PEC4: Final assessment

Marco Russo - Silvia Gamundi Sumando

Enero, 2025

Contents

1 Información del Estudiante

Nombre	Marco Russo
Email	mrussorb@uoc.edu
GitHub	https://github.com/marcusRB/uoc-ub-scientific-programming
LinkedIn	https://www.linkedin.com/in/marcusrb/
Fecha	January 6, 2026

Nombre	Silvia Gamundi Sumando
Email	sgamundis@uoc.edu
GitHub	https://github.com/
LinkedIn	https://www.linkedin.com/in/
Fecha	January 6, 2026

2 Sección 1. Contexto y objetivo del estudio. Datos (1 punto)

El dataset elegido es `Bacteremia` Heinze, G. (2023). Bacteremia [Data set]. In PLoS One (Version S2, Vol. 9, Number 9, p. e106765). Zenodo. https://doi.org/10.5281/zenodo.7554815

El resto de informaciones extracto del sitio oficial:

> The data set consists of 14,691 observations from different patients with the clinical suspicion to suffer from bacteremia, for whom a a blood culture analysis was performed at the Vienna General Hospital, Austria, between January 2006 and December 2010. It contains the results of the blood culture analysis for bacteremia and the values of 51 potential predictors of bacteremia. To protect data privacy our version of this data was slightly modified compared to the original version, and this modified version was cleared by the Medical University of Vienna for public use (DC 2019-0054). Details on the meaning of the variables can be found in the data dictionary. The original version of the data set was used by Ratzinger et al (2014) to develop a model for screening bacteremic patients based on highly standardizable laboratory variables. This public version has been used by Gregorich et al (2021).

Basada en la descripción oficial del mismo, se indican que existen 14,691 observaciones de diferentes pacientes que podrían ser afectos de `bacteriemia`. De la información disponible en Wikipedia:Bacteriemia, la bacteriemia es la presencia de bacterias en la sangre. La sangre es normalmente un medio estéril, por lo tanto la detección de bacterias es indicativa de infección.

Es importante entender este punto respecto al **diagnóstico**, muchas personas se recuperan completamente de la bacteriemia. Sin embargo, la bacteriemia es grave y puede provocar sepsis. Cuando tiene sepsis, el daño a los órganos principales puede ser irreversible.

Entre las **causas**, la entrada de bacterias en el torrente sanguíneo puede ser producto de una infección localizada (ej: neumonía, absceso en piel o mucosas), o por interrupción de la piel como barrera defensiva. Se destacan las intervenciones quirúrgicas, utilización de dispositivos invasivos (catéteres, sondas, asistencia mecánica respiratoria), heridas accidentales, o quemaduras.

La infección suele empezar en los pulmones, el tracto genitourinario, gastrointestinal o los tejidos blandos, entre ellos la piel de pacientes con úlceras. También puede ser secundaria a una intervención dental en pacientes de alto riesgo, especialmente en los que tienen prótesis intravasculares.

Respecto a las **consecuencias**, dependen del tipo de bacteria y el estado del paciente. La respuesta inmunológica a la infección puede causar sepsis y devenir en shock séptico. También puede ocurrir que la sangre transporte las bacterias a otros tejidos, que podrán ser infectados. Ejemplos incluyen endocarditis, osteomielitis, y meningitis. El tratamiento es fundamental para erradicar a las bacterias y requiere el uso de antibióticos por vía intravenosa.

Tenemos un pequeño diccionario disponible del significado de cada característica del dataset, lo descargaremos y visualizaremos para entender mejor el contexto.

---

3 Sección 2. Prospección y preparación de los datos (2 puntos)

3.1 2.1 Descripción de los datos (1 punto)

ver LAB1

3.2 2.2 Preguntas "objetivo" (1 punto)

ver LAB3 y PEC1

---

4 Sección 3. Análisis exploratorio de los datos (2,5 puntos)

4.1 3.1 Análisis descriptivo y gráfico (1 punto)

ver LAB2

4.2 3.2. Ejercicios de inferencia y simulación (1,5 puntos)

ver LAB3, LAB4, PEC2

---

5 Sección 4. Modelos de aprendizaje automático (2,5 puntos)

ver LAB5 y PEC3

---

6 Sección 5. Visualización (1,5 puntos)

ver LAB6 y PEC3

---

7 Sección 6. Conclusiones (0,5 puntos)

---

8 Bibliografía

References

1. Wolberg,W.H., and Mangasarian,O.L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. In Proceedings of the National Academy of Sciences, 87, 9193-9196.

- Size of data set: only 369 instances (at that point in time)

- Collected classification results: 1 trial only
- Two pairs of parallel hyperplanes were found to be consistent with 50% of the data
- Accuracy on remaining 50% of dataset: 93.5%
- Three pairs of parallel hyperplanes were found to be consistent with 67% of data
- Accuracy on remaining 33% of dataset: 95.9%

2. Zhang,J. (1992). Selecting typical instances in instance-based learning. In Proceedings of the Ninth International Machine Learning Conference (pp. 470-479). Aberdeen, Scotland: Morgan Kaufmann.

- Size of data set: only 369 instances (at that point in time)
- Applied 4 instance-based learning algorithms
- Collected classification results averaged over 10 trials
- Best accuracy result:
- 1-nearest neighbor: 93.7%
- trained on 200 instances, tested on the other 169
- Also of interest:
- Using only typical instances: 92.2% (storing only 23.1 instances)
- trained on 200 instances, tested on the other 169

Blake, C.L. & Merz, C.J. (1998). UCI Repository of Machine Learning Databases. Irvine, CA: University of California, Irvine, Department of Information and Computer Science. Formerly available from http: //www.ics.uci.edu/~mlearn/MLRepository.html.

- https://shiny.posit.co/py/templates/
