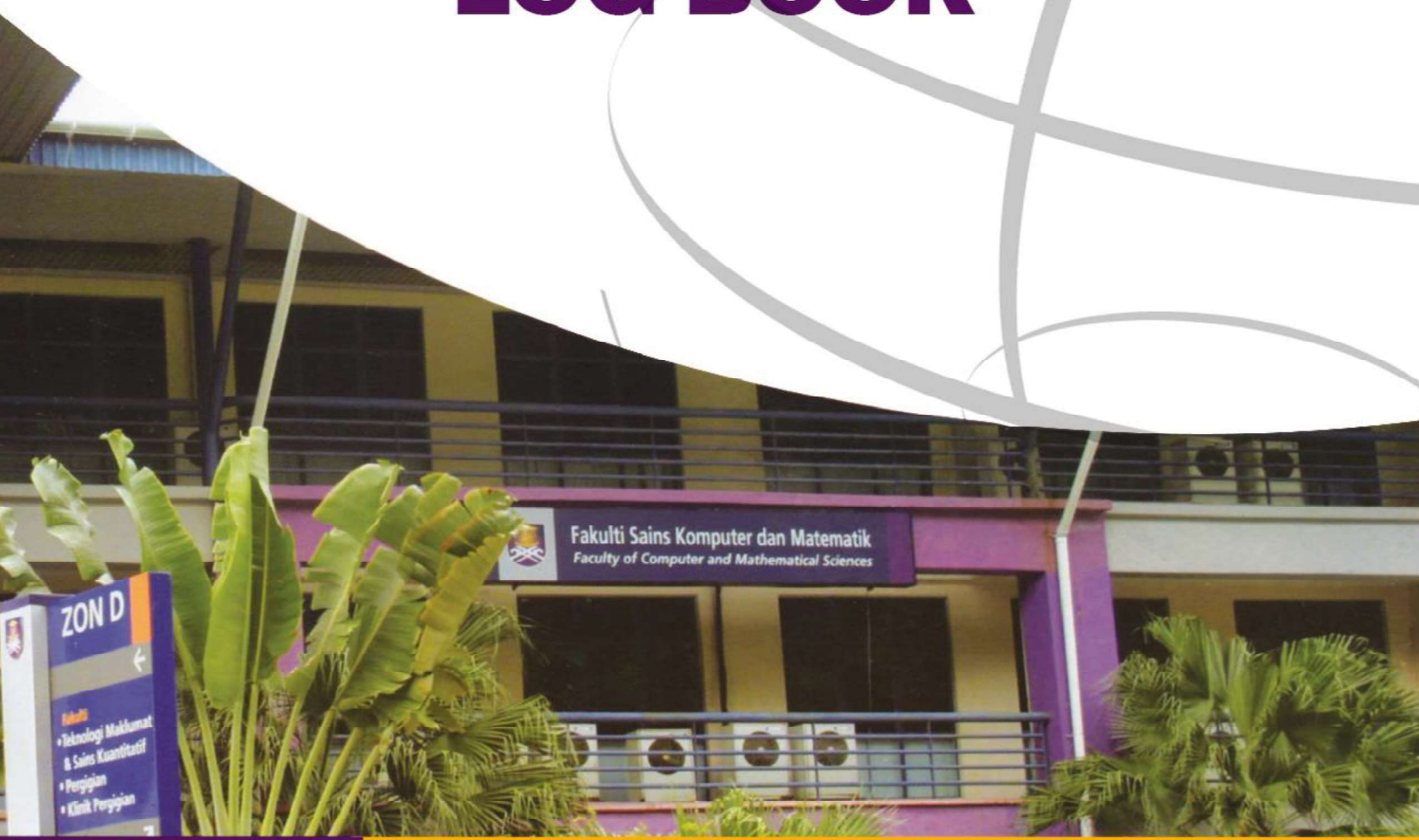
UNIVERSITI TEKNOLOGI MARA

Fakulti Sains Komputer Dan Matematik

PRACTICAL TRAINING LOG BOOK



Instructions

This book is issued to you to provide a history of your training and to act as a weekly record by the work on which you are engaged.

Student's responsibilities for keeping log book up-to-date

Immediately this book is issued to you, you should, I consultation with your Training Officer, complete the details required on the previous page.

It is your responsibility to make the main entries of the log book and keep it up to date. Entries must be regularly initialed by your Supervisor. You must ensure that;

1. It is available at your place of work during your training.
2. All entries, except sketches, are made in ink.
3. Entries are made within a week of the work to which they refer.
4. The book is handed to your Training Officer for retention on your return to UiTM and this will later be handed to the Head of School for grading.

Recording

The log book should contain the following information.

1. A neat concise description of each of your training location and the work on which you are engaged.
2. Relevant sketches, data and circuit diagrams.
3. References to textbooks, standards and other technical information related to the work being under taken.
4. Constructive comments on the work being undertaken and your considered opinion as to its value as training.

1. Student's name: MUHAMMAD AFFIQ BIN MOHD AZRIN
2. Date & Place of Birth: 27/02/1998 HOSPITAL UNIVERSITI KEBANGSAAN MALAYSIA (HUKM)
3. UiTM I/C No: 2017807094 980227-56-5003
4. Course: BACHELOR IN INFORMATION TECHNOLOGY (Hons.) INTELLIGENT SYSTEMS ENGINEERING
5. Year: OCTOBER 2020 - FEBRUARY 2021
6. Home address: NO18, JALAN BJ/30, TAMAN BAYU COURTYARD HOME, 43300, SRI KEMBANGAN, SELANGOR
7. Address during practical training: NO B-8-7, BLOK B PANGSAPURI KRISTAL HEIGHTS, JALAN KRISTAL 7/70, SEKSYEN 7, 40000 SHAH ALAM, SELANGOR
8. Place of training: AMBANK (M) BERHAD, WISMA AMFIRST, GROUND FLOOR JALAN STADIUM SS7/15, SS7, 47301 PETALING JAYA, SELANGOR
9. Name of Supervisor in-charge: TAN JUN SHENG
10. Duration of training: From: 28 SEPTEMBER 2020 to 31 DECEMBER 2021

FOR OFFICE USE ONLY

11. Remarks: (Dean/Course Tutor)

_____

_____

_____

_____

_____

_____

Week 1 (2/10)

1. Ice Breaking session.
2. Introduction to the working environment.
3. Tasks briefing.

Week 2 (9/10)

1. Apache Spark installation.
    a. Handling JAVA_HOME, SPARK_HOME, HADOOP_HOME to the correct user environment path in local machine.
    b. Execute spark-shell in cmd.
    c. TransmogrifAI library installation using spark-shell.

2. Configure Azure Databricks clients.
    a. Setup Databricks CLI and Databricks-Connect clients in virtual environment (*python vritualenv=project1*) and anaconda environment (*env=dbconnect*) respectively.
    b. Handling Databricks-Connect credentials such as server provider host, token, orgID and clusterID.
        i. *Heavyweight operations such as physical planning and execution must run on the servers in the cloud. Otherwise, the client could incur a lot of overhead reading data over the wide area network if it is not running co-located with the cluster.*
    c. Identify the correct version of clients' and cloud provider's Python, Apache Spark and Databricks.
3. Extensive exploration of Azure Databricks.
    a. Create clusters with different specifications to meet the requirements of preferred AutoML.
    b. Execute basic commands such as monitor, run, and terminate clusters using Databricks CLI.
    c. Execute example of spark (*.ipynb*) file in local machine using Databricks-Connect.
    d. Libraries installation in clusters using MAVEN, CRAN, PyPi repositories provider.

Week 3 (16/10)

4. Understanding MML Spark functions in current ML pipeline.
    a. Understanding data parallelism using Spark.
    b. Feature extractions and gradient boosting packages (*retrieved from the current JupyterHub server*) such as:
        i. *from pyspark.ml.feature import HashingTF, IDF*
        ii. *from pyspark.ml.feature import Tokenizer*
        iii. *from pyspark.ml.feature import CountVectorizer*
        iv. *import lightgbm as lgb (MMLspark)*
        v. *import xgboost as xgb (Python)*
5. Discover AutoMLs.
    a. Autogluon, ~~TransmogrifAI, Auto-SKLearn,~~ Auto-Keras, TPOT and Microsoft NNI.
    b. General comparison between AutoMLs.
        i. Alternative to MML Spark packages (*Sci-kit Learn, Dash*).
        ii. Install AutoMLs in separate local environments (*to prevent package conflicts*), identify dependencies and measure performance.

Week 4 (23/10)

6. Setup benchmark marking experiment template with a standardized dataset.
    a. Using Portuguese Bank Marketing (*41188 observations with 21 attributes, multivariate*) dataset retrieved from Kaggle as a standard dataset for the experiment.
        i. *The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).*
    b. Supervised learning (*classification*) problem with 70:30 dataset ratio.
        i. *from sklearn.model_selection import train_test_split*
    c. Measure errors such as (*error measurement might be different for every AutoML*):
        i. *RMSE*
        ii. *Accuracy*
        iii. *Precision*
7. Design machine learning pipeline scripts with the implementation of suitable AutoMLs .
    a. Data cleaning, feature pre-processing, feature selection, feature construction using Apache Spark (*Dask as alternative*).
        i. *Spark DataFrame converted into Pandas DataFrame before feed into AutoMLs.*
    b. Input vector values into AutoML for best model selection and parameter optimization (*hyperparameter tuning*).

Week 5 (30/10) extended to (2/11)

8. Explore H20, Sparkling Water and H20 AutoML
    a. **H2O** is a fully open source, distributed in-memory machine learning platform with linear scalability. H2O supports the most widely used statistical & machine learning algorithms including gradient boosted machines, generalized linear models, deep learning and more.
    b. **Sparkling Water** allows users to combine the fast, scalable machine learning algorithms of H2O with the capabilities of Spark.
    c. **H20 AutoML** trains the best model in the least amount of time to save time.
9. Installation.
    a. Version compatibility.
    b. Databricks supported.

Week 6 (5/11)

10. Integrates current prepped data from Spark to H20 ML algorithms.
    a. Converting an H2OFrame into an RDD.
    b. Converting an H2OFrame into a DataFrame.
    c. Converting an RDD[T] into an H2OFrame.
    d. Converting a DataFrame into an H2OFrame.
11. Comparison between dense vector and sparse vector
    a.
12. Manipulating default dataset using H2O scripts.
    a. XGBoost
    b. AutoML
    c. GBM
13. Export results to readable format

Week 7 (13/11) extended to Week 8 (20/11)

14. Articles reading
    a. Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools
    b. Benchmarking Automatic Machine Learning Frameworks
    c. Gartner: Magic Quadrant for Data Science and Machine Learning Platforms (2019, 2018)
15. Summarizing findings.
16. Finalizing data prep scripts using Spark.
17. Finalizing automl scripts for TPOT, H2O AutoML, AutoKeras.
    a. Enhanced automl scripts with defined functions for ROC, confusion matrix, SKlearn error metrics which can be used across multiple frameworks.

Week 9 (27/11)

18. Understanding Model Explanations frameworks
    a. LIME (Local Interpretable Model-agnostic Explanations)
    b. SHAP (Shapley Addictive explanations)
19. Understanding HashingTF, IDF libraries from Spark
    a. Example scripts
20. Data conversion from CSV to Parquet
    a. Improvise data prep pipeline.
21. Data conversion from Parquet to Delta (Delta Lake)
    a. Faster data streaming, batch processing, can be integrated with Apache Stream
    b. Improvise data prep pipeline.

Week 10 (4/12) extended to Week 12 (18/12)

22. AutoML Implementation. Final Decision
    a. Finalize AutoML results experiment.
    b. Presentation to the team and discussion.

Week 13 (25/12) extended to Week 14 (31/12)

23. Completion of internship report
    a. Submit to internship supervisor for proofing and validation.
    b. Submit to UiTM for checking.