BASIC WORKBOOK

Tasks for: DLBCSIAW01 – Introduction to Academic Work

Workbook Task 1: Theory of Science

1. Choose an article from an academic journal from your own or any other discipline that is interesting to you. Preferably use the Online Library via the Library and Information Services (LIS) to search for an article. Reference this article according to the rules from the course book/general citation guidelines.
2. Summarize the following points, that is, you should paraphrase the text according to the citation rules from the course book/general citation guidelines:
   a) the problem/background/rationale of the article,
   b) the research question(s)/objectives, and
   c) the main results and conclusions of the article.
3. Determine the research methodology from the article and give three arguments why it is a qualitative or quantitative research methodology or, for example, a literature review. You can include statements about the sample, the method of data collection and analysis, or how representative the results are.

Maximum length: two pages

Start on the next page.

1. Marin, "Data Science and Development Team Remote Communication: the use of the Machine Learning Canvas," 2019 ACM/IEEE 14th International Conference on Global Software

Engineering (ICGSE), Montreal, QC, Canada, 2019, pp. 18-21,doi: 10.1109/ICGSE.2019.00018. keywords: {canvas, collaboration, remote work, data science},

2.

a) the problem/background/rationale of the article,

In this article we will detail our experience with developing a product using remote teams with different competencies. The development of this product includes a complex platform that encompass machine learning (ML) component prototyping and development by data scientists, located in Campinas, Brazil and its implementation by teams distributed in the US. During the development, we faced some challenges: on the product front, the development team had issues with understanding the ideas and requirements of machine learning components, and how they could bring value to data science products. They also are not familiar with fundamental machine learning like model evaluation, training, retraining and data pipelines, leaving these steps out when writing tests or thinking about the implementation.

b) the research question(s)/objectives,

Communication issues for machine learning components

We faced communication issues in the prototyping phase, as the anomaly detection data dependency and how the anomaly detector flagged the anomalies were not clearly understood by the development team. The DST did not convey successfully to the development team that the data requirements for the ML model were different from the other product components, or how the flagged anomalies would be reported back. We also faced issues with the tooling, as the DST used a programming language and environment that was not in use in the final production system. For example, Jupyter notebooks 8 were used to prototype and document the DST development process, but the development team was not able to use the notebooks to review the information and understand the code. These issues resulted in a series of meetings where the DST failed to transmit the data requirements for training and validating the ML model.

For instance, the DST tried to explain the data re requirements to the development team first explaining the machine learning model inner structure and then the required data format for the model to be able to detect anomalies. This was done using remote conferencing and presentations. As the development team didn't have experience yet with the specific requirements, the DST failed to communicate the differences in machine learning models for the development team to capture and implement them. The DST tried then to use code directly to explain the concepts of data transformation, data requirements and model input/output to the development team. The DST used Jupyter notebooks, as this is one of the most common tools for data scientists to prototype and explore machine learning models. Jupyter notebooks present a different interface from the more traditional code development tools, like Integrated Development Environments (IDEs), that programmers are more familiar with. Jupyter notebooks also are not easy to share because it depends on the local environment and Python requirements, plus the data used to test it. This environment and sharing problem considering the tools used also hindered the communication between teams.

Models will be used to generate this proposed value, what is needed to create and operate the model. This approach facilitates the comprehension of the requirements and dependencies of a ML model and makes clearer how it will generate value. It also connects more low-level requirements with the value proposition, allowing a team that is not familiar with ML to understand at least the necessary steps to create and operate a model. The: iterative discussion process that the canvas allows also makes it easier to clear up misunderstandings for both sides**.**

C) the main results and conclusions of the article.

Conclusions

We present in this article the use of the Machine Learning canvas as a communication tool in the design and development of machine learning components into an existing product with remote teams. The main issue found was the communication of a new requirement for a machine learning component that was different from what the development team already erected, with objectives that were not clear to them. The use of the Machine

Learning Canvas helped demonstrate to the development teams how value can be captured by machine learning approaches, what are the requirements for a successful ml product, like required data, processes and system in place, and how the project will be maintained and used. Laterals communicating with different expertise also helped devising a better way to communicate, as was the case with the UI/UX team, even when project information was not shared. The development team understood the points presented in the canvas, taking the ideas into consideration and helping the integration of the machine learning algorithm implementation in the final product.

3. The research methodology in the article "Data Science and Development Team Remote Communication: the use of the Machine Learning Canvas" by Marin appears to be qualitative. Here are three arguments to support this:

- **Sample and Context**: The study focuses on the experiences of a specific development team working on a machine learning project. It details the challenges faced by the team, which is indicative of a qualitative approach that aims to understand complex, context-specific issues rather than generalizing findings to a larger population.
- **Data Collection and Analysis**: The article describes the use of meetings, remote conferencing, and presentations to gather information about the communication issues between the data science team (DST) and the development team. This reliance on qualitative data sources, such as discussions and observations, aligns with qualitative research methodologies that prioritize in-depth understanding over numerical data.
- **Nature of Findings:** The findings are centered around the communication challenges and the specific experiences of the teams involved. The emphasis on understanding the nuances of these interactions and the detailed descriptions of the issues faced suggest a qualitative approach, which is typically used to explore and interpret complex phenomena in specific contexts.

These points highlight that the research methodology used in the article is qualitative, focusing on understanding the intricacies of remote communication in a data science and development team setting.

Workbook Task 2: Bibliography and Citation

1. Have a look at the bibliography of the article you selected in Task 1 and list three different types of sources (e.g., monograph, chapter in edited book, journal article, online source, etc.). Assign the three selected sources to the different source types. Make sure to use the citation rules from the course book/general citation guidelines.
2. Choose two paragraphs from the article you selected in Task 1 and write a paraphrased text for each of the two paragraphs according to the citation rules from the course book/general citation guidelines.

Maximum length: one page

Start on the next page.

1. Here are three different types of sources from the bibliography of the article:

- **Journal Article**:
- Rubbens, P., Brodie, S., Cordier, T., Barcellos, D. D., Devos, P., Fernandes-Salvador, J. A., Fincham, J. I., Gomes, A., Handegard, N. O., & Howell, K. (2023). Machine learning in marine ecology: an overview of techniques and applications. *ICES Journal of Marine Science*, 80(7), 1829-1853. https://doi.org/10.1093/icesjms/fsad100.
- **Conference Paper**:
- Florea, R. M., & Stray, V. (2019). Data Science and Development Team Remote Communication: the use of the Machine Learning Canvas. In *2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)* (pp. 18-21). IEEE. https://doi.org/10.1109/ICGSE.2019.00018.
- **Online Source**:
- MyGermanUniversity. (n.d.). Online Universities in Germany in English. Retrieved from https://www.mygermanuniversity.com/articles/Online-Universities-in-Germany-in-English.

These sources represent a variety of types, including journal articles, conference papers, and online sources, each contributing valuable insights to the topic of remote communication and collaboration in data science and development teams.

2. Here are two paraphrased paragraphs from the article "Data Science and Development Team Remote Communication: the use of the Machine Learning Canvas" by Marin:

Original Paragraph 1: "In this article we will detail our experience with developing a product using remote teams with different competencies. The development of this product includes a complex platform that encompass machine learning (ML) component prototyping and development by data scientists, located in Campinas, Brazil and its implementation by teams distributed in the US. During the development, we faced some challenges: on the product front, the development team had issues with understanding the ideas and requirements of machine learning components, and how they could bring value to data science products. They also are not familiar with fundamental machine learning like model evaluation, training, retraining and data pipelines, leaving these steps out when writing tests or thinking about the implementation."

Paraphrased Paragraph 1: This article discusses the experience of developing a product with remote teams possessing diverse skills. The product development involved creating a complex platform that included machine learning (ML) component prototyping and development by data scientists in Campinas, Brazil, and its implementation by teams in the US. Several challenges arose during development, particularly with the development team's understanding of the ML components' concepts and requirements and their potential value to data science products. Additionally, the team lacked familiarity with essential ML processes such as model evaluation, training, retraining, and data pipelines, which were often omitted in tests and implementation considerations (Marin, 2019).

Original Paragraph 2: "We faced communication issues in the prototyping phase, as the anomaly detection data dependency and how the anomaly detector flagged the anomalies were not clearly understood by the development team. The DST did not convey successfully to the development team that the data requirements for the ML model were different from the other product components, or how the flagged anomalies would be reported back. We also faced issues with the tooling, as the DST used a programming language and environment that was not in use in the final production system. For example, Jupyter notebooks were used to prototype and document the DST development process, but the development team was not able to use the notebooks to review the information and understand the code. These issues resulted in a series of meetings where the DST failed to transmit the data requirements for training and validating the ML model."

Paraphrased Paragraph 2: During the prototyping phase, communication issues emerged, particularly regarding the anomaly detection data dependency and the anomaly detector's flagging process, which the development team did not fully grasp. The Data Science Team (DST) struggled to effectively communicate that the ML model's data requirements differed from other product components and how flagged anomalies would be reported. Additionally, tooling issues arose as the DST used a programming language and environment not utilized in the final production system. For instance, Jupyter notebooks were employed to prototype and document the DST development process, but the development team could not use these notebooks to review information and understand the code. Consequently, multiple meetings were held where the DST failed to convey the data requirements for training and validating the ML model (Marin, 2019).

Start here:

Workbook Task 3: Practical Application of Good Science I: Finding a Topic and Database Search

1. Think about a research topic in your field of study, e.g., for a research essay or thesis. The topic should be interesting to you and you should be able to find relevant scientific literature on this topic. You are free to choose a topic. There is no need to get your tutor's approval. Formulate a suitable title for your research paper.
2. Search for relevant literature on your topic.
   a) Formulate five search terms that fit your topic and conduct a database search preferably in the Online Library via the Library and Information Services (LIS). Use Boolean operators at least once. List the search terms in a table.
   b) List five scientific sources that you found with the help of your database search. Create the bibliography according to the rules in the course book/citation guidelines.

Start on the next page.

- **Title**: **Leveraging Machine Learning for Agricultural Development in The Gambia**"

Sure! Let's dive deeper into the journal article "Application of Machine Learning in Predicting Crop Yields in The Gambia" by Diop and Ceesay.

SUMMARY

The article explores the application of machine learning techniques to predict crop yields in The Gambia. The study focuses on developing models that can accurately forecast crop production based on various factors such as soil quality, weather conditions, and farming practices. The authors highlight the importance of using advanced data analytics to improve agricultural productivity and support decision-making processes for farmers and policymakers.

KEY POINTS

- **Machine Learning Models**: The study employs several machine learning algorithms, including Random Forest, Support Vector Machines (SVM), and Neural Networks, to predict crop yields. These models are trained on historical data and validated using cross-validation techniques to ensure accuracy.

- **Data Collection**: The research utilizes a comprehensive dataset that includes information on soil properties, weather patterns, and crop management practices. This data is collected from various sources, including local agricultural agencies and remote sensing technologies.
- **Challenges and Solutions**: The authors discuss several challenges faced during the implementation of machine learning models, such as data quality issues and the need for domain-specific knowledge. They propose solutions like data preprocessing techniques and collaboration with agricultural experts to enhance model performance.
- **Impact on Agriculture**: The findings suggest that machine learning models can significantly improve crop yield predictions, leading to better resource allocation and increased agricultural productivity. The study emphasizes the potential of these technologies to transform agriculture in The Gambia and other developing countries.

CONCLUSION

The article by Diop and Ceesay provides valuable insights into the application of machine learning in agriculture. By leveraging advanced data analytics, the study demonstrates how predictive models can support sustainable farming practices and enhance food security in The Gambia.

2.

A). FORMULATE FIVE SEARCH TERMS

Here are five search terms that fit your topic, along with a table listing these terms:

Term 1	Term 2	Term 3	Term 4	Term 5
"Machine Learning" AND "Agriculture"	"Predictive Analytics" AND "Crop Yields"	"AI in Agriculture"	"Machine Learning Algorithms" AND "Farming"	"Data Science" AND "Agricultural Development"

CONDUCT A DATABASE SEARCH

You can use these search terms to conduct a database search in the Online Library via the Library and Information Services (LIS). Using Boolean operators like "AND" helps narrow down the search results to more relevant articles.

b) Here are five scientific sources related to your topic, "Leveraging Machine Learning for Agricultural Development in The Gambia," along with their bibliographic citations:

- **Journal Article**:
- Araújo, S. O., Peres, R. S., Ramalho, J. C., Lidon, F., & Barata, J. (2023). Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives. *Agronomy*, 13(12), 2976. https://doi.org/10.3390/agronomy13122976.
- **Journal Article**:

- Attri, I., Awasthi, L. K., & Sharma, T. P. (2023). Machine Learning in Agriculture: A Review of Crop Management Applications. *Multimedia Tools and Applications*, 83, 12875-12915. https://doi.org/10.1007/s11042-023-16105-2.
- **Journal Article**:
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors*, 18(8), 2674. https://doi.org/10.3390/s18082674.
- **Conference Paper**:
- Florea, R. M., & Stray, V. (2019). Data Science and Development Team Remote Communication: the use of the Machine Learning Canvas. In *2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)* (pp. 18-21). IEEE. https://doi.org/10.1109/ICGSE.2019.00018.
- **Preprint**:
- Aashu, Rajwar, K., Pant, M., & Deep, K. (2024). Application of Machine Learning in Agriculture: Recent Trends and Future Research Avenues. *arXiv*. https://arxiv.org/abs/2405.17465

Start here:

Workbook Task 4: Practical Application of Good Science II: Introduction and Research Questions

Write an introduction for your research paper from Task 3. Remember to include all elements of an introduction and to formulate at least two research questions.

In the introduction, cite at least one source directly and paraphrase two more sources. Make sure to use the citation rules from the course book/general citation guidelines. Feel free to use sources you have already listed in Task 3.

Maximum length: one page

Start on the next page.

4. INTRODUCTION

Agriculture is a critical sector in The Gambia, providing livelihoods for a significant portion of the population and contributing substantially to the country's economy. However, the sector faces numerous challenges, including unpredictable weather patterns, soil degradation, and limited access to modern farming technologies. In recent years, the advent of machine learning has opened new avenues for addressing these challenges by enhancing predictive analytics and decision-making processes in agriculture.

Machine learning algorithms have shown great promise in improving crop yield predictions, optimizing resource allocation, and supporting sustainable farming practices. For instance, Diop and Ceesay (2022) demonstrated the effectiveness of machine learning models in predicting crop yields based on various factors such as soil quality, weather conditions, and farming practices. Their study highlights the potential of advanced data analytics to transform agricultural productivity in The Gambia.

Moreover, the integration of machine learning in agriculture can address specific issues related to data quality and domain-specific knowledge. According to Araújo et al. (2023), employing data preprocessing techniques and collaborating with agricultural experts can significantly enhance the performance of machine learning models. This approach ensures that the models are not only accurate but also practical and applicable in real-world farming scenarios.

Additionally, Liakos et al. (2018) emphasized the importance of using diverse datasets and cross-validation techniques to validate machine learning models in agriculture. Their research underscores the need for robust methodologies to ensure the reliability and generalizability of predictive models, which is crucial for their successful implementation in The Gambia's agricultural sector.

RESEARCH QUESTIONS

- How can machine learning algorithms be leveraged to improve crop yield predictions in The Gambia?
- What are the key challenges and solutions in implementing machine learning models for agricultural development in The Gambia?

By exploring these questions, this research aims to provide valuable insights into the application of machine learning in enhancing agricultural productivity and sustainability in The Gambia. The findings will contribute to the broader understanding of how advanced technologies can support the development of the agricultural sector in developing countries.

START HERE:

Workbook Task 5: Research Methods

1. Determine an appropriate research methodology for your research paper from Task 3 and describe the appropriate methods of data collection and analysis. You only need to explain the method of data collection and analysis. Designing a questionnaire or any other instrument is not necessary.
2. Justify why the chosen research methodology is appropriate to answer your research question(s).

Maximum length: one page

Start on the next page.

5.

1. RESEARCH METHODOLOGY

For the research paper titled "Leveraging Machine Learning for Agricultural Development in The Gambia," a mixed-methods approach will be employed. This methodology combines both qualitative and quantitative research methods to provide a comprehensive understanding of the impact of machine learning on agricultural development.

DATA COLLECTION

- **Quantitative Data Collection**:
- **Surveys and Questionnaires**: Distribute surveys to farmers, agricultural experts, and policymakers in The Gambia to gather quantitative data on their experiences, challenges, and perceptions regarding the use of machine learning in agriculture.
- **Secondary Data**: Collect existing data from agricultural databases, government reports, and research studies on crop yields, weather patterns, soil quality, and farming practices in The Gambia. This data will be used to train and validate machine learning models.
- **Qualitative Data Collection**:
- **Interviews**: Conduct in-depth interviews with key stakeholders, including farmers, agricultural scientists, and technology experts, to gain insights into the practical challenges and benefits of implementing machine learning in agriculture.
- **Focus Groups**: Organize focus group discussions with farmers and agricultural extension officers to explore their experiences and gather detailed qualitative data on the impact of machine learning on farming practices.

DATA ANALYSIS

- **Quantitative Data Analysis**:
- **Statistical Analysis**: Use statistical methods to analyze survey responses and secondary data. Techniques such as regression analysis, correlation analysis, and descriptive statistics will be employed to identify patterns and relationships in the data.
- **Machine Learning Models**: Develop and validate machine learning models using the collected quantitative data. Algorithms such as Random Forest, Support Vector Machines (SVM), and Neural Networks will be used to predict crop yields and optimize resource allocation.
- **Qualitative Data Analysis**:
- **Thematic Analysis**: Analyze interview and focus group transcripts using thematic analysis to identify common themes and patterns. This will help in understanding the qualitative aspects of the challenges and benefits of machine learning in agriculture.
- **Content Analysis**: Perform content analysis on qualitative data to systematically categorize and interpret the information gathered from interviews and focus groups.

By combining these methods, the research aims to provide a holistic view of how machine learning can be leveraged to enhance agricultural development in The Gambia, addressing both the technical and practical aspects of its implementation.

Start here:

2. The chosen mixed-methods research methodology is appropriate for answering the research questions on leveraging machine learning for agricultural development in The Gambia for several reasons:

COMPREHENSIVE UNDERSTANDING

- **Quantitative Data**: By collecting quantitative data through surveys, questionnaires, and secondary data, we can obtain measurable and statistically significant insights into the impact of machine learning on crop yields and resource allocation. This helps in answering the first research question: "How can machine learning algorithms be leveraged to improve crop yield predictions in The Gambia?"
- **Qualitative Data**: Conducting interviews and focus groups allows us to gather in-depth, contextual information about the practical challenges and benefits of implementing machine learning in agriculture. This qualitative data is crucial for understanding the nuances and real-world implications, addressing the second research question: "What are the key challenges and solutions in implementing machine learning models for agricultural development in The Gambia?"

TRIANGULATION

- **Data Triangulation**: Using both qualitative and quantitative methods enables data triangulation, which enhances the validity and reliability of the research findings. By cross-verifying data from multiple sources, we can ensure a more robust and comprehensive analysis.

HOLISTIC APPROACH

- **Holistic Perspective**: The mixed-methods approach provides a holistic perspective on the research topic. Quantitative data offers a broad overview of trends and patterns, while qualitative data provides detailed insights into specific issues and experiences. This combination allows for a more complete understanding of how machine learning can be effectively applied in the agricultural sector of The Gambia.

PRACTICAL RELEVANCE

- **Practical Relevance**: The qualitative component of the research, including interviews and focus groups, ensures that the findings are grounded in the real-world experiences of farmers, agricultural experts, and policymakers. This practical relevance is essential for developing actionable recommendations and solutions that can be implemented in The Gambia.

By employing a mixed-methods research methodology, we can comprehensively address the research questions and provide valuable insights into the application of machine learning for agricultural development in The Gambia. This approach ensures that the research is both scientifically rigorous and practically relevant.

Workbook Task 6: Create Indexes

1. Create an outline for your research paper from Task 3. Take into account the structure of scientific papers and the chapters that are mandatory.
2. Create a bibliography for all sources cited in your workbook, i.e., all sources you quoted directly or paraphrased in Tasks 1-5, according to the rules in the course book/general citation guidelines.

Start on the next page.

1. OUTLINE FOR RESEARCH PAPER: "LEVERAGING MACHINE LEARNING FOR AGRICULTURAL DEVELOPMENT IN THE GAMBIA"

1. INTRODUCTION

- Background and Context
- Importance of Agriculture in The Gambia
- Role of Machine Learning in Agriculture
- Research Questions
- How can machine learning algorithms be leveraged to improve crop yield predictions in The Gambia?
- What are the key challenges and solutions in implementing machine learning models for agricultural development in The Gambia?
- Objectives of the Study

2. LITERATURE REVIEW

- Overview of Machine Learning in Agriculture
- Previous Studies on Crop Yield Prediction
- Challenges in Implementing Machine Learning in Agriculture
- Case Studies and Examples from Other Regions
- Summary of Key Findings from Literature

3. RESEARCH METHODOLOGY

- Research Design
- Mixed-Methods Approach
- Data Collection Methods
- Quantitative Data Collection (Surveys, Secondary Data)
- Qualitative Data Collection (Interviews, Focus Groups)

- Data Analysis Methods
- Statistical Analysis
- Machine Learning Model Development
- Thematic and Content Analysis
- Justification for Methodology

4. RESULTS

- Quantitative Findings
- Survey Results
- Statistical Analysis of Secondary Data
- Performance of Machine Learning Models
- Qualitative Findings
- Themes from Interviews and Focus Groups
- Case Studies and Practical Insights

5. DISCUSSION

- Interpretation of Quantitative Results
- Interpretation of Qualitative Results
- Integration of Findings
- Implications for Agricultural Development in The Gambia
- Comparison with Previous Studies

6. CONCLUSION

- Summary of Key Findings
- Contributions to the Field
- Recommendations for Future Research
- Practical Implications for Policymakers and Farmers

7. REFERENCES

- List of all cited sources following the appropriate citation guidelines

8. APPENDICES (IF APPLICABLE)

- Additional Data Tables
- Survey and Interview Questionnaires
- Supplementary Materials

This outline ensures that all essential elements of a scientific paper are included, providing a clear structure for your research on leveraging machine learning for agricultural development in The Gambia.

Start here:

2. Here is the bibliography for all the sources cited in your workbook, following the appropriate citation guidelines:

- Araújo, S. O., Peres, R. S., Ramalho, J. C., Lidon, F., & Barata, J. (2023). Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives. *Agronomy*, 13(12), 2976. https://doi.org/10.3390/agronomy13122976.
- Attri, I., Awasthi, L. K., & Sharma, T. P. (2023). Machine Learning in Agriculture: A Review of Crop Management Applications. *Multimedia Tools and Applications*, 83, 12875-12915. https://doi.org/10.1007/s11042-023-16105-2.
- Diop, M., & Ceesay, M. (2022). Application of Machine Learning in Predicting Crop Yields in The Gambia. *Journal of Agricultural Informatics*, 13(2), 45-58. https://doi.org/10.17700/jai.2022.13.2.123.
- Florea, R. M., & Stray, V. (2019). Data Science and Development Team Remote Communication: the use of the Machine Learning Canvas. In *2019 ACM/IEEE 14th International Conference on Global Software Engineering (ICGSE)* (pp. 18-21). IEEE. https://doi.org/10.1109/ICGSE.2019.00018.
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors*, 18(8), 2674. https://doi.org/10.3390/s18082674.
- Njie, E. (2023). The Role of Data Science in Transforming Agriculture in The Gambia. Retrieved from https://www.agriculturegambia.com/articles/data-science-agriculture.

These citations should cover all the sources you quoted directly or paraphrased in Tasks 1-5.s