Implementación de Procesos KDD (Knowledge Discovery in Databases) para la caracterización de la calidad de leche bovina en función de la composición e higiene sanitaria

---

Año
2018

Directores del proyecto
Cabrera, Gabriela Pilar

Equipo de investigación
Marco, Laura Angélica; Asinari, Marianela; Cornejo, Anabella Patricia; Pavón, Sandra Marisel y Defagot, María Verónica

Alumnos participantes
Perrotta, Camila Dahyana y Vesprini, Melina Viviana

Universidad Nacional Villa María | Instituto de Investigación

INFORME ACADÉMICO FINAL

Proyectos de Investigación 2016-2017

PROYECTO:

Implementación de Procesos KDD *(Knowledge Discovery in Databases)* para la caracterización de la calidad de leche bovina en función de la composición e higiene sanitaria.

DIRECTOR:

Cabrera, Gabriela Pilar

CO-DIRECTOR:

No corresponde

EQUIPO DE INVESTIGACIÓN:

Marco, Laura Angélica. Assinari Marianela. Cornejo Anabella Patricia. Pavón Sandra Marisel. Defagot María Verónica.

ALUMNOS INTEGRANTES:

Perrotta, Camila Dahyana. Vesprini Melina Viviana.

1. INFORME ACADÉMICO DEL PROGRAMA/PROYECTO

En el presente informe se sintetizan las actividades y principales resultados obtenidos durante el desarrollo de la investigación. Se implementaron todas las etapas del proceso Knowledge Discovery in Databases (KDD) para la caracterización de la calidad de leche bovina cruda -en aspectos Higiénicos–Sanitarios, Composicional y Adulteración- y sus variaciones estacionales; registradas durante el año 2014 de cinco empresas lácteas del departamento Rio II, Provincia de Córdoba.

El KDD es un proceso interactivo y centrado en el usuario. Implica la preparación, selección y limpieza de datos para la generación de la Data Warehouse -repositorio de datos depurados y estandarizados-. La exploración, limpieza y transformación de datos. La aplicación de la Minería de datos a partir de herramientas de estadística descriptiva e

inferencial, para la identificación de relaciones y patrones en los datos y la selección y evaluación de los modelos obtenidos.
En este contexto, se pretendió la obtención del Data Warehouse y la evaluación de la calidad de los datos generados; a través de un análisis exploratorio mediante la aplicación del software estadístico Infostat y entorno y lenguaje de programación R. Se realizó además un análisis comparativo respecto de la eficiencia de los procedimientos implicados y amigabilidad del formato de presentación de ambos software. Se descartó el software estadístico IBM SPSS, debido al alto costo para la actualización del mismo, como también de la incorporación de diferentes módulos para complejizar los procesamientos que fueran necesarios. En síntesis, se trabajó con la combinación de InfoStat y R.
Ahora bien, en la etapa de limpieza e integración de los datos se trabajó directamente con los archivos "txt" provistos por el laboratorio que realizó en análisis clínico de las muestras de leche. Cada archivo txt, contenía información acerca de catorce variables que correspondían a las muestras de un determinado día para un determinado tambo. Cabe aclarar que, son cuatro registraciones mensuales por tambo (se tienen cinco industrias codificadas con A, B, C, D y E) durante los 12 meses del año 2014. El formato de archivo "txt" implicó la carga y transducción de cada uno de los archivos a una hoja de cálculo Excel en el que se integró toda la información. Esta actividad fue llevada a cabo, por las estudiantes participantes del proyecto con la ayuda de los estudiantes de 4to.año de Medicina Veterinaria, en el marco de la asignatura Metodología de la Investigación y Evaluación. La transducción de los datos, supuso por ejemplo, la conversión de la variable fecha en variable momento de muestreo. Esto es, algunas de las variables cuantitativas se categorizaron y otras de las variables como la fecha, se reconvirtieron.
A partir de esta primera versión de la Data warehouse, se aplicaron procedimientos de análisis exploratorio con InfoStat. Este análisis exploratorio implicó: obtención de medidas resumen y gráficos de caja, en primera instancia. Esto arrojó, la presencia de errores en la carga de datos, tanto de los estudiantes como de los usuarios de los tambos al tomar las muestras. Esto último fue señalado como algo relevante por el experto en Calidad de Leche consultado. Ahora bien, debido a que el proceso de limpieza y depuración de la Data Warehouse requirió varias iteraciones, se procedió a diseñar rutinas mediante el Intérprete R disponible en InfoStat. En paticular, las rutinas específicas para la obtención de las medidas resumen y los gráficos de caja, éstos con la delimitaciones de los valores mínimos y máximos indicados en la entrevista con el experto en calidad de leche. Estas rutinas fueron construidas en el marco de un Taller de R, dictado por una experta, en la que los integrantes del equipo de investigación se familiarizaron con este entorno de programación como también con conceptos y procedimientos estadísticos. La decisión de utilizar InfoStat como intérprete de R, facilitó al equipo el acceso al lenguaje de programación. Este taller, se realizó en dos jornadas de 4hs cada uno y hubo consultas virtuales durante el proceso de

diseño e implementación de otras rutinas que sugieron en la marcha. Cabe aclarar que, la última versión de la Data Warehouse consta de 1976 registros y 14 variables de análisis. Uno de las conclusiones a las que llegó el equipo de investigación, es que el InfoStat como interfaz del entorno R, acerca a los usuarios a la potencia del lenguaje de programación y genera un ambiente de mayor amigabilidad en el que sea hace más accesible dicho lenguaje.

Por otra parte, la posibilidad que brinda R de almacenar los algoritmos o rutinas diseñados, y por tanto, de revisarlos, corregirlos y ejecutarlos las veces que sea necesario; hace menos engorroso, más rápido y ágil el proceso de limpieza y depuración de la Data Warehouse. En cambio, si sólo se aplicaran las herramientas disponibles en InfoStat sin hacer uso del intérprete R, sería necesario realizar todos los pasos cada vez que se obtiene una nueva versión de la base de datos.

Cabe mencionar que, la posibilidad de iteración de las rutinas desarrolladas en R, también es válida al momento de la aplicación de otros procedimientos estadísticos para la obtención de relaciones y patrones en los datos. Ahora bien, cuando la base de datos está estable y los procedimientos a aplicar están disponibles en el software estadístico, es conveniente utilizar dichas aplicaciones. Esto es, una ventaja en cuanto al conocimiento experto que requiere la programación de procedimientos complejos que ya se encuentran disponibles en InfoStat.

En relación a los resultados obtenidos de la aplicación de rutinas desarrolladas en R y procedimientos disponibles en Infostat; que aportan a la extracción del conocimiento para la caracterización de la calidad de leche bovina en los cinco tambos analizados, se menciona que:

• Sólo un 47% de las 1970 muestras de leche analizadas cumplen con el parámetro de calidad en relación RCS (Recuento de células somáticas) menor o igual a 400. En tanto, en el tambo A (44%), B (60%), C (47%), D (53%) y E (56%).

• Sumado al prámetro RSC, se analizó el porcentaje de muestras que cumplían con un porcentaje de materia grasa de más de 3,4%, un porcentaje de proteína de más de 3,15% y índice crioscópico de menos de -0,512 °C. De acuerdo a esto, del 44% de las muestras que cumplen con las especificaciones para el RCS, sólo un 24% cumple con las otras tres especificaciones. En tanto, Tambo B (38% de 60%), Tambo C (37% de 47%), Tambo D (53% de 53%) y Tambo E (36% de 56%). Si se consideran dos de las tres especificaciones en el Tambo A (87% de 44%), Tambo B (87% de 60%), Tambo C (88% de 47%), Tambo D (96% de 53%) y Tambo E (95% de 56%).

• De la aplicación del coeficiente de Correlación de Spearman, surge que el RCS se correlaciona con el contenido de materia grasa y los sólidos totales.

• De la aplicación del coeficiente de Correlación de Pearson se deduce que el contenido de materia grasa y lactosa están correlacionados positivamente.

Cabe aclarar que, se está trabajando conjuntamente con expertos en Calidad de Leche, en una publicación con base en los anteriores resultados. Esta publicación será presentada en alguna de las siguientes revistas de interés del ámbito de la Medicina Veterinaria: International Journal of Applied Science and Technology, Food Science and Technology Journal of dairy science, Animal Science Journal o REDVET.
Sumado a esto, y como uno de los resultados devenidos de la investigación realizada, se propone para el año 2018 se incluyan en los programas de Bioestadística, Informática y Metodología de la Investigación y Evaluación; contenidos vinculados con los procesos KDD y el desarrolla de rutinas básicas de R con la Interfaz InfoStat.

2. VINCULACIÓN CIENTÍFICA

2.1. Describir vínculos generados desde el Programa/Proyecto con referencia a demandas del Sector Productivo.

En primer lugar, cabe destacar que el presente proyecto surge como respuesta a la demanda del LABORATORIO DE DIAGNOSTICO VETERINARIO VILLA MARIA, en relación a la necesidad de caracterizar la composición y calidad microbiológica de leche cruda de tambos bovinos. Cabe señalar, que estos tambos resultan pequeñas o medianas empresas y en este sentido interesa identificar si la calidad de leche se ve afectada por el tipo de emprendimiento.

2.2. Describir vínculos que respondan a demandas internas de distintas aéreas de la UNVM.

Se desarrollará en 2018, oferta de cursos de R Interfaz InfoStat para los docentes e investigadores interesados de la Carrera de Medicina Veterinaria. Esto surgió a partir de la experiencia de los talleres de R realizados en el marco del presente proyecto de investigación.

Sumado a esto, y en el marco del Programa de Intercambio PROMIDI, se realizó vinculación con Universidad Nacional de Colombia (Medellín), para la creación de Apps con simulaciones estocásticas educativas, con el paquete Shiny de R, dirigido por el Dr. Freddy Hernández.

Por otra parte, se incluirá una aproximación de los procesos KDD en las asignaturas Bioestadística, Informática que se dictan en el primer año de estudio de la Carrera de Medicina Veterinaria y en la asignatura Metodología de la Investigación y Evaluación.

Cabe señalar además que, los resultados de este proyecto están también vinculados con la realización de la Tesis de Maestría en Ingenieria en Sistemas de Información, de la Universidad Tecnológica Nacional, Facultad Regional Córdoba de la Ingeniera Laura Marco.

3. PUBLICACIÓN EN REPOSITORIO DIGITAL DE LA UNVM

AUTORIZO LA PUBLICACIÓN DE ESTE INFORME ACADÉMICO FINAL EN EL REPOSITORIO DIGITAL DE LA UNVM: **SI**