|  |  |
| --- | --- |
| **University of Aberdeen**  **School of Natural and Computing Sciences**  **Department of Computing Science**  **MSc in Artificial Intelligence**  **2019 – 2020** | |
| **Assessment Item 2 of 2 Briefing Document – Individually Assessed (no teamwork)** | |
| **Title: CS551G - Data Mining and Visualisation** | Note: This assessment accounts for 50% of your total mark of the course. |
| **Learning Outcomes**  On successful completion of this component a student will have demonstrated competence in the following areas:   * Using a non-trivial dataset, plan, execute and evaluate significant experimental investigations using multiple data mining, visualisation and machine learning strategies. * Gain experience on using distributed processing ecosystems for data mining. | |
| **Information for Plagiarism:** The source code and your report may be submitted for plagiarism check (e.g., Turnitin). Please refer to the slides available at MyAberdeen for more information about avoiding plagiarism before you start working on the assessment. Please also read the following information provided by the university: <https://www.abdn.ac.uk/sls/online-resources/avoiding-plagiarism/>  **Application Problem Definition: Classification of Pressurised Water Reactor Status**  The objective of this assessment is to analyse a large dataset concerning pressurised water reactor data, specifically on the properties involved in the fuel assemblies cluster vibrations, alterations of thermal and hydraulic parameters, etc. For over 70 years, the nuclear power industry – in the UK and worldwide – have primarily focused on the technological evolution of reliable nuclear power plants to produce electricity. By monitoring pressurised water reactors (a type of nuclear reactor), whilst operating at nominal conditions, it is possible to collect valuable insight and extract knowledge for early detection of abnormal events. Various types of fluctuations and perturbations can be caused by the turbulent nature of flow in the core, mechanical vibrations within the reactor, the boiling coolant and stochastic character (random noise). The dataset can be downloaded from MyAberdeen. It is based on data from a research project that investigates how to detect anomalies and events in pressurised water reactors. The dataset includes two classes (normal/abnormal condition) and a number of features, which will need to be utilised throughout this assessment. The class membership of each row is stored in the field ‘Status’. The task is to develop a set of classification models for automatically classifying reactors as normal or abnormal, based on their parameters/features. No prior knowledge of the domain problem is needed or assumed to fulfil the requirements of this assessment.  Feature information in the dataset include:   * Various vibration measurements in different parts of the reactor (four features) * Various pressure values in different parts of the reactor (four features) * Power levels in different parts of the reactor (four features)   Status refers to the condition of the nuclear reactor, or in other words, we consider this to be our label/annotation for the sake of all implementations (first column). As this is a binary classification task, all implementations should treat the problem as such.  Unit of measurement or range of values of each feature are not relevant. However, features can be at different scales and/or measured in different units. | |
| **Report Guidance & Requirements**  Your report must conform to the below structure and include the required content as outlined in each section. Each subtask has its own marks allocated. You must supply a written report, along with the corresponding code, containing all distinct sections/subtasks that provide a full critical and reflective account of the processes undertaken.  This assessment focuses on big data mining processes and visualisation, which as you already know, does not necessarily imply that the datasets should be massive, but rather that the focus is on distributed processing that can be scaled on demand in a fault tolerant way. Therefore, we will focus on the fundamentals of distributed data mining and the use of the Apache ecosystem. The Apache ecosystem covers a number of different things needed in the realm of data mining and machine learning, including but not limited to distributed real-time computational systems, streaming dataflow engines, distributed analytics and machine learning platforms to name a few. The following tasks require you to expand and elaborate upon the principles of big data mining, different components of the Apache ecosystem and some aspects on how such techniques can be used in real-life problems.  **Task 1: Description of Distributed Learning Big Data Ecosystem (10/50)**  Using your own words, the lecture material and any other relevant sources, explain the distributed big data processing ecosystem. Your description should cover the following points at a technical level (max 500 words):   * Distributed File Systems (e.g. HDFS) * Resource Manager and Scheduler (e.g. YARN) * Volume, Velocity, Variety, Value, Veracity (Big Data Characteristics) * Fault Tolerance and Resilience * Data Lakes * Apache Spark * Apache Hive * Spark ML/PySpark * SparkSQL * Docker containers (e.g. Kubernetes)   **Task 2: Develop distributed models in Apache spark to classify nuclear reactors (40/50)**  The problem we aim at tackling has been clearly described and defined earlier. This task includes a number of subtasks, each of which bears its own marks.  **Subtasks**:   1. Create an Apache spark environment and load the dataset provided. You may use CoLab or a jupyter notebook. Please create a table providing summary statistics of this dataset, i.e. mean values, standard deviations, min/max values, median values and 25%/75% percentile values. Comment on whether there are any missing values present throughout. The column that contains the labels (normal/abnormal) should not be part of the table. (**3 marks**). 2. Visualise the data as follows: create two plots, i.e. box plot and a density plot. The box plot shall include the two classes (normal/abnormal) in the x-axis, and the “Vibration\_sensor\_1” in the y-axis. The density plot shall include the feature “Vibration\_sensor\_2” with the graphs of both classes appearing in the same plot. Please elaborate on what information one can obtain from each of these two plots. (**7 marks**). 3. Using the Apache spark ecosystem, such as pyspark, sparksql or any other component needed, please develop and train a binary logistic regression model to classify the condition of nuclear reactors, i.e. normal or abnormal, based on how they operate. Split the dataset provided into a training (75%) and test (25%) sets. Please use the training set to train your developed model keeping the test set only for evaluating its performance in unseen data (**15 marks**). 4. Use the following three metrics to report the model’s performance, i.e. Precision/Recall, Accuracy and Area under the curve (AUROC). When reporting performance, please only use the test set created by yourselves (**5 marks**). 5. Repeat steps 3 and 4 but using a gradient boosted tree classifier this time (**10 marks**). | |
| **Useful Information**   * Please describe and justify each step that is needed to reproduce your results by using code-snippets, screenshots and plots. When using screenshots or plots generated in Python please make sure they are clearly readable. * As the provided dataset is a subset of a real-life problem, the performance expected might not be as high as you might think. Therefore, as long as your implementations and justifications are correct the performance achieved will not have any effect on your marks whatsoever. * If you use open source code, you must point out where it was obtained from (even if the sources are online tutorials or blogs) and detail any modifications you have made to it in your tasks. You should mention this in both your code and report. *Failure to do so will result in zero marks being awarded on related (sub)tasks.*   **Marking Criteria**   * Quality of the report, including structure, clarity, and brevity. * Reproducibility. How easy is it for another MSc AI student to repeat your work based on your report and code? * Quality of your experiments, including design and result presentation (use of figures and tables for better reporting). * Configured to complete the task and the parameter tuning process (if needed). * In-depth analysis of the results generated, including critical evaluation, insights into data, and significant conclusions. * Quality of the source code, including the documentation of the code. | |
| **Submission Instructions**  You should submit a PDF version of your report along with your code via MyAberdeen by 23:59 on Sunday 26th April 2020. The name of the PDF file should have the form “CS551G\_Assessment2\_< your Surname>\_<your first name>\_<Your Student ID>”. For instance, “CS551G\_Assessment2\_Smith\_John\_4568985.pdf”, where 4568985 is your student ID.  You should submit your code and any associated files along with your report. If you have additional files that you wish to include then these should also be included in your submission.  If you have more than two files to submit, please compress all your files into one “zip” file (other format of compression files will not be accepted). Please try to make your submission files less than 10MB as you may have issues when uploading large files to MyAberdeen.  Any questions pertaining to any aspects of this assessment, please address them to the course coordinator Georgios Leontidis, [georgios.leontidis@abdn.ac.uk](mailto:georgios.leontidis@abdn.ac.uk). | |