

Magíster en Data Science

Curso: machine Learning

Pauta segunda evaluación

Instrucciones generales

1. El trabajo deberá ser desarrollado en parejas. Sólo en casos excepcionales se aceptarán grupos de máximo tres integrantes.
2. Se recomienda escoger temáticas sobre las cuales dispongan de datos relacionados con su contexto profesional (datos con los que se trabaje a diario) y/o académico (por ejemplo, tema de tesis). También pueden utilizar los datos empleados para la primer avealuacióin Al final de la pauta se expone un listado de datos abiertos para el sector público.
3. Se deben desarrollar **las dos secciones** que contempla la pauta. Si lo desea puede disponer de un conjunto de datos para resolver ambas secciones, siempre y cuando se respeten las restricciones de especificación correspondientes. Si también lo desea, puede escoger un conjunto de datos para la primera sección (clasificadores) y otro para el modelamiento mediante regresión lineal.
4. Entregables: reporte en formato .doc, .docx o .pdf con el desarrollo de cada sección. Un archivo .ipynb con el código completo de los procedimientos empleados para cada sección (es decir, son dos archivos con el código)
5. Plazo entrega: **Lunes 16 de Diciembre (por confirmar con Dirección)** por correo a la dirección [rodrigofernandez@udd.cl](mailto:rodrigofernandez@udd.cl) hasta las 23:59 hrs.

Pauta reporte

1. *DECISION TREES & RANDOM FOREST*
2. Escoja un problema que pueda ser abordado mediante el diseño, implementación y validación de algoritmos de aprendizaje supervisado de **clasificación**.
   1. Contextualice el problema en no más de una plana
   2. Formule una pregunta de valor que justifique el abordaje mediante algoritmos de Machine Learning
3. Descripción de los datos
   1. Características generales de las instancias y características de sus datos a incluir dentro de su análisis.
   2. Resumen de los datos a incluir dentro del análisis. Mediante la función .info() para resumen del tipo de datos utilizados y mediante la función .describe para describir su comportamiento univariado.
4. Elabore un gráfico de matriz entre los datos que utilizará en el análisis y describa asociaciones visuales que le parezcan relevantes.
5. Especificar, implementar, validar, comparar y decidir el algoritmo de clasificación. Los algoritmos a comparar son el Decision Tree y Random Forest.
   1. Implementación: sea claro en señalar los argumentos y parámetros utilizados en cada modelo.
   2. Validación mediante método hold out para al menos una partición (entrenamiento/validación) aleatoria. Obtención de las métricas de desempeño de cada clasificador (Accuracy, Precission, Recall y f1-score – con y sin ponderación -). En el caso de Decision Tree evalué “podar el árbol” si es necesario y vuelva a comparar resultados ¿Se reduce el overfitting?
   3. Comparación del desempeño de ambos algoritmos. Considere las restricciones de especificación para cada algoritmo en su respuesta.
   4. Conclusiones respecto del clasificador seleccionado.
6. *APRENDIZAJE NO SUPERVISADO*
7. Escoja un problema que pueda ser abordado mediante el diseño, implementación y validación de algoritmos de aprendizaje supervisado. Recuerde que no es necesario contar con una hipótesis que oriente este tipo de enfoques.
   1. Contextualice el problema en no más de una plana
   2. Formule una pregunta de valor que justifique el abordaje mediante algoritmos de Machine Learning. Recuerde que como se trata de algoritmos de aprendizaje no supervisado, puede considerar una pregunta lo suficientemente abierta para tratarla desde una perspectiva exploratoria
8. Descripción de los datos
   1. Características generales de las instancias y características de sus datos a incluir dentro de su análisis.
   2. Resumen de los datos a incluir dentro del análisis. Mediante la función .info() para resumen del tipo de datos utilizados y mediante la función .describe para describir su comportamiento univariado.
9. Elabore un gráfico de matriz entre los datos que utilizará en el análisis y describa asociaciones visuales que le parezcan relevantes.
10. Especificar, implementar, validar, comparar y decidir entre el algoritmo de cluster (kmeans) y análisis de componentes principales (pca), con el cual considere que le permita una mejor exploración de los datos
    1. Implementación: sea claro en señalar los argumentos y parámetros utilizados en el modelo.
    2. Validación: método de Elbow para cluster y porcentaje acumulado de la varianza multivariada explicada en el caso del PCA.
    3. Describa los resultados de ambos algoritmos. En el caso de cluster, en qué cuadrantes se concentran las mejor particiones de los datos y en el caso del pca mediante la matriz de correlaciones entre los componentes y las características.

Anexo: sugerencias de repositorios de datos abiertos para el sector público

Ministerio de Educación

<http://datosabiertos.mineduc.cl/>

Encuesta CASEN

[http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/casen\_2017.php#](http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/casen_2017.php)

Servicio Electoral

<https://www.servel.cl/estadisticas-2/>

Ministerio de Energía

<http://energiaabierta.cl/>

Ministerio de Economía

<https://www.economiaabierta.cl/home>

Datos abiertos públicos generales

<http://datos.gob.cl/>