al. 2015 y Abeer Badr et al. 2014, que predicen el rendimiento académico mediante el uso de técnicas de clasificación y minería de datos, en el presente trabajo más que predecir, se identificaron algunos factores que intervienen en la aprobación de los estudiantes en las carreras de pregrado en la Universidad Católica Boliviana San Pablo, regional Cochabamba.

1.1 Aprendizaje Supervisado

En este estudio se utilizaron técnicas del aprendizaje supervisado, que se refiere al hecho de dotar al algoritmo de un conjunto de datos de entrenamiento donde se dan las "respuestas correctas" y se espera que el algoritmo con nuevos datos encuentre una "respuesta adecuada", teniendo la idea de que hay una relación entre la entrada y la salida, de acuerdo a STANFORD, 2016.

En este trabajo se aplicaron árboles de decisión para descubrir la información que almacenan los datos.

1.2 Árboles de decisión

Los árboles de decisión son estructuras de datos jerárquicas que permiten obtener de forma visual las reglas de decisión bajo las cuales responden los estudiantes, a partir de datos históricos almacenados.

Un árbol de decisiones está compuesto por nodos de decisión. Dada una entrada, en cada nodo, se aplica una prueba y se toma una de las ramas dependiendo del resultado, este proceso empieza en la raíz y se repite recursivamente hasta llegar a un nodo hoja, en este punto el valor escrito en el nodo hoja constituye la salida, según explica Ethem Alpaydın, 2010. Además aclara que la estructura del árbol no se fija a priori, pero el árbol crece, las ramas y las hojas se agregan durante el aprendizaje dependiendo de la complejidad del problema inherente a los datos.

En este estudio se utilizaron árboles CHAID (del inglés *Chi-squared Automatic Interaction Detection*) ya que "producen resultados intuitivamente atractivos y estadísticamente válidos" de acuerdo al trabajo de Baron y Phillips, 1994. El algoritmo CHAID analiza todos los valores de cada variable predictora potencial a través del Chi-cuadrado, el cual refleja cuan relacionadas están las variables. A partir

de aquí, según Sanz Arazuri *et al.* 2010 se selecciona el predictor más significativo para formar la primera partición en el árbol de decisión.

De acuerdo al sitio de IBM Knowledge Center, 2017, si el predictor cuenta con más de dos categorías, se compararán estas categorías y se contraerán las que no presenten diferencias en los resultados. Este proceso de fusión de categorías se detiene cuando todas las categorías restantes difieren entre sí en el nivel de comprobación especificado.

La técnica de segmentación jerárquica utilizada por el algoritmo CHAID, según Magidson 1994, permite establecer una relación jerárquica de las variables explicativas en función de su nivel de significación a la hora de explicar la variable dependiente.

2 Metodología

El proceso que dirige este proyecto es conocido como KDD (*Knowledge Discovery in Databases*) y se refiere el proceso de descubrir de conocimiento de bases de datos, se lo usa en trabajos de investigación como el de Brijesh *et al*, 2011, que analizan el rendimiento académico usando técnicas de minería de datos. A continuación, se detallan cada una de las fases del proceso.

2.1 Selección de datos

Los datos se extrajeron mediante consultas a la base de datos del Sistema de Información Académico Administrativo, SIAA. De más de quinientas tablas que contiene el SIAA, se seleccionaron 15 tablas, que fueron relacionadas para mantener la integridad en los datos. Durante los años 2012 al 2015 se cursaron 8 semestres regulares, además de algunas materias dictadas en invierno y verano, se recopilaron, integraron y depuraron más de 68 mil registros en ese periodo. Cada registro representa la calificación de un estudiante de una carrera en una materia y un semestre dado.

En la Tabla 1 se presentan las variables que se seleccionaron para el estudio, luego de varias iteraciones en el proceso de extracción e integración.

Tabla 1. Lista de variables recopiladas para el estudio

Nombre	Categoría	Descripción
Facultad	Nominal	Se agruparon las carreras en EXA, HUM y ECO
Carrera	Nominal	Nombre de la carrera
Semestre	Nominal	Semestre en que se cursa la materia
Sigla	Nominal	Sigla de la materia
Doc_Sexo	Nominal	Género del docente
Doc_Estado_Civil	Nominal	Estado civil del docente
Doc_Edad	Discreta	Edad del docente
ExpEnMat	Discreta	Número de veces que un docente ha dictado una materia en el pasado
Período	Nominal	Primer semestre, segundo semestre, invierno o verano
Num_Alumnos_Inscritos	Discreta	Cantidad de alumnos inscritos
Espacio	Discreta	Espacio libre en la materia, diferencia entre el cupo y el número de alumnos inscritos
Est_Sexo	Nominal	Género del estudiante
Est_Estado_Civil	Nominal	Estado civil del estudiante
Est_Edad	Discreta	Edad del estudiante
Est_Nacionalidad	Nominal	Nacionalidad del estudiante
Inscripción	Discreta	Diferencia entre fecha de inscripción e inicio de clases, expresada en días
Turno_Colegio	Nominal	Diurno, vespertino o nocturno
Tipo	Nominal	Privado, estatal o CEMA (Centro de Educación Media Acelerada)
Localidad_Colegio	Nominal	Ciudad donde se encuentra el colegio
Cantidad_Materias	Discreta	Cantidad de materias que el estudiante toma en el semestre
Hora_Inicio	Discreta	Hora de inicio de la materia
Repetición	Discreta	Indica si el alumno está repitiendo la materia, por segunda o más ocasiones.
Continua	Continua	Nota de evaluación continua
Final1	Continua	Nota de primera sesión de examen final
Final2	Continua	Nota de segunda sesión de examen final
Total	Continua	Nota resultado la suma de la nota continua y la nota del examen final correspondiente
Aprobación	Nominal	Variable dependiente : S o N

2.2 Preparación de datos

En esta etapa se utilizaron diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, Han y Kamber, 2001.

Se verificó que las variables tanto dependientes como independientes no contengan valores nulos o fuera de rango y se obtuvieron más de 57 mil registros de calificaciones. Cuando los datos encontrados fueron erróneos se decidió excluirlos como sugiere IBM, 2017, para incluir en el análisis solo datos correctos.

2.3 Transformación

Una de las primeras variables generadas fue la nota total que es una variable dependiente, que está en el rango de 0 a 100 puntos, es un dato métrico resultado de sumar la evaluación continua que está valorada sobre 50 puntos, con la calificación del primer o segundo examen final también valorado sobre 50 puntos cada uno. Hay que tomar en cuenta que el estudiante se puede presentar a la primera sesión o a la segunda sesión, y en caso de ser necesario a ambas sesiones.

Por otro lado, se generó la variable dependiente Aprobación que es categórica, que tiene uno de dos posibles valores [s/n] que representa que el estudiante si aprobó la materia con una nota total de más de 50 puntos o que no aprobó la materia con una nota total menor a 51 puntos, este dato se calculó en base a la nota total.

Otras variables que se generaron mediante consultas SQL a partir de datos ya existentes son: el espacio en aula, repetición de materias, cantidad de materias que se cursan en el semestre, inscripción, edad del estudiante, edad y experiencia del docente en la materia.

2.4 Descripción de datos

A continuación, se listan las carreras que son parte del estudio, junto a los registros de 4 años de notas de estudiantes y el porcentaje de registros de aprobación con una nota total mayor a 50 puntos.

Tabla 2. Registros de notas de estudiantes agrupados por carrera de materias impartidas desde el 2012 al 2015

Código	Carrera	Registros de notas	Registros con nota aprobación	Porcentaje registros con aprobación
INQ	Ing. Química	698	377	54%
ING	Ing. de Sistemas	3 329	1 898	57%
CPU	Contaduría Pública	5 751	3 566	62%
INC	Ing. Civil	3 113	1 992	64%
DER	Derecho	5 299	3 762	71%
ITL	Ing. de Telecomunicaciones	1 387	985	71%
ADM	Administración de Empresas	16 869	12 146	72%
COM	Comunicación Social	7 312	5 411	74%
ICO	Ing. Comercial	4 771	3 531	74%
IMA	Ing. Medio Ambiental	2 199	1 671	76%
IND	Ing. Industrial	1 999	1 559	78%
FYL	Filosofía y Letras	1 053	821	78%
PSI	Psicología	2 454	1 914	78%
IFI	Ing. Financiera	1 420	1 179	83%
IME	Ing. Mecatrónica	214	193	90%

Las carreras de Ingeniería Química, Ingeniería de Sistemas y Contaduría Pública son las que tienen registros con el menor porcentaje de aprobación. Cabe mencionar que Ingeniería Mecatrónica es una carrera nueva que tiene registros desde el año 2013.

El tipo de colegio del que viene el alumno determina su procedencia. Los estudiantes que provienen de instituciones privadas tienen mejores porcentajes de aprobación con 72%, los que proceden de colegios estatales alcanzan un 63% y si provienen del CEMA, Centro de Educación Media Acelerada, logran solo el 38% de aprobación.

Se tiene el registro de notas de 29 314 varones y de 28 554 mujeres. Los porcentajes de aprobación varían de acuerdo al género de los estudiantes ya que el 66% de varones aprobaron sus asignaturas, frente al 74% de mujeres.

En la siguiente figura se aprecia que el porcentaje de aprobación más alto se da cuando el estudiante ha tomado 6 materias en el semestre, con un 80%.

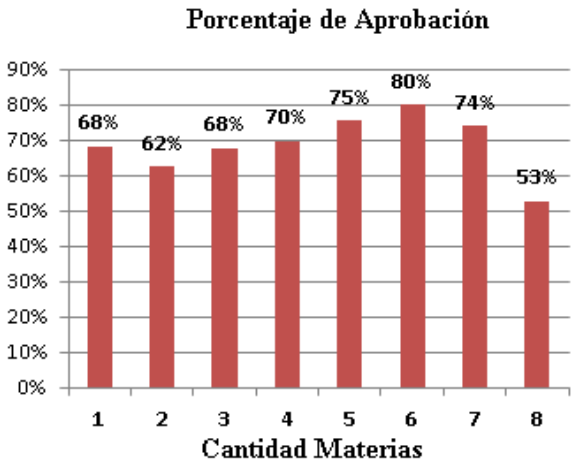


Figura 1: Aprobación por cantidad de materias tomadas en un semestre

Al visualizar los porcentajes de aprobación de acuerdo al período académico en el que se dicta la materia, se aprecia un incremento en la aprobación en las materias que se dictan en invierno y verano.

Tabla 3. Aprobación por período académico en que se cursa una materia

Período académico	Aprobación
Primer semestre	69%
Invierno	82%
Segundo semestre	72%
Verano	80%

Los datos muestran que el 99% de los estudiantes son de nacionalidad boliviana y casi el 100% de los estudiantes son solteros.

2.5 Minería de datos

Una vez que se seleccionaron, integraron, prepararon y depuraron los datos, empieza el trabajo de minería de datos, según Cesar Pérez Lopez, 2007, "La minería de datos es un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos".

En este estudio se aplicaron árboles de decisión CHAID, para explicar la variable objetivo Aprobación que tiene una salida binaria [s/n] que representa si un estudiante aprueba o no una materia. Luego de aplicar el algoritmo CHAID a los datos de entrenamiento equilibrados, obtenidos de la mitad de la muestra aleatoriamente, se

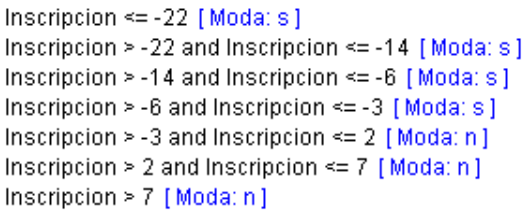
genera un árbol que permite establecer una relación jerárquica de las variables predictoras que explican la variable dependiente Aprobación.

Finalmente se aplica el árbol generado en la otra mitad de la muestra con los datos de prueba para verificar la precisión del árbol.

3 Resultados

El árbol de decisión CHAID encuentra que la primera partición es la inscripción, que resulta ser el predictor más significativo. La inscripción es la diferencia entre la fecha de matriculación y la fecha de inicio del semestre, representada en días.

El algoritmo CHAID agrupó en siete categorías la variable Inscripción. Se aprecia que la moda en este árbol es: s, significa que, SI aprobó la materia, si la inscripción se realiza 3 días antes del inicio del semestre. En este árbol se trabajó con un el nivel de significación para la fusión de categorías de 0,05.



En la Tabla 4 se muestra la cantidad de registros y los porcentajes de aprobación de los datos de entrenamiento en cada categoría.

Tabla 4. Cantidad de registros y porcentajes de aprobación por categoría

Inscripción	<= -22	<-22 y <= -14	>-14 y <= -6	>-6 y <= -3	>-3 y <=2	>2 y <=7	>7
Registros	2 294	4 892	4 950	2 181	3 425	1 474	1 247
Aprobación	64 %	58 %	54 %	50 %	44 %	37 %	33 %

Para conseguir un árbol más simple de interpretar, se cambió el nivel de significación para la fusión de categorías a 0,01. El resultado fue un árbol con menos nodos y con las categorías de las variables predictoras más fusionadas, en este caso la inscripción sigue siendo la primera partición del árbol y el predictor más significativo, ahora agrupada en solo dos categorías.

Inscripcion <= -3 [Moda: s]
Inscripcion > -3 [Moda: n]

En este segundo árbol la moda continúa siendo aprobación si la inscripción se realiza 3 días antes del inicio del semestre, con 56 % de aprobados. Si la inscripción se da después del tercer día, solo el 39 % de los registros de entrenamiento son de aprobación.

Al analizar las ramas del último árbol, se aprecia en la Figura 2: que, si la inscripción es menor o igual a -3 y el alumno está repitiendo la materia la moda es No aprobar. En el caso de que sea la primera vez que cursa la materia en las carreras de Contaduría Pública, Ingeniería Civil, Ingeniería de Sistemas e Ingeniería Química la moda es No aprobar, excepto cuando la materia inicia pasadas las 11 de la mañana y la edad del estudiante es menor o igual a 21 años. En caso que la materia inicie antes de las 11 de la mañana solo se aprueba si hay espacio libre mayor a 4.

Para el resto de las carreras la moda es aprobar si la inscripción es temprana, y es la primera vez que se cursa la materia.

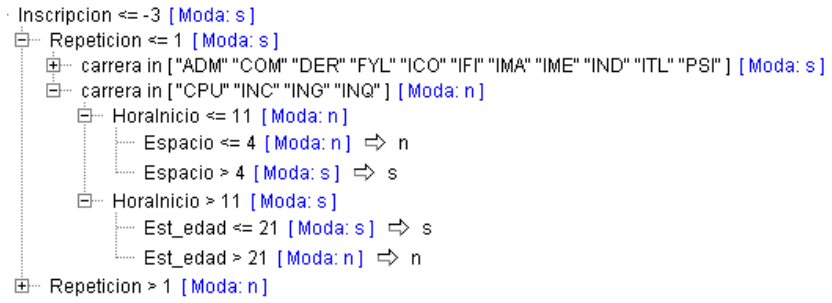


Figura 2: Sub-árbol con Inscripción menor o igual a -3

La Figura 3: muestra el segundo sub-árbol, donde se aprecia la relación jerárquica de las variables explicativas cuando la inscripción es mayor a -3.

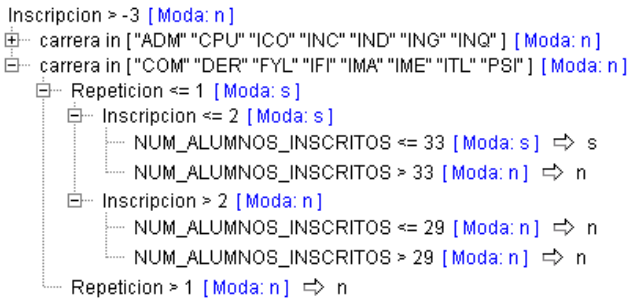


Figura 3: Sub-árbol con Inscripción mayor a -3

Si la inscripción es posterior a tres días antes del inicio del semestre, la moda es reprobar en todos los casos, salvo que se tengan menos o igual a 33 alumnos inscritos en la materia, no se haya repetido la materia y se esté en carreras como: Comunicación, Derecho, Filosofía, Ing. Financiera, Ambiental, Mecatrónica, Telecomunicaciones o Psicología.

4 Discusión

En este estudio se detectaron patrones en la matriculación, características de los estudiantes y docentes que afectan en el rendimiento académico, sin embargo no se cuenta con datos como el ambiente familiar o social, la vocación del estudiante y sus motivaciones, que son temas trascendentales a la hora de analizar el rendimiento académico, este es solo un primer paso con el objetivo de apoyar a estudiantes, docentes y administrativos a mejorar las decisiones que se tomen en torno al rendimiento académico en la Universidad.

5 Conclusiones

Los árboles de decisión son simples de interpretar, pero pueden dar resultados muy específicos si no se configura correctamente la división y fusión de categorías. Permiten visualizar las reglas de decisión, por lo tanto, favorecen el análisis de los datos ya que se pueden ver las condiciones bajo las cuales se realiza la selección.

En cuanto a los factores que intervienen en el rendimiento académico de los estudiantes se puede concluir que:

- En ambos árboles CHAID, independiente del nivel de fusión, la inscripción es la primera partición y por lo tanto el predictor más significativo. Si bien la inscripción temprana no garantiza por si sola la mejora del rendimiento académico, junto a un mayor espacio libre en aula, un horario de clases acorde a las necesidades del estudiante y no haber reprobado antes la materia son variables que afectan en el rendimiento académico de los estudiantes en la Universidad. Estos resultados coinciden con estudios como el de Corral, 2006.
- De acuerdo con OXFORD 2015, las horas de inicio tempranas para la escuela a menudo no concuerdan con los ritmos circadianos del cuerpo. Tomando en cuenta este y otros estudios no sorprende que el rendimiento de los estudiantes mejore cuando la hora de inicio de clases se da pasadas las 11 am., especialmente para estudiantes menores a 21 años, ya que según Troxel 2017, los horarios de inicio antes de las 8:30 de la mañana impiden a los adolescentes dormir durante el momento de sus vidas que más lo necesitan.

- Un número de inscritos menor también ayuda a mejorar las oportunidades de aprobar una materia. La investigación de Aitken 1982, encontró que la nota promedio en la universidad se incrementa si mejora la calidad del ambiente físico en el cual el estudiante realizaba su trabajo académico. Esto indica que cursos concentrados no ofrecen las condiciones ideales de estudio.
- Es interesante observar que la edad o género del docente no son significativos en los árboles estudiados, y que en los niveles inferiores del árbol a mayor experiencia docente en una materia se produce un incremento en la reprobación. Este resultado coincide con Martí 2012, lo que sugiere que los docentes con mayor experiencia son más exigentes.
- La edad de los estudiantes ideal para aprobar asignaturas en la Universidad es antes de los 22 años. Este resultado coincide con el estudio de Kotsiantis *et al*, 2005, donde indican que los estudiantes mayores tienden a reprobar materias en la Universidad. Esto puede deberse a que con el pasar de los años las personas deben empezar a trabajar o deciden formar familia, lo que los aleja de la concentración y motivación necesarias para culminar sus estudios.

Referencias Bibliográficas

[1] Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby (2014). *Data Mining: A prediction for Student's Performance Using Classification Method,* World Journal of Computer Application and Technology. Vol. 2(2), pp. 43 - 47

[2] Aitken M. (1982) *A personality profile of the college student procrastinator.* Doctoral dissertation, University of Pittsburgh.

[3] Barón, S. y Phillips, D. (1994). *Attitude Survey Data Reduction Using CHAID: An Example in Shopping Centre Market Research.* En Hooley, G. J. y Hussey M. K., Quantitative Methods in Marketing (75-88). Londres: Academic Press.

[4] Brijesh Kumar, Saurabh (2011). *Mining educational data to analyze students' performance,* IJACSA Vol. 2, No. 6.

[5] Cesar Pérez Lopez, (2007). *Minería de datos. Técnicas y herramientas.* Editorial Paraninfo, España, Madrid.

[6] Corral Verdugo, Xochitl Díaz Núñez (2006). *Factores asociados a la reprobación de los estudiantes de la universidad de Sonora*, X Congreso nacional de investigación educativa, México.

[7] Ethem Alpaydın (2010). *Introduction to Machine Learning*, Segunda Edición, The MIT Press Cambridge, Massachusetts

[8] Han , Kamber (2006). *Data Mining: concepts and techniques, 2nd edition*

[9] IBM Knowledge center (2017). . Recuperado el 20 de noviembre de 2017 de: https://www.ibm.com/support/knowledgecenter

[10] K. P. Shaleena , Shaiju Paul (2015). *Data mining techniques for predicting student performance,* Engineering and Technology (ICETECH), IEEE.

[11] Kotsiantis, Pintelas, P.E (2005). *Predicting students marks in Hellenic Open University, ICALT 2005.* Fifth IEEE international conference on advanced learning technologies, Kaohsiung, pp.

[12] Magidson and Vermunt (2005*). An Extension of the CHAID Tree-based Segmentation Algorithm to Multiple Dependent Variables*. Statistical Innovations Inc., USA and Department of Methodology and Statistics, Tilburg University,

[13] Magidson, J. (1994). *The CHAID Approach to Segmentation Modeling: Chi-square Automatic Interaction Detection.* En Bagozzi, R. P., (Ed.) Advanced Methods of Marketing Research (pp. 118-159). Oxford: Blackwell

[14] Martí Ballester (2012). *¿Influyen las características del profesor en el rendimiento académico del estudiante?* Departamento de Economía de la Empresa Universitat Autònoma de Barcelona

[15] OXFORD University (2015). *Wake-up call over sleep and public health needed performance.*

[16] Sanz Arazuri, Eva, Ponce de León Elizondo, Ana (2010). *Claves en la aplicación del algoritmo Chaid. Un estudio del ocio físico deportivo universitario*. Revista de Psicología del Deporte, Universitat de les Illes Balears, España

[17] STANFORD, Andrew Yan-Tak (2016), *Machine Learning Course*. Recuperado el 12 de octubre de 2016 de: *https://es.coursera.org/instructor/andrewng*

[18] Troxel (2017). *Senior Behavioral and Social Scientist at RAND and Adjunct Professor of Psychiatry and Psychology at the University of Pittsburgh*. Recuperado el 7 de octubre de 2017 de: *https://www.ted.com/talks/wendy_troxel_why_school_should_start_later_for_teens#t-508869*