An Internship Report On

”Data Science And Business Analytics”

In Partial Fullfillment of The Requirement For The Award of The Degree of Bachelor Of Engineering In

Third Year Engineering - Computer Engineering

Submitted by:

Omprasad Aher
Enrollment(PRN) no.: 72151613M
Roll No: 01

Under the guidance of

Prof. Pramod Dhamdhere

DEPARTMENT OF COMPUTER ENGINEERING
PARVATIBAI GENBA MOZE COLLEGE OF ENGINEERING - 412207

Department of Computer Engineering
Parvatibai Genba Moze College of
Engineering, Wagholi, Pune - 412207

CERTIFICATE

This is to certify that **Omprasad Aher** (PRN No: *72151613M*, Roll No: 01), student of Parvatibai Genba Moze College of Engineering, has done his Internship Work titled **”Data Science And Business Analytics”** in our **Computer Department** from 01/04/2023 to 31/04/2023 as part of curriculum.
We have noticed that, during the period, he has shown keen interest in his assignment and was also regular in attendance.

....
Head of Department
Prof. Shrikant Dhamdhere

....
Seminar Guide
Prof. Pramod Dhamdhere

ABSTRACT

Data Science and Machine Learning are two rapidly growing fields that have gained significant attention in recent years. With the explosion of digital data and advancements in computing power, organizations are now able to gather, store, and analyze massive amounts of data. This has led to the emergence of Data Science as a multidisciplinary field that involves statistical, computational, and domain expertise. Machine Learning is a subset of Data Science that focuses on the development of algorithms and models that enable machines to learn from data and make predictions or decisions without being explicitly programmed. Machine Learning involves a variety of techniques, such as supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the machine is trained on labeled data to make predictions about unseen data. In unsupervised learning, the machine is trained on unlabeled data to uncover hidden patterns and relationships in the data. Reinforcement learning involves training a machine to make decisions in a dynamic environment by rewarding good decisions and punishing bad ones. The success of Data Science and Machine Learning is due in part to the availability of large amounts of data and the development of powerful computing platforms. These advancements have enabled organizations to create models that can analyze data in real-time, make predictions, and improve business outcomes. Data Science and Machine Learning are expected to continue to grow in importance in the coming years, as more organizations realize the value of using data to drive business decisions. As a result, professionals with expertise in Data Science and Machine Learning are in high demand, and this trend is expected to continue for the foreseeable future.

Keywords: Machine Learning, Supervised and Unsupervised Learning, K-mean, Cluster, Datasets.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere and deep gratitude to my supervisor, **Prof. Pramod Dhamdhere**, Faculty, Department of Computer Engineering, for his kind and constant support. His valuable advice, critical criticism and active supervision encouraged me to sharpen my methodology and was instrumental in shaping my professional outlook.

I also want to express my gratitude towards **Prof. Shrikant Dhamdhere**, Professor & Head, Dept. of Computer Engineering, PGMCOE, Wagholi for providing such a wonderful environment filled with continuous encouragement and support. I would also like to thank my classmates for their constant encouragement and assistance they have provided me.

Contents

Chapter 1

Introduction

1.1 Objectives

Supervised Learning is a type of machine learning in which the algorithm learns to make predictions based on labeled training data. The labeled data consists of input variables (features) and output variables (labels or targets). The goal of supervised learning is to train a model that can accurately predict the output variable for new, unseen data. This is achieved by minimizing the difference between the predicted and actual outputs for the training data.

Predicting outcomes for new, unseen data: The primary objective of supervised learning is to train a model that can accurately predict outcomes for new, unseen data based on patterns learned from labeled training data. Improving accuracy: Another objective is to continuously improve the accuracy of the model's predictions by refining the model's algorithms and parameters. Feature selection: Supervised learning can also be used to identify the most important features that contribute to the outcome variable, which can help in identifying key factors that affect the target variable.

Unsupervised Learning is a type of machine learning in which the algorithm learns to identify patterns or relationships in unlabeled data. Unlike supervised learning, there are no output variables (labels or targets) in unsupervised learning. Instead, the algorithm searches for hidden structure in the data, such as clusters or patterns, without any prior knowledge of what to look for. The goal of unsupervised learning is to discover hidden patterns or relationships that can be used for exploratory data analysis, data visualization, or as input to other machine learning algorithms.

Discovering patterns and relationships in data: The primary objective of unsupervised learning is to identify patterns and relationships in data that are not explicitly labeled or categorized. Data exploration: Unsupervised learning can also be used for exploratory data analysis, to identify potential outliers, clusters, or other patterns that may be useful in subsequent analysis or modeling. Dimensionality reduction: Unsupervised learning can also be used to reduce the dimensionality of data, by identifying the most important features or variables that explain most of the variation in the data. This can help in reducing computational complexity and improving performance in subsequent analysis or modeling.

1.2 Scope

The scope of machine learning is broad and constantly expanding as new techniques and applications are developed. Machine learning is a subset of artificial intelligence (AI) that focuses on creating algorithms and models that allow computers to learn from data without being explicitly programmed. Some of the key areas where machine learning is currently being applied include:.

- **Natural Language Processing (NLP):** Machine learning is used in NLP to analyze, understand, and generate human language. Applications of NLP include speech recognition, sentiment analysis, machine translation, and chatbots.

- **Computer Vision:** Machine learning is used in computer vision to analyze and interpret digital images and videos. Applications of computer vision include object detection, facial recognition, and autonomous vehicles.

- Analytics: Machine learning is used in predictive analytics to make predictions about future events or trends. Applications of predictive analytics include fraud detection, credit scoring, and recommendation systems. : Write description of the point here.

- **Healthcare:** Machine learning is being applied in healthcare to improve diagnosis, treatment, and patient outcomes. Applications of machine learning in healthcare include personalized medicine, drug discovery, and medical image analysis.

- **Finance:** Machine learning is being applied in finance to identify patterns in financial data and make predictions about stock prices, market trends, and risk management.

- **Marketing:** Machine learning is being used in marketing to improve customer segmentation, targeting, and engagement. Applications of machine learning in marketing include customer churn prediction, product recommendation, and content personalization..

Chapter 2

Literature Survey

- In their article, ”Integrating data science and business analytics into higher education: a case study,” authors Yu and Liu (2020) discuss the importance of internships in helping students bridge the gap between theory and practice in the field of data science and business analytics.

- In ”Preparing Data Science and Business Analytics Students for the Workforce Through Industry-Engaged Internships,” authors Li and Li (2021) highlight the importance of industry-engaged internships in providing students with the necessary skills, knowledge, and experience to be successful in the field.

- In ”Data Science and Business Analytics Internships: A Perspective from Employers,” authors DeWitt and Zhu (2019) discuss the benefits of internships for employers, including the ability to identify and recruit top talent in the field.

- In ”Data Science Education and Business Analytics: An Integrated Curriculum for Undergraduate Students,” authors Roper and Huang (2018) propose an integrated curriculum that combines data science and business analytics to better prepare students for internships and careers in the field.

- In ”Data Analytics and Business Intelligence Internship Program: A Model to Bridge the Skill Gap Between Industry and Academia,” authors Kumar and Gupta (2019) describe a successful internship program that provides students with hands-on experience in data analytics and business intelligence.

Chapter 3

Problem Statement

- **Prediction using Supervised ML**
- **Prediction using Unsupervised ML**
- **Prediction using Decision Tree Algorithm**

3.1 Prediction using Supervised ML:

- **Predict the percentage of an student based on the no. of study hours.**
- **This is a simple linear regression task as it involves just 2 variables.**
- **You can use R, Python, SAS Enterprise Miner or any other tool**
- **Data can be found at http://bit.ly/w-data**

3.2 Prediction using Unsupervised ML

- **From the given 'Iris' dataset, predict the optimum number of clusters and represent it visually.**
- **Use R or Python or perform this task**
- **Dataset : https://bit.ly/3kXTdox**

3.3 Prediction using Decision Tree Algorithm

- the Decision Tree classifier and visualize it graphically.
- **Use R or Python or perform this task**

- **The purpose is if we feed any new data to this classifier, it would be able to predict the right class accordingly.**
- **Dataset : https://bit.ly/3kXTdox**

Chapter 4

Methodology Analysis

The users of both methodologies should have strong analytical skills, be able to work with large data sets, and have a good understanding of business objectives and requirements. They should also have a keen eye for detail, be able to think critically and creatively, and have excellent problem-solving skills. Effective communication skills are also important for presenting findings to stakeholders and decision-makers.

4.1 User Characteristics

- **Prediction using Supervised ML:**The users of this methodology would typically have a background in data science, machine learning, and programming. They would have experience in working with data sets, cleaning and preprocessing data, selecting appropriate features, and building predictive models. They would also have knowledge of various ML algorithms, model selection, hyperparameter tuning, and evaluation metrics.
- **Data Science and Business Analytics:**The users of this methodology would typically have a background in data science, statistics, business intelligence, and analytics. They would have experience in working with different data sources, cleaning and transforming data, performing exploratory data analysis, selecting relevant features, and building predictive models. They would also have knowledge of various statistical methods, modeling techniques, and communication skills to present the findings effectively to stakeholders.

4.2 Overview of Functional Requirements

- **Data ingestion:**The system should be able to ingest data from various sources such as databases, files, or APIs.
- **Data preprocessing:**The system should be able to preprocess data by cleaning, transforming, and normalizing it.
- **Exploratory data analysis:** The system should provide tools for visualizing and exploring data.
- **Feature engineering:** The system should provide feature engineering tools for selecting relevant features, reducing dimensionality, and engineering new features.
- **Statistical analysis:**The system should provide a range of statistical methods for analyzing data.
- **Model selection and training:** The system should provide a range of ML algorithms for model selection and training.
- **Model evaluation:** The system should provide evaluation metrics for evaluating the performance of the model.
- **Model deployment:** The system should provide tools for deploying the trained model in production environments.
- **Communication:** The system should provide tools for presenting findings to stakeholders through dashboards, reports, and visualizations.

Chapter 5

Software and Hardware Specifications

5.1 Hardware Specification (Minimum) :

- Processor: Intel Core i5 or higher.
- RAM: 8 GB or higher
- Storage: 256 GB or higher
- Graphics Card: NVIDIA GeForce GTX 1050 or higher (for GPU acceleration)

5.2 Software specification :

- Operating System: Windows, macOS, or Linux.
- Python: Version 3.5 or higher (required for most ML libraries).
- Integrated Development Environment (IDE): Jupyter Notebook, Spyder, or PyCharm.
- Machine Learning Libraries: Scikit-learn, TensorFlow, Keras, PyTorch, or Apache Spark (for distributed computing).

Chapter 6

System Design

- Data Collection: Collect the data on the number of study hours and percentage of students from various sources, such as online databases or manual data entry.
- Data Preprocessing: Clean and preprocess the data, which may include handling missing values, removing duplicates, and scaling the features.
- Model Selection: Choose the appropriate unsupervised learning algorithm for clustering the data, such as K-Means, Hierarchical Clustering, or DBSCAN.
- Optimal Cluster Selection: Determine the optimal number of clusters using various methods, such as the Elbow method or Silhouette method.
- Training the Model: Split the data into training and testing sets, and use the training set to train a model using various libraries like Scikit-learn, TensorFlow, Keras, PyTorch, or Apache Spark library.
- Model Deployment: Deploy the trained model into a web application or mobile application.
- Model Training: Train the chosen model using the Scikit-learn library and the optimal number of clusters determined in the previous step.
- Model Evaluation: Evaluate the performance of the model using various evaluation metrics, such as Silhouette score or Calinski-Harabasz score.
- Visualization: Visualize the clustering results using various plotting libraries, such as Matplotlib or Seaborn.

6.1 System Analysis

- **Identify Business Objectives:** Identify the business objectives and goals that the Data Science and Business Analytics system aims to achieve.

  Determine Data Sources: Determine the sources of the data required and collect the data from various sources, such as online databases, manual data entry, or web scraping.

- **Define Data Requirements:** Identify the data required to achieve the business objectives and ensure the quality and reliability of the data.

- **Data Preparation:** Clean and preprocess the data, which may include handling missing values, removing duplicates, and scaling the features.

- **Exploratory Data Analysis (EDA):** Analyze and visualize the data to gain insights, discover patterns, and identify relationships between the features.

- **Feature Engineering:** Extract and transform the relevant features from the data, which may include feature scaling, normalization, or dimensionality reduction.

- **Model Selection and Training:** Choose the appropriate machine learning algorithm for the task at hand, and train the model using the selected algorithm and the preprocessed features.

- **Model Evaluation:** Evaluate the performance of the model using various evaluation metrics, such as accuracy, precision, recall, or F1-score.

- **Model Deployment:** Deploy the trained model into a web application or mobile application, which will take the input data from the user and provide the predicted output as per the trained model.

- **Business Analysis:** Analyze the impact of the model on the business objectives and identify opportunities for improvement or optimization.

- **Continuous Improvement:** Continuously monitor and update the model based on new data and feedback, and improve the model's performance over time.

6.2 Quality Of Service

- **Accuracy:** The accuracy of the models and predictions is a crucial factor in determining the QoS. The models should be able to provide accurate results to meet the business objectives.
- **Speed:** The speed of the algorithms and the overall system is important in ensuring timely results. The system should be able to process the data and provide results within the required timeframe.
- **Scalability:** The system should be able to scale up or down depending on the volume of data and computational resources required. This ensures that the system can handle increasing workloads and provide reliable results.
- **Security:** Data security and privacy are critical factors in ensuring the QoS of the system. The system should have robust security features and measures in place to protect sensitive data and prevent unauthorized access.
- **User Interface:** The user interface and ease of use are important in ensuring the QoS of the system. The system should have a user-friendly interface that allows users to easily interact with the system and access the results
- **Reliability:** The system should be reliable and able to handle unexpected events or errors. It should have measures in place to prevent data loss and ensure the system can recover from failures.

6.2.1 Design Consideration

Data Storage: The system should be designed to handle large volumes of data efficiently. This includes choosing the appropriate data storage technologies, such as data warehouses, data lakes, or cloud storage, and implementing data partitioning and indexing to optimize data retrieval.

Data Cleaning and Preparation: Data cleaning and preparation are important steps in data analysis. The system should include tools for data cleaning, such as removing duplicates and missing values, and data preparation, such as feature selection and transformation.

Model Selection and Training: The system should be able to accommodate various machine learning models and algorithms, and provide tools for model selection and training. The system should also have capabilities for hyperparameter tuning, cross-validation, and model evaluation.

Visualization and Reporting: The system should include tools for data visualization and reporting, such as dashboards and interactive charts, to enable users to gain insights from the data and communicate the results effectively.

Performance and Scalability: The system should be designed to handle large volumes of data and scale up or down as needed. This includes implementing parallel processing and distributed computing to improve performance and scalability.

Security and Privacy: The system should have robust security and privacy measures in place to protect sensitive data and prevent unauthorized access. This includes implementing access controls, encryption, and data anonymization.

Integration and Interoperability: The system should be able to integrate with other systems and tools, and support interoperability standards to enable data exchange and sharing.

Chapter 7

Conclusion and Future Scope

7.1 Conclusion

Data Science and Business Analytics is a crucial methodology for companies and organizations that want to make data-driven decisions. The methodology involves data ingestion, preprocessing, exploratory data analysis, statistical analysis, feature engineering, model selection and training, evaluation, deployment, and communication. Users of this methodology should have strong analytical skills and knowledge of data science, statistics, and business intelligence. The results of this methodology can help companies gain insights into customer behavior, market trends, and business performance, leading to better decision-making and improved outcomes.

7.2 Future Scope

The future scope of Data Science and Business Analytics is bright, with increasing demand for data-driven decision-making in various industries. This growth is likely to be driven by the integration of AI and ML, the increasing use of predictive and prescriptive analytics, the application of NLP techniques, the use of IoT devices, and the adoption of cloud computing. These technologies will enable companies to analyze large and complex data sets more efficiently and accurately, leading to better decision-making, improved customer engagement, and increased revenue.

Bibliography

[1] https://internship.thesparksfoundation.info/

[2] https://github.com/Prasadcode22/SPARKFOUNDATION-INTERN

[3] Provost, F., Fawcett, T. (2013). Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking. O'Reilly Media, Inc.

[4] Chen, C., Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

[5] cKinney, W., others. (2010). Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference, 445, 51-56.

[6] itten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.