PROGRAMMING FOR DATA ANALYTICS 1
PDAN8411/w
MODULE GUIDE 2025
(First Edition: 2022)

The Independent Institute of Education (Pty) Ltd is registered with the Department of Higher Education and Training as a private higher education institution under the Higher Education Act, 1997 (reg. no. 2007/HE07/002). Company registration number: 1987/004754/07.

DID YOU KNOW?

Student Portal

The full-service Student Portal provides you with access to your academic administrative information, including:

- an online calendar,
- timetable,
- academic results,
- module content,
- financial account, and so much more!

Module Guides or Module Manuals

When you log into the Student Portal, the ‘Module Information’ page displays the ‘Module Purpose’ and ‘Textbook Information’ including the online ‘Module Guides or ‘Module Manuals’ and assignments for each module for which you are registered.

Supplementary Materials

For certain modules, electronic supplementary material is available to you via the ‘Supplementary Module Material’ link.

Module Discussion Forum

The ‘Module Discussion Forum’ may be used by your lecturer to discuss any topics with you related to any supplementary materials and activities such as ICE, etc.

To view, print and annotate these related PDF documents, download Adobe Reader at following link below:
www.adobe.com/products/reader.html

IIE Library Online Databases

The following Library Online Databases are available. These links will prompt you for a username and password. Use the same username and password as for student portal. Please contact your librarian if you are unable to access any of these. Here are links to some of the databases:

Library Website	**This library website gives access to various online resources and study support guides** **[Link]**
LibraryConnect (OPAC)	**The Online Public Access Catalogue. Here you will be able to search for books that are available in all the IIE campus libraries.** **[Link]**
EBSCOhost	**This database contains full text online articles.** **[Link]**
EBSCO eBook Collection	**This database contains full text online eBooks.** **[Link]**
SABINET	**This database will provide you with books available in other libraries across South Africa.** **[Link]**
DOAJ	**DOAJ is an online directory that indexes and provides access to high quality, open access, peer-reviewed journals.** **[Link]**
DOAB	**Directory of open access books.** **[Link]**
IIESPACE	**The IIE open access research repository** **[Link]**
Emerald	**Emerald Insight** **[Link]**
HeinOnline	**Law database** **[Link]**
JutaStat	**Law database** **[Link]**

Table of Contents

Using this Guide

This guide has been developed to support your use of the prescribed material for this module. There may be occasions when the prescribed material does not provide sufficient detail regarding a particular idea or principle. In such instances, additional detail may be included in the guide. This guide should not, however, be used as a stand-alone textbook, as the bulk of the information that you will need to engage with will be covered in the prescribed material. You will not pass this module if you only use the module guide to study from.

Various activities and revision questions are included in the learning units of this guide. These are designed to help you to engage with the subject matter as well as to help you prepare for your assessments.

Introduction

In this practical module, you will be introduced to the Python programming language. Python is one of the main languages used in Data Analytics, and lots of libraries are available to assist with easily developing sophisticated data driven applications.

You will use Python to implement supervised and unsupervised learning, you will learn how to represent data, you will evaluate and improve models, you will implement algorithm chains and work with text data.

We hope you will enjoy the module and learn lots of practically useful skills.

Module Resources	
Prescribed Book for this Module	*Please note that this module guide is intended to support your learning – the content of this module should be sourced from the prescribed material. You will not succeed in this module if you focus on this module guide only.* Muller, A.C. and Guido, S. (2016.) *Introduction to Machine Learning with Python, A Guide for Data Scientists*. O'Reilly Media, Inc. ISBN: 978-1-449-36941-5
Recommended Additional Reading	*The following titles include information related to this module and may be consulted as additional resources. Please note, however, that you will not be tested on any content from these titles.* Severance, C.R. 2016. *Python for Everybody. [Online]* Available at: https://www.py4e.com/html3/ [Accessed 24 January 2022].

Module Purpose
This module deepens and extends existing programming and development of knowledge and skills into the specialist field of Data Analytics. Students use appropriate software tools to retrieve, prepare, explore, model data and present the results to solve business problems.

Module Outcomes	
MO1	Develop applications that employ logical regression to solve complex problems and perform statistical testing.
MO2	Solve complex problems using deep learning networks.
MO3	Develop applications to identify patterns in large, unstructured data sets.
MO4	Create applications that apply data visualisation.
MO5	Create applications that perform statistical testing.

Pacer and Assessment Brief Applicable to Module: MODULE PDAN8411

<table>
<tr><th colspan="6">Module Pacer</th></tr>
<tr><td>Code</td><td>PDAN8411</td><td>Hour Sessions</td><td>52</td><td rowspan="2">Credits</td><td rowspan="2">15</td></tr>
<tr><td>Code</td><td>PDAN8411w</td><td>Hour Sessions</td><td>12</td></tr>
<tr><td colspan="2">Learning Unit 1</td><td colspan="2">Theme: Introduction to Python</td><td colspan="2">Notes on this LU</td></tr>
<tr><td colspan="2">PDAN8411
Sessions: 1–4

PDAN8411w:
Sessions: 1

Related Outcomes:
MO3
MO4</td><td colspan="2">Learning objectives:

<u>Theme 1: Why Machine Learning?</u>

LO1: Identify the types of problems that can be solved with machine learning.

<u>Theme 2: Introduction to Python and Git</u>

LO2: Identify the essential Python libraries used for machine learning.
LO3: Defend the use of Python for machine learning.
LO4: Justify the use of version control in machine learning.
LO5: Create a program to visualise data using graphs.</td><td colspan="2">This unit introduces machine learning, as well as the Python programming language and the libraries and tools that you will use throughout the module.

This unit covers Chapter 1 of the prescribed textbook, as well some online resources.</td></tr>
</table>

Learning Unit 2	**Theme: Supervised Learning**	*Notes on this LU*
PDAN8411 Sessions: 5–12 PDAN8411w: Sessions: 2–3 Related Outcomes: MO1 MO2 MO5	**Learning objectives:** <u>Theme 1: Classification, Regression, Generalisation, Overfitting and Underfitting</u> LO1: Classify machine learning problems as classification or regression problems. LO2: Distinguish between underfitting and overfitting. <u>Theme 2: Supervised Machine Learning Algorithms and Uncertainty Estimates</u> LO3: Compare different supervised learning algorithms. LO4: Apply supervised learning algorithms to solve problems. LO5: Calculate uncertainty estimates from classifiers.	*This learning unit introduces several supervised machine learning algorithms.* *This unit covers Chapter 2 of the prescribed textbook.*

Learning Unit 3	**Theme: Unsupervised Learning**	*Notes on this LU*
PDAN8411 Sessions: 13–20 PDAN8411w: Sessions: 4–5 Related Outcomes: MO3	**Learning objectives:** <u>Theme 1: Pre-processing and Scaling</u> LO1: Compare unsupervised transformations and clustering. LO2: Identify challenges in unsupervised learning. LO3: Compare different kinds of pre-processing. LO4: Apply pre-processing to data. LO5: Assess the effect of pre-processing on data. <u>Theme 2: Dimensionality Reduction, Feature Extraction, and Manifold Learning</u> LO6: Compare different unsupervised learning algorithms. LO7: Apply unsupervised learning algorithms to solve problems. <u>Theme 3: Clustering</u> LO8: Compare the different clustering algorithms. LO9: Apply clustering algorithms to data sets.	*This unit introduces unsupervised learning algorithms.* *This unit covers Chapter 3 of the prescribed textbook.*

Learning Unit 4	**Theme: Representing Data and Engineering Features**	*Notes on this LU*
PDAN8411 Sessions: 21–24 PDAN8411w: Sessions: 6 Related Outcomes: MO3 MO4	**Learning objectives:** <u>Theme 1: Representing Data and Engineering Features</u> LO1: Apply one hot encoding to data. LO2: Critically examine methods for enriching a feature representation. LO3: Apply methods for enriching feature representation. LO4: Compare strategies for evaluating features. LO5: Apply strategies for evaluating features.	*This learning unit will explain how to represent any arbitrary data in a way that can be used for machine learning.* *This unit covers Chapter 4 of the prescribed textbook.*

Learning Unit 5	**Theme: Model Evaluation and Improvement**	*Notes on this LU*
PDAN8411 Sessions: 25–32 PDAN8411w: Sessions: 7–8 Related Outcomes: MO3	**Learning objectives:** <u>Theme 1: Cross Validation</u> LO1: Justify the use of cross validation for evaluating generalisation performance. LO2: Apply cross validation to assess model performance. <u>Theme 2: Grid Search</u> LO3: Criticise the performance of simple grid search. LO4: Apply grid search with cross-validation to improve model performance. <u>Theme 3: Evaluation Metrics and Scoring</u> LO5: Choose the best metrics for evaluating a model. LO6: Apply evaluation metrics in model selection.	*In this learning unit you will learn how to evaluate and improve models in supervised learning.* *This unit covers Chapter 5 of the prescribed textbook.*

Learning Unit 6	**Theme: Algorithm Chains and Pipelines**	*Notes on this LU*
PDAN8411 Sessions: 33–40 PDAN8411w: Sessions: 9–10 Related Outcomes: MO3	**Learning objectives:** <u>Theme 1: Pipelines</u> LO1: Apply pipelines in Python to chain multiple steps. LO2: Plan the steps used in a pipeline. <u>Theme 2: Grid-Searching Pre-Processing Steps and Models</u> LO3: Assess the suitability of grid-search for choosing pre-processing steps. LO4: Apply-grid searching to selecting pre-processing steps. LO5: Assess the suitability of grid-search for determining which model to use. LO6: Apply grid-search to determine which model to use.	*This learning unit explorers how to chain multiple processing steps and machine learning models to create a complete machine learning application.* *This unit covers Chapter 6 of the prescribed textbook.*

Learning Unit 7	**Theme: Working with Text Data**	*Notes on this LU*
PDAN8411 Sessions: 41–48 PDAN8411w: Sessions: 11–12 Related Outcomes: MO3	**Learning objectives:** <u>Theme 1: Working with Text Data</u> LO1: Differentiate between the types of string data. LO2: Classify data as one of the types of string data. LO3: Use a bag of words to represent text data. LO4: Differentiate between stop words and meaningful words. LO5: Apply the term frequency–inverse document frequency method to text data. LO6: Apply advanced tokenisation to text data. LO7: Apply stemming to text data. LO8: Apply lemmatisation to text data. LO9: Apply Latent Dirichlet Allocation to text data.	*This learning unit explains how to extract useful information out of text data such as email messages.* *This unit covers Chapter 7 of the prescribed textbook.*

Assessments

Integrated Curriculum Engagement (ICE)	
Minimum number of ICE activities to complete	4
Weighting towards the final module mark	10%

Assignments/Projects	Part 1	Part 2	POE
Weighting	25%	30%	35%
Duration	10 hours	10 hours	10 hours
Submit after	Learning Unit 3	Learning Unit 5	
Learning Units covered	LU 1 to 3	LU 1 to 5	All
Resources required	Python	Python	Python

Assessment Preparation Guidelines		
	Format of the Assessment ***(The Focus/Approach/ Objectives)***	**Preparation Hints** ***(How to Prepare, Resources to Use, etc.)***
Part 1	This assignment will assess your understanding of Learning Units 1 to 3 of this module and will be a practical data analytics programming task.	• Ensure that you work through all the relevant activities, exercises and revision questions on Learn and in your textbook. • Pay attention to the instructions and to the mark allocations of each question to ensure that you are able to meet the requirements. • Make sure that you have mastered the objectives in Learning Units 1 to 3.
Part 2	This assignment will assess your understanding of Learning Units 1 to 5 of this module and will be a practical data analytics programming task.	• Ensure that you work through all the relevant activities, exercises and revision questions on Learn and in your textbook. • Pay attention to the instructions and to the mark allocations of each question to ensure that you are able to meet the requirements. • Make sure that you have mastered the objectives in Learning Units 1 to 5.
Portfolio of Evidence (POE)	This assignment will assess your understanding of all Learning Units of this module and will be a practical data analytics programming task.	• Ensure that you work through all the relevant activities, exercises and revision questions on Learn and in your textbook. • Pay attention to the instructions and to the mark allocations of each question to ensure that you are able to meet the requirements. • Make sure that you have mastered the objectives in all Learning Units.

Glossary of Key Terms for this Module

Term	Definition	*My Notes*
Algorithm	"A set of steps that are followed in order to solve a mathematical problem or to complete a computer process." (Meriam-Webster, 2021)	
Git	A distributed version control system.	
Machine Learning	"Machine learning is a subset of artificial intelligence (AI). It is focused on teaching computers to learn from data and to improve with experience – instead of being explicitly programmed to do so." (SAP, n.d.)	
Python	Programming language used in machine learning.	
Supervised Machine Learning	"Supervised Machine Learning is an algorithm that learns from labeled training data to help you predict outcomes for unforeseen data. In Supervised learning, you train the machine using data that is well 'labeled.'" (Johnson, 2021)	
Unsupervised Machine Learning	"Unsupervised learning is a machine learning technique in which models are not supervised using training dataset." (JavaTpoint, 2021)	

Learning Unit 1: Introduction to Python	
Learning Objectives: • Identify the types of problems that can be solved with machine learning. • Identify the essential Python libraries used for machine learning. • Defend the use of Python for machine learning. • Justify the use of version control in machine learning. • Create a program to visualise data using graphs.	**My notes**
Material used for this learning unit: • Chapter 1 of the prescribed textbook • Online resources: ○ Python for Everybody ○ Getting Started with Jupyter Notebook ○ Introduction to Git for Data Science	
How to prepare for this learning unit: • Read through the material • Install Python	

1 Introduction

This unit introduces machine learning, as well as the Python programming language and the libraries and tools that you will use throughout the module.

After reading through Chapter 1 in the prescribed textbook, you will already know about some of the libraries that will be used throughout the module. However, it would be beneficial to get a better understanding of the basics of Python before we jump into the machine learning programs in the next learning unit. Work through the recommended additional reading in some detail before attempting the activities.

If you are familiar with programming languages like C# and Java, you will encounter similar concepts in Python. However, there are also some significant differences. For example, indentation (white space) is meaningful in Python. So, it is worthwhile spending some time getting used to the differences.

2 Recommended Additional Reading

Read through the following sources in the given order:

Severance, C.R. 2016. *Python for Everybody.* [Online] Available at: https://www.py4e.com/html3/ [Accessed 24 January 2022].

Elance, P. 2019. Graph Plotting in Python. [Online] Available at: https://www.tutorialspoint.com/graph-plotting-in-python [Accessed 24 January 2022].

Singhal, G. 2019. *Getting Started with Jupyter Notebook*. [Online] Available at: https://www.pluralsight.com/guides/jupyter-notebook-getting-started [Accessed 24 January 2022].

Luvsandorj, Z. *Introduction to Git for Data Science*. [Online] Available at: https://towardsdatascience.com/introduction-to-git-for-data-science-ca5ffd1cebbe [Accessed 24 January 2022].

Tyagi, H. 2021. *Programming, Math, and Statistics You Need to Know for Data Science and Machine Learning*. [Online] Available at: https://www.freecodecamp.org/news/first-steps-to-learn-data-science-or-ml-after-the-roadmap/ [Accessed 24 January 2022].

3 Recommended Digital Engagement and Activities

After reading through the recommended additional reading, there are also video lessons available from the Python for Everybody website (https://www.py4e.com/lessons), with worked exercises.

4 Activities

4.1 Activity 1.1

Create a mind map of the tools and libraries that will be useful for machine learning in Python. Include the purpose of each tool/library, and an example of how you would use it.

If you do not have a diagram tool handy, try the free online tools at https://www.diagrams.net/ [Accessed 24 January 2022].

4.2 Activity 1.2

Create a GitHub repository where you can store all the code you write during this module.

Search online for open data sets that you can use for creating a small program. You only need a few data points to be able to draw a graph, so look for something simple.

Using Python, write a program that displays a data set as a bar graph, and another data set as a line graph. Explain why a bar graph and a line graph are good choices for chosen data sets.

Learning Unit 2: Supervised Learning	
Learning Objectives: • Classify machine learning problems as classification or regression problems. • Distinguish between underfitting and overfitting. • Compare different supervised learning algorithms. • Apply supervised learning algorithms to solve problems. • Calculate uncertainty estimates from classifiers.	**My notes**
Material used for this learning unit: • Chapter 2 of the prescribed textbook	
How to prepare for this learning unit: • Read through the prescribed textbook chapter. • Sign up for a free Microsoft Learn account.	

1 Introduction

This learning unit introduces several supervised machine learning algorithms. Work through the examples in the textbook to gain an understanding of the concepts.

In this learning unit, we will be work through some modules on Microsoft Learn. Although the modules can be accessed without signing into Microsoft Learn, you would miss out on earning XP and levels and badges. So go ahead and sign up for your free Microsoft Learn account at: https://docs.microsoft.com/en-gb/learn/ [Accessed 24 January 2022].

2 Recommended Additional Reading

Brownlee, J. 2016. *Overfitting and Underfitting With Machine Learning Algorithms*. [Online] Available at: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/ [Accessed 24 January 2022].

Zakharchuk, I. 2021. *Generalization, Overfitting, and Underfitting in Supervised Learning*. [Online] Available at: https://medium.com/mlearning-ai/generalization-overfitting-and-underfitting-in-supervised-learning-a21f02ebf3df [Accessed 24 January 2022].

Fumo, D. 2017. *Types of Machine Learning Algorithms You Should Know*. [Online] Available at: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861 [Accessed 24 January 2022].

Uddin, S., Khan, A., Hossain, E. and Moni, M.A. 2019. *Comparing different supervised machine learning algorithms for disease prediction*. [Online] Available at: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8 [Accessed 24 January 2022].

3 Activities

3.1 Activity 2.1

Work through the following module on Microsoft Learn:

Build a machine learning model
https://docs.microsoft.com/en-us/learn/modules/machine-learning-model-nasa/ [Accessed 24 January 2022].

In this activity, you get to work with Nasa data to predict whether a rocket launch will be successful or not. This activity makes use of a decision tree.

What did you learn from this activity?

3.2 Activity 2.2

Work through the following learning path on Microsoft Learn. Note that you will need an Azure student subscription for some parts of this learning path. Your lecturer may decide to assign one or more specific modules from this learning path for you to do.

Create machine learning models
https://docs.microsoft.com/en-us/learn/paths/create-machine-learn-models/ [Accessed 24 January 2022].

What did you Learn from this learning path? Is there anything that you think will be generally useful for you in your career?

Learning Unit 3: Unsupervised Learning	
Learning Objectives: • Compare unsupervised transformations and clustering. • Identify challenges in unsupervised learning. • Compare different kinds of pre-processing. • Apply pre-processing to data. • Assess the effect of pre-processing on data. • Compare different unsupervised learning algorithms. • Apply unsupervised learning algorithms to solve problems. • Compare the different clustering algorithms. • Apply clustering algorithms to data sets.	**My notes**
Material used for this learning unit: • Chapter 3 of the prescribed textbook.	
How to prepare for this learning unit: • Read through the prescribed textbook chapter.	

1 Introduction

This unit introduces unsupervised learning algorithms. Work through the textbook examples to gain an understanding of the concepts.

2 Recommended Additional Reading

Johnson, D. 2021. *Unsupervised Machine Learning: What is, Algorithms, Example*. [Online] Available at: https://www.guru99.com/unsupervised-machine-learning.html [Accessed 24 January 2022].

Durmus, M. 2018. *Importance of Unsupervised Learning in data preprocessing*. [Online] Available at: https://www.aisoma.de/ importance-of-unsupervised-learning-in-data-preprocessing/ [Accessed 24 January 2022].

Awasthi, S. 2021. *Five Most Popular Unsupervised Learning Algorithms*. [Online] Available at: https://dataaspirant.com/unsupervised-learning-algorithms/ [Accessed 24 January 2022].

Brownlee, J. 2021. *10 Clustering Algorithms With Python*. [Online] Available at: https://machinelearningmastery.com/clustering-algorithms-with-python/ [Accessed 24 January 2022].

Li, B. and Lu, P. 2021. *Component: K-Means Clustering*. [Online] Available at: https://docs.microsoft.com/en-us/azure/machine-learning/component-reference/k-means-clustering [Accessed 24 January 2022].

Carnes, B. 2021. *Python scikit-learn Tutorial – Machine Learning Crash Course*. [Online] Available at: https://www.freecodecamp.org/news/learn-scikit-learn/ [Accessed 24 January 2022].

3 Activities

3.1 Activity 3.1

Work through the following module on Microsoft Learn. Note that you will need access to an Azure student account to work through this activity.

Create a Clustering Model with Azure Machine Learning designer
https://docs.microsoft.com/en-us/learn/modules/create-clustering-model-azure-machine-learning-designer/ [Accessed 24 January 2022].

3.2 Activity 3.2

Work through the following module on Microsoft Learn. Note that you will need access to an Azure student account to work through this activity.

Microsoft Azure AI Fundamentals: Explore decision support
https://docs.microsoft.com/en-us/learn/paths/explore-fundamentals-of-decision-support/ [Accessed 24 January 2022].

Learning Unit 4: Representing Data and Engineering Features	
Learning Objectives: • Apply one hot encoding to data. • Critically examine methods for enriching a feature representation. • Apply methods for enriching feature representation. • Compare strategies for evaluating features. • Apply strategies for evaluating features.	**My notes**
Material used for this learning unit: • Chapter 4 of the prescribed textbook.	
How to prepare for this learning unit: • Read through the prescribed textbook chapter.	

1 Introduction

This learning unit will explain how to represent any arbitrary data in a way that can be used for machine learning. Work through the textbook examples to learn about the concepts.

2 Recommended Additional Reading

Brownlee, J. 2017. *Why One-Hot Encode Data in Machine Learning?* [Online] Available at: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/ [Accessed 24 January 2022].

Tannenbaum, L. 2018. *ML Intro 5: One hot Encoding, Cyclic Representations, and Normalization*. [Online] Available at: https://towardsdatascience.com/ml-intro-5-one-hot-encoding-cyclic-representations-normalization-6f6e2f4ec001 [Accessed 24 January 2022].

Brownlee, J. 2020. *How to Choose a Feature Selection Method For Machine Learning*. [Online] Available at: https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/ [Accessed 24 January 2022].

Scikit-learn. 2021. *sklearn.preprocessing.OneHotEncoder*. [Online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html [Accessed 24 January 2022].

David, D. 2021. *How to Improve Machine Learning Model Performance by Combining Categorical Features.* [Online] Available at: https://www.freecodecamp.org/news/improve-machine-learning-model-performance-by-combining-categorical-features/ [Accessed 24 January 2022].

Carnes, B. 2021. *Python scikit-learn Tutorial – Machine Learning Crash Course*. [Online] Available at: https://www.freecodecamp.org/news/learn-scikit-learn/ [Accessed 24 January 2022].

3 Activities

3.1 Activity 4.1

Work through the following tutorial on feature engineering and feature selection:

David, D. 2021. *Machine Learning Tutorial – Feature Engineering and Feature Selection For Beginners*. [Online] Available at: https://www.freecodecamp.org/news/feature-engineering-and-feature-selection-for-beginners/ [Accessed 24 January 2022].

3.2 Activity 4.2

Work through the instructions on the below page to create features for data in SQL Server.

Create features for data in SQL Server using SQL and Python https://docs.microsoft.com/en-us/azure/architecture/data-science-process/create-features-sql-server [Accessed 24 January 2022].

3.3 Activity 4.3

Work through the instructions on the below page to learn about using one-hot encoding.

One-Hot Encoding in Python with Pandas and Scikit-Learn https://stackabuse.com/one-hot-encoding-in-python-with-pandas-and-scikit-learn/ [Accessed 24 January 2022].

Learning Unit 5: Model Evaluation and Improvement

Learning Objectives:	**My notes**
• Justify the use of cross validation for evaluating generalisation performance. • Apply cross validation to assess model performance. • Criticise the performance of simple grid search. • Apply grid search with cross-validation to improve model performance. • Choose the best metrics for evaluating a model. • Apply evaluation metrics in model selection.	
Material used for this learning unit: • Chapter 5 of the prescribed textbook.	
How to prepare for this learning unit: • Read through the prescribed textbook chapter.	

1 Introduction

In this learning unit, you will learn how to evaluate and improve models in supervised learning. Work through the examples in the textbook to see model evaluation and improvement in action.

2 Recommended Additional Reading

scikit-learn. 2021. *3.1. Cross-validation: evaluating estimator performance.* [Online] Available at: https://scikit-learn.org/stable/modules/cross_validation.html [Accessed 24 January 2022].

scikit-learn. 2021. *Parameter estimation using grid search with cross-validation*. [Online] Available at: https://scikit-learn.org/stable/auto_examples/model_selection/plot_grid_search_digits.html [Accessed 24 January 2022].

Srivastava, T. 2019. *11 Important Model Evaluation Metrics for Machine Learning Everyone should know*. [Online] Available at: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/ [Accessed 24 January 2022].

Li, B., Buck, A. and Lu, P. 2021. *Evaluate Model component*. [Online] Available at: https://docs.microsoft.com/en-us/azure/machine-learning/component-reference/evaluate-model [Accessed 24 January 2022].

3 Activities

3.1 Activity 5.1

The below four tutorials all build on one another. Parts 3 and 4 are about model evaluation. Work through all of these tutorials to get hands on experience with model evaluation.

Practical Machine Learning Tutorial: Part.1 (Exploratory Data Analysis)
https://towardsdatascience.com/practical-machine-learning-tutorial-part-1-data-exploratory-analysis-c13d39b8f33b
[Accessed 24 January 2022].

Practical Machine Learning Tutorial: Part.2 (Build Model & Validate)
https://towardsdatascience.com/practical-machine-learning-tutorial-part-2-build-model-validate-c98c2ddad744 [Accessed 24 January 2022].

Practical Machine Learning Tutorial: Part.3 (Model Evaluation-1)
https://towardsdatascience.com/practical-machine-learning-tutorial-part-3-model-evaluation-1-5eefae18ec98 [Accessed 24 January 2022].

Practical Machine Learning Tutorial: Part.4 (Model Evaluation-2)
https://towardsdatascience.com/practical-machine-learning-tutorial-part-4-model-evaluation-2-764d69f792a5 [Accessed 24 January 2022].

Learning Unit 6: Algorithm Chains and Pipelines	
Learning Objectives: • Apply pipelines in Python to chain multiple steps. • Plan the steps used in a pipeline. • Assess the suitability of grid-search for choosing pre-processing steps. • Apply-grid searching to selecting pre-processing steps. • Assess the suitability of grid-search for determining which model to use. • Apply grid-search to determine which model to use.	**My notes**
Material used for this learning unit: • Chapter 6 of the prescribed textbook.	
How to prepare for this learning unit: • Read through the prescribed textbook chapter.	

1 Introduction

This learning unit explores how to chain multiple processing steps and machine learning models to create a complete machine learning application. Work through the examples in the textbook to learn about algorithm chains and pipelines.

2 Recommended Additional Reading

scikit-learn. 2021. *sklearn.pipeline.Pipeline*. [Online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html [Accessed 24 January 2022].

Brownlee, J. 2020. *How to Grid Search Data Preparation Techniques*. [Online] Available at: https://machinelearningmastery.com/grid-search-data-preparation-techniques/ [Accessed 24 January 2022].

Malik, U. 2019. *Cross Validation and Grid Search for Model Selection in Python*. [Online] Available at: https://stackabuse.com/cross-validation-and-grid-search-for-model-selection-in-python/ [Accessed 24 January 2022].

3 Activities

3.1 Activity 6.1

Work through the following tutorial:

Azure Machine Learning Notebook Code and run as pipeline — Automate using Azure Data Factory
https://medium.com/analytics-vidhya/azure-machine-learning-notebook-code-and-run-as-pipeline-automate-using-azure-data-factory-4128c270bb0e [Accessed 24 January 2022].

3.2 Activity 6.2

Work through the following tutorial:

ML Pipeline with Grid Search in Scikit-Learn
https://towardsdatascience.com/ml-pipelines-with-grid-search-in-scikit-learn-2539d6b53cfb [Accessed 24 January 2022].

Learning Unit 7: Working with Text Data	
Learning Objectives: • Differentiate between the types of string data. • Classify data as one of the types of string data. • Use a bag of words to represent text data. • Differentiate between stop words and meaningful words. • Apply the term frequency–inverse document frequency method to text data. • Apply advanced tokenisation to text data. • Apply stemming to text data. • Apply lemmatisation to text data. • Apply Latent Dirichlet Allocation to text data.	**My notes**
Material used for this learning unit: • Chapter 7 of the prescribed textbook.	
• .	

1 Introduction

This learning unit explains how to extract useful information out of text data such as email messages. Work through the textbook examples to learn how to work with text data.

2 Recommended Additional Reading

Mujtaba, H. 2020. *An Introduction to Bag of Words (BoW) | What is Bag of Words?* [Online] Available at: https://www.mygreatlearning.com/blog/bag-of-words/ [Accessed 24 January 2022].

Scott, W. 2019. *TF-IDF from scratch in python on a real-world dataset*. [Online] Available at: https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089 [Accessed 24 January 2022].

Jabeen, H. 2018. *Stemming and Lemmatization in Python*. [Online] Available at: https://www.datacamp.com/community/tutorials/stemming-lemmatization-python [Accessed 24 January 2022].

Kapadia, S. 2019. *Topic Modeling in Python: Latent Dirichlet Allocation (LDA)*. [Online] Available at: https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0 [Accessed 24 January 2022].

3 Activities

3.1 Activity 7.1

Work through the following learning path on Microsoft Learn. Note that you will need access to an Azure student account to work through this activity. Your lecturer may give you specific modules in this learning path to do.

Microsoft Azure AI Fundamentals: Explore natural language processing
https://docs.microsoft.com/en-za/learn/paths/explore-natural-language-processing/?WT.mc_id=api_CatalogApi [Accessed 24 January 2022].

3.2 Activity 7.2

Read the following article:

Methods - Text Feature Extraction with Bag-of-Words Using Scikit Learn
https://colab.research.google.com/github/RPI-DATA/course-intro-ml-app/blob/master/content/notebooks/16-intro-nlp/03-scikit-learn-text.ipynb#scrollTo=CtVIabzzb64m [Accessed 24 January 2022].

Then work through the below tutorial to see a bag of words in action.

An introduction to Bag of Words and how to code it in Python for NLP
https://www.freecodecamp.org/news/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04/ [Accessed 24 January 2022].

Bibliography

JavaTpoint, 2021. *Unsupervised Machine Learning.* [Online] Available at: https://www.javatpoint.com/unsupervised-machine-learning [Accessed 24 January 2022].

Johnson, D., 2021. *Supervised Machine Learning: What is, Algorithms with Examples.* [Online] Available at: https://www.guru99.com/supervised-machine-learning.html [Accessed 24 January 2022].

Meriam-Webster, 2021. *algorithm.* [Online] Available at: https://www.merriam-webster.com/dictionary/algorithm [Accessed 15 November 2021].

SAP, n.d.. *What is machine learning?.* [Online] Available at: https://www.sap.com/insights/what-is-machine-learning.html [Accessed 24 January 2022].

Intellectual Property

Plagiarism occurs in a variety of forms. Ultimately though, it refers to the use of the words, ideas or images of another person without acknowledging the source using the required conventions. The IIE publishes a Quick Reference Guide that provides more detailed guidance, but a brief description of plagiarism and referencing is included below for your reference. It is vital that you are familiar with this information and the Intellectual Integrity Policy before attempting any assignments.

Introduction to Referencing and Plagiarism

What is 'Plagiarism'?

'Plagiarism' is the act of taking someone's words or ideas and presenting them as your own.

What is 'Referencing'?

'Referencing' is the act of citing or giving credit to the authors of any work that you have referred to or consulted. A 'reference' then refers to a citation (a credit) or the actual information from a publication that is referred to.

Referencing is the acknowledgment of any work that is not your own, but is used by you in an academic document. It is simply a way of giving credit to and acknowledging the ideas and words of others.

When writing assignments, students are required to acknowledge the work, words or ideas of others through the technique of referencing. Referencing occurs in the text at the place where the work of others is being cited, and at the end of the document, in the bibliography.

The bibliography is a list of all the work (published and unpublished) that a writer has read in the course of preparing a piece of writing. This includes items that are not directly cited in the work.

A reference is required when you:

- Quote directly: when you use the exact words as they appear in the source;
- Copy directly: when you copy data, figures, tables, images, music, videos or frameworks;
- Summarise: when you write a short account of what is in the source;
- Paraphrase: when you state the work, words and ideas of someone else in your own words.

It is standard practice in the academic world to recognise and respect the ownership of ideas, known as intellectual property, through good referencing techniques. However, there are other reasons why referencing is useful.

Good Reasons for Referencing

It is good academic practice to reference because:

- It enhances the quality of your writing;
- It demonstrates the scope, depth and breadth of your research;
- It gives structure and strength to the aims of your article or paper;
- It endorses your arguments;
- It allows readers to access source documents relating to your work, quickly and easily.

Sources

The following would count as ‘sources’:

- Books,
- Chapters from books,
- Encyclopaedias,
- Articles,
- Journals,
- Magazines,
- Periodicals,
- Newspaper articles,
- Items from the Internet (images, videos, etc.),
- Pictures,
- Unpublished notes, articles, papers, books, manuscripts, dissertations, theses, etc.,
- Diagrams,
- Videos,
- Films,
- Music,
- Works of fiction (novels, short stories or poetry).

What You Need to Document from the Hard Copy Source You are Using

(Not every detail will be applicable in every case. However, the following lists provide a guide to what information is needed.)

You need to acknowledge:

- The words or work of the author(s),
- The author(s)'s or editor(s)'s full names,
- If your source is a group/ organisation/ body, you need all the details,
- Name of the journal, periodical, magazine, book, etc.,
- Edition,
- Publisher's name,
- Place of publication (i.e. the city of publication),
- Year of publication,
- Volume number,
- Issue number,
- Page numbers.

What You Need to Document if you are Citing Electronic Sources

- Author(s)'s/ editor(s)'s name,
- Title of the page,
- Title of the site,
- Copyright date, or the date that the page was last updated,
- Full Internet address of page(s),
- Date you accessed/ viewed the source,
- Any other relevant information pertaining to the web page or website.

Referencing Systems

There are a number of referencing systems in use and each has its own consistent rules. While these may differ from system-to-system, the referencing system followed needs to be used consistently, throughout the text. Different referencing systems cannot be mixed in the same piece of work!

A detailed guide to referencing, entitled Referencing and Plagiarism Guide is available from your library. Please refer to it if you require further assistance.

When is Referencing Not Necessary?

This is a difficult question to answer – usually when something is 'common knowledge'. However, it is not always clear what 'common knowledge' is.

Examples of 'common knowledge' are:

- Nelson Mandela was released from prison in 1990;
- The world's largest diamond was found in South Africa;
- South Africa is divided into nine (9) provinces;
- The lion is also known as 'The King of the Jungle'.
- $E = mc^2$
- The sky is blue.

Usually, all of the above examples would not be referenced. The equation $E = mc^2$ is Einstein's famous equation for calculations of total energy and has become so familiar that it is not referenced to Einstein.

Sometimes what we think is 'common knowledge', is not. For example, the above statement about the sky being blue is only partly true. The light from the sun looks white, but it is actually made up of all the colours of the rainbow. Sunlight reaches the Earth's atmosphere and is scattered in all directions by all the gases and particles in the air. The smallest particles are by coincidence the same length as the wavelength of blue light. Blue is scattered more than the other colours because it travels as shorter, smaller waves. It is not entirely accurate then to claim that the sky is blue. It is thus generally safer to always check your facts and try to find a reputable source for your claim.

Important Plagiarism Reminders

The IIE respects the intellectual property of other people and requires its students to be familiar with the necessary referencing conventions. Please ensure that you seek assistance in this regard before submitting work if you are uncertain.

If you fail to acknowledge the work or ideas of others or do so inadequately this will be handled in terms of the Intellectual Integrity Policy (available in the library) and/ or the Student Code of Conduct – depending on whether or not plagiarism and/ or cheating (passing off the work of other people as your own by copying the work of other students or copying off the Internet or from another source) is suspected.

Your campus offers individual and group training on referencing conventions – please speak to your librarian or ADC/ Campus Co-Navigator in this regard.

Reiteration of the Declaration you have signed:

1. I have been informed about the seriousness of acts of plagiarism.
2. I understand what plagiarism is.
3. I am aware that The Independent Institute of Education (IIE) has a policy regarding plagiarism and that it does not accept acts of plagiarism.
4. I am aware that the Intellectual Integrity Policy and the Student Code of Conduct prescribe the consequences of plagiarism.

5. I am aware that referencing guides are available in my student handbook or equivalent and in the library and that following them is a requirement for successful completion of my programme.
6. I am aware that should I require support or assistance in using referencing guides to avoid plagiarism I may speak to the lecturers, the librarian or the campus ADC/ Campus Co-Navigator.
7. I am aware of the consequences of plagiarism.

Please ask for assistance prior to submitting work if you are at all unsure.

