Machine Learning Assignment Spring 2024

Forest Fire Prediction

Project Objective: Develop predictive models that will identify whether a forest fire will happen.

Context:

Note: **this is a fictional scenario. The numbers will not reflect actual forest/weather dynamics or geography.**

Our national forestry commission has asked you to build a machine learning model capable of classifying forests as either “fire likely” (1) or “fire unlikely” (0) in a particular time window based on details about the forest and surrounding land and air. For each forest, the commission have recorded details such as the air humidity and the tree density. You have been provided with a dataset and a data dictionary that describes the data.

Your Task:

You must develop logistic regression, decision tree and neural network models that will identify whether a fire will occur. You can use Orange, Python, R, or any machine learning package of your choice. The data for the assignment is in a file **forestdata.csv**, which you can download from the same place you found this document. The data dictionary is given at the end of this document. You must follow the correct methodology to use the data to build and test your models.

What to Submit:

You must submit a **single page** infographic poster showing the results of your analysis. Create the poster using a word processor (like Word) or a presentation package (like PowerPoint). Set the page size to A3 and use a 12pt font. You can choose the layout, but you must include:

1. Put your student number (not your name) at the top of the poster
2. A list of the steps you took to carry out the project, including details of the train / validate / test split that you chose;
3. A table showing which variables you used and whether your model treats them as numeric (continuous or discrete) or categorical. Explain one consequence of your choices;
4. A single example of how you used a histogram to detect an error in the data and what you did to fix that error. State the data cleaning operation you carried out;
5. A table showing the different models and hyper parameters you trained, along with the

correct metric for each; Add a sentence on how you chose the hyper parameter values.
6. A justification of the choice of the final model and a confusion matrix showing its results on the correct data split. Add a comment on what the confusion matrix tells us about how useful the model will be.
7. One or more references that are used to justify decisions that you made and cite any materials that you used (such as illustrations).

Important: Make sure your poster has the six sections listed above. Missing sections will cost you a lot of marks.

Save your poster as a PDF and submit it on canvas. Check the canvas page for the deadline and other submission rules.

Marks:

Marks will be allocated as follows:

Mark Range	**Requirements / Achievement**
0 – 39 (clear fail)	Many sections missing, no models built, little work completed
40 – 49 (marginal fail)	Most sections completed, but with methodological errors and poor or no interpretation
50 – 59 (pass)	All sections completed with correct methodology, but with poor or no interpretation of results
60 – 69 (merit)	All sections completed with correct methodology, good interpretation of results, good justification of decisions
70 – 100 (distinction)	All sections completed with correct methodology, excellent interpretation of results, good justification of decisions with references, excellent presentation, excellent insight.

Any Questions?

If you have any questions about this assignment, please post them on Teams under "Assessment Questions". If circumstances outside your control mean you are unable to meet the deadline, please apply for an extension under "Extension Request" on Canvas.

MOST IMPORTANT OF ALL!!

Read this part very carefully. You must work on this assignment completely on your own. Do not ask classmates for advice and do not share your answers with anybody. Do not copy any work from the internet. If you do look up something online, make sure you include a reference to it. Marks will not be given to work that is clearly not your own.

More details on academic integrity can be found here:
https://canvas.stir.ac.uk/courses/14736/pages/academic-integrity-and-academic-writing-support

Statement on the use of AI

For this coursework, the ethical and intentional use of Generative Artificial Intelligence Tools (AI) is permitted (with the exception of the use of AI for the specific purpose of writing or amending code and generate figures, which is not permitted as this assignment tests your ability to write code or visual tools to implement machine learning pipelines).

Whenever AI tools are used you should:

- Cite as a source, any AI tool used in completing your assignment.
- Acknowledge how you have used AI in your work.
- Using AI without citation or against assessment guidelines falls within the definition of plagiarism or cheating, depending on the circumstances, under the current Academic Integrity Policy, and will be treated accordingly. Making false or misleading statements as to the extent, and how AI was used, is also an example of “dishonest practice” under the policy.
- AI tools that might be relevant for this assignment include: ChatGPT and Copilot.
- Please contact Dr Brownlee or Dr Adeel for specific guidance.

Data Dictionary:

Variable	Description
collector.id	The ID of the staff member preparing the data on a given row
c.score	A score measuring the relative carbohydrate makeup of the tree species present
l.score	A score measuring the relative mass of wood to leaves in the tree species present
rain	The amount of rain over the previous three days, normalised
tree.age	The average age of trees in the forest, normalised
surface.litter	An index for the amount of litter/trash found in the forest (potential sources of ignition)
wind.intensity	A measure of average wind speeds over the past 24 hours
humidity	Air humidity at time of measurement
tree.density	Number of trees per square metre, normalised
month	Month in the year (1=January, 12=December)
time.of.day	Time period in the day when the record is taken
fire	Whether a fire happens (0 = no fire, 1 = fire)