|  |  |
| --- | --- |
| **Chapters** | **Table of Contents** |
|  | **About the Company** |
| * + 1.1 Company Profile   + 1.2 Industry Focus |
|  | **About the Domain** |
| * + 2.1 Artificial Intelligence   + 2.2 Data Science   + 2.3 Applications   + 2.4 Future Scope |
|  | **Tools and Technologies Used** |
| * + 3.1 Python   + 3.2 Jupyter Notebook   + 3.3 Libraries (Pandas, NumPy, Scikit-learn, etc.) |
|  | **Tasks Performed** |
| * + 4.1 Week-wise Summary   + 4.2 Glimpse of Project Execution |
|  | **Mini Projects** |
| * + 5.1 Handwritten Digits Recognition   + 5.2 FIFA 20 Player Analysis   + 5.3 PUBG Match Winner Prediction   + 5.4 Heart Disease Prediction   + 5.5 Outputs |
|  | **Conclusion** |

# CHAPTER 1:

# ABOUT THE COMPANY

**1.1 COMPANY PROFILE**

| **Aspect** | **Details** |
| --- | --- |
| **Organization Name** | Rubixe™ |
| **Organization Logo** |  |
| **Established** | 2015 |
| **Type** | Private Technology Company |
| **Ownership** | Privately Held |
| **Headquarters** | Bangalore, Karnataka, India |
| **Key Functions** | - Artificial Intelligence Solutions - Business Automation - Data Science & Analytics Consulting - AI-based Product Development |
| **Product Specialization** | - Predictive Analytics - AI-Driven Business Intelligence - Custom ML and NLP Tools |
| **Key Domains Served** | - Retail - Finance - Healthcare - Manufacturing |
| **Notable Projects** | - AI Chatbots for customer support - Fraud detection systems - Personalized recommendation engines |
| **Global Role** | - AI and automation consulting for clients across India, the Middle East, and parts of Europe |
| **Vision** | To deliver intelligent and adaptive technology that transforms businesses through innovation. |
| **Mission** | To enable businesses to leverage the power of disruptive technologies like AI, Machine Learning, and IoT to drive digital transformation. |

|  |
| --- |
|  |
| Fig 1.1.1 Company Building |

|  |
| --- |
|  |
| Fig 1.1.2 Company Lobby |

|  |
| --- |
|  |
| Fig 1.1.3 Company Break Room |

|  |
| --- |
|  |
| Fig 1.1.4 Company Work Space |

|  |
| --- |
|  |
| Fig 1.1.5 Company Work Space |

## 1.2 Industry Focus

Rubixe serves clients across a broad spectrum of industries, reflecting the versatility of its AI-driven solutions. The company’s industry focus is not limited to one sector; instead, it includes major domains such as **Healthcare**, **Finance and Banking**, **Information Technology**, **Retail**, **Manufacturing**, **Education**, **Agriculture**, and more​[goodfirms.co](https://www.goodfirms.co/company/rubixe#:~:text=%2A%20Healthcare%20%26%20Medical%20,5)​[goodfirms.co](https://www.goodfirms.co/company/rubixe#:~:text=%2A%20E,5). For instance, in healthcare, Rubixe might develop predictive analytics for patient data, whereas in finance it could implement AI for fraud detection or algorithmic trading support. In retail and e-commerce, Rubixe’s solutions could involve personalized recommendation systems and inventory optimization, while in manufacturing, they might focus on predictive maintenance and automation of assembly lines. This wide industry coverage (from **startups** to **large enterprises**) indicates Rubixe’s adaptive approach in applying AI to different contexts. The company’s team includes domain experts who understand the unique challenges of each sector, ensuring that AI solutions are effective and relevant to the client’s field. By maintaining a diverse industry focus, Rubixe not only expands its market reach but also enriches its expertise, as lessons learned in one domain can often inspire innovative applications in another.

# CHAPTER 2:

# ABOUT THE DOMAIN

This chapter provides an overview of the technical domain central to the internship: **Artificial Intelligence and Data Science**. It covers the fundamental concepts of AI and data science, discusses some of their real-world applications, and reflects on the future scope of these fields. Understanding the domain is crucial, as the internship work involved applying AI and data science techniques to solve practical problems in various projects. The following sections define Artificial Intelligence and Data Science, outline key application areas, and consider emerging trends and future developments.

## 2.1 Artificial Intelligence

**Artificial Intelligence (AI)** refers to the simulation of human intelligence processes by machines, especially computer systems​[techtarget.com](https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence#:~:text=Artificial%20intelligence%20,recognition%20%20and%20%2039). In essence, AI enables computers to perform tasks that would typically require human intelligence, such as learning from experience, understanding language, recognizing patterns, and making decisions. Core subfields of AI include **machine learning** (where algorithms improve through data exposure), **natural language processing** (enabling machines to understand and generate human language), **computer vision** (interpreting visual information from the world), and **expert systems** (which use logic and rules to mimic human experts)​[techtarget.com](https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence#:~:text=by%20machines%2C%20especially%20computer%20systems,speech%20recognition%20and%20machine%20vision). AI systems can be either narrow AI (designed for a specific task, like a speech recognition assistant or a chess-playing program) or general AI (an as-yet unachieved state where a machine possesses broad intelligence comparable to a human). Over the years, AI techniques have advanced significantly, enabling the development of systems that can **learn**, **perceive**, and **adapt** to new information. During the internship projects, AI concepts were applied, for example, in training models to recognize handwriting and predict outcomes, demonstrating the practical use of algorithms that enable machine intelligence.

## 2.2 Data Science

**Data Science** is an interdisciplinary field focused on extracting knowledge and insights from large data sets and applying those insights to solve complex problems​[en.wikipedia.org](https://en.wikipedia.org/wiki/Data_science#:~:text=Data%20science%20is%20an%20interdisciplinary,11). It blends techniques from statistics, computer science, and domain expertise to collect, process, and analyze data. Key components of data science include data collection (gathering relevant data from various sources), data wrangling or preprocessing (cleaning and transforming raw data into a usable format), exploratory data analysis (using statistical methods and visualization to discover patterns or anomalies in the data), and modeling (building predictive or descriptive models using machine learning or statistical algorithms). Data science also involves validating models and interpreting results to inform decision-making. A data scientist must not only build accurate models but also understand the context of the data—this is where domain knowledge comes into play. In this internship, data science practices were central: each project began with understanding and exploring the dataset at hand (whether images of digits, football player statistics, game outcomes, or medical records) and then applying appropriate modeling techniques to draw conclusions or make predictions. The interdisciplinary nature of data science was evident, as success required a mix of programming skills (for implementation in Python), mathematical understanding (for algorithm correctness), and contextual knowledge (to make sense of results in domains like sports or healthcare).

## 2.3 Applications

AI and data science have a wide array of applications across different industries, many of which are becoming integral to how modern organizations operate. A few notable application areas include:

* **Healthcare:** AI and data science are used for diagnosing diseases, predicting patient outcomes, and personalizing treatment plans. For example, machine learning models can analyze medical images (X-rays, MRIs) to detect anomalies or predict the likelihood of conditions such as cancer or heart disease. In this internship’s Heart Disease Prediction project, the application of data science in healthcare was evident, as a model was developed to predict cardiac risk based on patient data. Such models can assist doctors in early detection of high-risk patients, potentially saving lives by enabling proactive care.
* **Finance:** In the financial industry, AI algorithms detect fraudulent transactions, assess credit risk, automate trading strategies, and provide customer service through chatbots. Data science enables banks and financial firms to analyze vast quantities of transactions and client data to find patterns indicative of fraud or to segment customers for tailored services. The ability of AI to learn from historical financial data helps in making predictions about market trends or default probabilities, thereby supporting more informed and efficient financial decision-making.
* **Retail and E-commerce:** Businesses leverage AI for recommendation systems (suggesting products to customers based on their browsing and purchase history), inventory management, and demand forecasting. Data science techniques analyze customer behavior data to optimize pricing strategies and marketing campaigns. For instance, an AI system can predict which products are likely to be popular in the next season and ensure they are stocked appropriately.
* **Transportation and Automotive:** A high-profile application of AI is in autonomous vehicles (self-driving cars) where computer vision and deep learning are used to navigate roads and make real-time decisions. Additionally, AI optimizes logistics and supply chain operations by predicting the fastest delivery routes and minimizing fuel consumption.
* **Manufacturing:** AI-driven predictive maintenance systems analyze sensor data from machinery to predict failures before they occur, thus avoiding costly downtime. Robotics powered by AI also enable automation of complex manufacturing tasks with precision and adaptability.
* **Entertainment and Sports Analytics:** Streaming services and media companies use data science to analyze viewer preferences and recommend content. In sports, AI and data analysis are used to evaluate player performance and devise game strategies. One of the internship projects, the FIFA 20 Player Analysis, mirrors real-world sports analytics where player data is examined to glean insights on performance and value.

These examples barely scratch the surface; AI and data science are also transforming education (through personalized learning platforms), agriculture (through crop and soil monitoring), and government (through smart cities and policy planning), among others. In every application, the goal is to harness data to make better decisions, automate tasks, and create intelligent behavior in software or machines. The projects completed during the internship each correspond to a real-world application of AI: computer vision in the digit recognition task, sports analytics in the FIFA data task, game outcome prediction in the PUBG task, and medical risk prediction in the heart disease task. This demonstrates the versatility of AI and data science techniques across domains.

## 2.4 Future Scope

The future of Artificial Intelligence and Data Science is both exciting and expansive. As computational power grows and algorithms become more sophisticated, AI systems are expected to achieve even higher levels of performance and autonomy. One emerging trend is the development of more generalized AI models (such as **large language models** and advanced neural networks) that can perform a wider range of tasks and learn with less human supervision. These models are increasingly capable of understanding context and generating human-like responses, which could revolutionize fields like customer service, content creation, and beyond.

In the realm of data science, the volume of data generated globally is increasing exponentially (often referred to as the data deluge). This creates a growing opportunity to extract insights, but also necessitates advancements in **big data** technologies, cloud computing, and data engineering to handle and process data efficiently. The future will likely see data science integrated even more into business decision processes, with real-time analytics and dashboards guiding strategic and operational choices in organizations.

Another important aspect of the future of AI is its integration with IoT (Internet of Things) and edge computing. As more devices become smart and interconnected, AI algorithms will be deployed on distributed devices (from smartphones to industrial sensors) to provide instant insights and automation at the source of data. This could enable smarter homes, smarter cities, and responsive industrial systems.

In terms of societal impact, AI and data science are poised to significantly influence the job market and skill requirements. There is an increasing demand for professionals skilled in AI/ML, and educational curricula are evolving to include these topics from early stages. However, the rise of AI also brings challenges that will shape its future: **ethical considerations** and **governance**. Issues such as data privacy, algorithmic bias, and transparency of AI decisions are under scrutiny. Future AI systems will need to be developed with fairness and accountability in mind, leading to the growth of fields like AI ethics and explainable AI.

Overall, the scope of AI and data science is set to expand into virtually every field. Continuous research and development are making AI solutions more powerful and accessible. For a practitioner like an intern entering this field, it means the learning is ongoing—new tools, techniques, and best practices emerge regularly. The projects done during the internship provide a foundation, but staying updated will be crucial. In the near future, one can expect more automation of routine tasks, more accurate predictive models improving decision-making, and AI tackling complex problems (like climate modeling or drug discovery) that were previously beyond reach. The trajectory suggests that AI and data science will remain at the forefront of technological innovation, driving progress in the coming years.

# CHAPTER 3:

# TOOLS AND TECHNOLOGIES USED

A variety of software tools, programming platforms, and libraries were used throughout the internship to accomplish the tasks and projects. As the internship work centered on data analysis and machine learning, the tools chosen are standard in the data science community. This chapter describes the main tools and technologies employed, namely the Python programming language, the Jupyter Notebook environment, and several important Python libraries such as Pandas, NumPy, Scikit-learn, etc. Each of these played a critical role in enabling the intern to implement solutions efficiently and effectively.

## 3.1 Python

|  |
| --- |
| Image result for python interface |
| Fig 3.1.1 Python interface |

**Python** was the primary programming language used during the internship. Python is a high-level, interpreted language known for its simplicity and readability, which makes it particularly well-suited for data science and AI projects. One of the key advantages of Python is its extensive collection of libraries and frameworks for scientific computing and machine learning (for example, Pandas for data manipulation, or TensorFlow/PyTorch for deep learning). Python’s syntax is concise and clear, allowing developers to prototype ideas quickly with relatively few lines of code. During the internship projects, Python was used for tasks such as loading and preprocessing datasets, implementing machine learning algorithms, and evaluating model performance. The language’s interactive nature (especially when used in notebooks) helped in iteratively developing solutions: the intern could write a portion of code, run it, inspect the output, and adjust as needed in a rapid development cycle. Additionally, Python’s strong community support and documentation meant that any issues encountered could be resolved by referencing existing resources. Overall, Python’s role was foundational, providing the programming backbone for all analyses and experiments conducted in the internship.

## 3.2 Jupyter Notebook

|  |
| --- |
| Getting started with Jupyter Notebooks |
| Fig 3.2.1 Jupyter interface |

The **Jupyter Notebook** was the primary development environment for the projects. Jupyter Notebook is an open-source web-based interactive platform that allows developers to combine live code, visualizations, and narrative text in a single document. This format was extremely useful for the internship, as it facilitated a literate programming approach: one can execute code step by step and document the thought process alongside the results. Each project was organized in a dedicated Jupyter notebook, as per the internship requirements. Using notebooks, the intern performed data exploration by writing code to, for example, display the first few rows of a dataset or plot distributions, and then immediately viewed the results (tables, charts) inline. The notebooks also made it convenient to tune machine learning models and record observations; for instance, the intern could run multiple training experiments in sequence and compare outcomes all within one notebook environment. Another advantage is the ease of presentation—completed notebooks with outputs and markdown explanations can serve as final reports or records of the work. This was helpful for sharing progress with mentors and for final submission, as the notebook contained both the code and the interpretation of results (including any challenges faced and how they were addressed). The interactive and incremental nature of Jupyter notebooks significantly enhanced productivity and clarity during the internship.

## 3.3 Libraries (Pandas, NumPy, Scikit-learn, etc.)

Several specialized Python libraries were utilized to carry out data science tasks efficiently. The major libraries and frameworks used during the internship include:

* **Pandas:** This library was used for data manipulation and analysis. Pandas provides the DataFrame structure, which is excellent for handling tabular data (similar to spreadsheets or SQL tables). With Pandas, the intern could easily load datasets (from CSV files, for example), clean data (handle missing values, filter rows, create new columns), and perform aggregations or summary statistics. Throughout the projects, Pandas was essential for preparing data before feeding it into machine learning models—such as selecting relevant features for the FIFA player data or merging information in the PUBG match data.
* **NumPy:** NumPy is the fundamental package for scientific computing with Python, providing support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays. In the context of this internship, NumPy was frequently used under the hood (many other libraries like Pandas and Scikit-learn are built on NumPy arrays for performance). Directly, it was used for tasks such as numerical computations, generating random numbers (e.g., splitting data into random train/test sets before using library functions), and implementing any algorithmic logic that required array manipulations. NumPy’s efficiency in handling computations made operations on large datasets (like thousands of game records or image pixel arrays) feasible and fast.
* **Scikit-learn:** This is a powerful machine learning library that was central to building and evaluating models for the projects. Scikit-learn provides a consistent interface for a wide range of machine learning algorithms—for classification, regression, clustering, dimensionality reduction, etc. During the internship, Scikit-learn was used to implement models such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), decision trees, random forests, logistic regression, and clustering algorithms like K-Means. It also offers utilities for splitting data, cross-validation, and computing evaluation metrics. For example, in the Handwritten Digits project, Scikit-learn’s SVM and KNN classifiers were used to recognize digits, and accuracy scores were obtained easily through its API. Similarly, in the Heart Disease project, a logistic regression model (from Scikit-learn) might have been used to predict disease presence, along with metrics like precision and recall to gauge performance. The library’s reliability and ease of use significantly streamlined the experimentation with different algorithms.
* **Matplotlib and Seaborn:** These libraries were used for data visualization. Matplotlib is a plotting library that allows the creation of static, interactive, and animated visualizations in Python. Seaborn is built on Matplotlib and provides a higher-level interface for drawing attractive statistical graphics. Visualizing data was a critical part of the internship projects, especially during exploratory analysis and result interpretation. For instance, using Matplotlib/Seaborn, the intern created histograms of player ages, scatter plots of player rating versus age for the FIFA analysis, and correlation heatmaps of features in the Heart Disease dataset. These visuals helped in understanding data distributions, identifying trends, and presenting findings.
* **TensorFlow/Keras (and other frameworks, if applicable):** In some cases, more advanced frameworks like Keras (which uses TensorFlow as a backend) were utilized for building neural network models. This was particularly relevant to the Handwritten Digits Recognition project, where a simple neural network (multilayer perceptron) was trained to improve accuracy on classifying digit images. Keras provides a user-friendly way to define and train deep learning models. While the internship primarily focused on classical machine learning via Scikit-learn, the use of Keras for the digit classification task introduced the intern to the basics of deep learning (such as constructing neural network layers and using an optimizer to adjust weights based on training data).

In summary, these libraries formed the toolkit that enabled effective handling of all aspects of the projects—from reading and exploring data (Pandas, NumPy) to building predictive models (Scikit-learn, TensorFlow) and visualizing results (Matplotlib, Seaborn). Mastery of these tools is fundamental for any data science project, and the internship provided ample opportunity to apply and strengthen skills in using each of them.

# CHAPTER 4:

# TASKS PERFORMED

The internship was structured over 15 weeks, and each week had specific objectives and deliverables. This section provides a **week-wise summary** of the tasks performed, the learning activities, and the progress made throughout the internship. The work gradually progressed from introductory orientation and training in the initial weeks to more complex project work in the later weeks. By breaking down the experience week by week, it becomes clear how the intern’s responsibilities evolved and how the mini projects were distributed across the internship timeline. The summary below highlights key activities and accomplishments for each week.

**4.1 Week-wise Summary**

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 1** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 17-02-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – exploring MNIST dataset and loading into notebook | Consistently delivered tasks for week 1, with initiative and attention to detail. |
| **2** | 18-02-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – preprocessing images and visualizing data | Consistently delivered tasks for week 1, with initiative and attention to detail. |
| **3** | 19-02-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – applying basic classifiers like KNN and SVM | Consistently delivered tasks for week 1, with initiative and attention to detail. |
| **4** | 20-02-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – implementing neural networks | Consistently delivered tasks for week 1, with initiative and attention to detail. |
| **5** | 21-02-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – comparing models and selecting best performer | Consistently delivered tasks for week 1, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Prof. Reeba Rani Internal Guide** | **Prof. Ashwini M Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 2** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 24-02-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – exploring MNIST dataset and loading into notebook | Consistently delivered tasks for week 2, with initiative and attention to detail. |
| **2** | 25-02-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – preprocessing images and visualizing data | Consistently delivered tasks for week 2, with initiative and attention to detail. |
| **3** | 26-02-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – applying basic classifiers like KNN and SVM | Consistently delivered tasks for week 2, with initiative and attention to detail. |
| **4** | 27-02-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – implementing neural networks | Consistently delivered tasks for week 2, with initiative and attention to detail. |
| **5** | 28-02-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – comparing models and selecting best performer | Consistently delivered tasks for week 2, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 3** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 03-03-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – exploring MNIST dataset and loading into notebook | Consistently delivered tasks for week 3, with initiative and attention to detail. |
| **2** | 04-03-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – preprocessing images and visualizing data | Consistently delivered tasks for week 3, with initiative and attention to detail. |
| **3** | 05-03-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – applying basic classifiers like KNN and SVM | Consistently delivered tasks for week 3, with initiative and attention to detail. |
| **4** | 06-03-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – implementing neural networks | Consistently delivered tasks for week 3, with initiative and attention to detail. |
| **5** | 07-03-2025 | 10:00 AM | 05:00 PM | Worked on handwritten digits recognition – comparing models and selecting best performer | Consistently delivered tasks for week 3, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Prof. Masrath Begum Internal Guide** | **Prof. vimla K  Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 4** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 10-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – exploring FIFA player dataset | Consistently delivered tasks for week 4, with initiative and attention to detail. |
| **2** | 11-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – analyzing player attributes and skills | Consistently delivered tasks for week 4, with initiative and attention to detail. |
| **3** | 12-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – clustering based on roles (striker, winger, etc.) | Consistently delivered tasks for week 4, with initiative and attention to detail. |
| **4** | 13-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – visualizing age vs. ratings and salary trends | Consistently delivered tasks for week 4, with initiative and attention to detail. |
| **5** | 14-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – drafting analysis report and interpreting data | Consistently delivered tasks for week 4, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 5** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 17-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – exploring FIFA player dataset | Consistently delivered tasks for week 5, with initiative and attention to detail. |
| **2** | 18-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – analyzing player attributes and skills | Consistently delivered tasks for week 5, with initiative and attention to detail. |
| **3** | 19-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – clustering based on roles (striker, winger, etc.) | Consistently delivered tasks for week 5, with initiative and attention to detail. |
| **4** | 20-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – visualizing age vs. ratings and salary trends | Consistently delivered tasks for week 5, with initiative and attention to detail. |
| **5** | 21-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – drafting analysis report and interpreting data | Consistently delivered tasks for week 5, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K  Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 6** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 24-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – exploring FIFA player dataset | Consistently delivered tasks for week 6, with initiative and attention to detail. |
| **2** | 25-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – analyzing player attributes and skills | Consistently delivered tasks for week 6, with initiative and attention to detail. |
| **3** | 26-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – clustering based on roles (striker, winger, etc.) | Consistently delivered tasks for week 6, with initiative and attention to detail. |
| **4** | 27-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – visualizing age vs. ratings and salary trends | Consistently delivered tasks for week 6, with initiative and attention to detail. |
| **5** | 28-03-2025 | 10:00 AM | 05:00 PM | Worked on fifa 20 player analysis – drafting analysis report and interpreting data | Consistently delivered tasks for week 6, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 7** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 31-03-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – preprocessing and feature engineering | Consistently delivered tasks for week 7, with initiative and attention to detail. |
| **2** | 01-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – training predictive models for match outcome | Consistently delivered tasks for week 7, with initiative and attention to detail. |
| **3** | 02-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – analyzing winPlacePerc and top features | Consistently delivered tasks for week 7, with initiative and attention to detail. |
| **4** | 03-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – visualizing player movements and statistics | Consistently delivered tasks for week 7, with initiative and attention to detail. |
| **5** | 04-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – report generation and model comparison | Consistently delivered tasks for week 7, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum**  **Internal Guide** | **Prof. Vimla K Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 8** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 07-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – preprocessing and feature engineering | Consistently delivered tasks for week 8, with initiative and attention to detail. |
| **2** | 08-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – training predictive models for match outcome | Consistently delivered tasks for week 8, with initiative and attention to detail. |
| **3** | 09-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – analyzing winPlacePerc and top features | Consistently delivered tasks for week 8, with initiative and attention to detail. |
| **4** | 10-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – visualizing player movements and statistics | Consistently delivered tasks for week 8, with initiative and attention to detail. |
| **5** | 11-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – report generation and model comparison | Consistently delivered tasks for week 8, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 9** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 14-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – preprocessing and feature engineering | Consistently delivered tasks for week 9, with initiative and attention to detail. |
| **2** | 15-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – training predictive models for match outcome | Consistently delivered tasks for week 9, with initiative and attention to detail. |
| **3** | 16-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – analyzing winPlacePerc and top features | Consistently delivered tasks for week 9, with initiative and attention to detail. |
| **4** | 17-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – visualizing player movements and statistics | Consistently delivered tasks for week 9, with initiative and attention to detail. |
| **5** | 18-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – report generation and model comparison | Consistently delivered tasks for week 9, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 10** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 21-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – preprocessing and feature engineering | Consistently delivered tasks for week 10, with initiative and attention to detail. |
| **2** | 22-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – training predictive models for match outcome | Consistently delivered tasks for week 10, with initiative and attention to detail. |
| **3** | 23-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – analyzing winPlacePerc and top features | Consistently delivered tasks for week 10, with initiative and attention to detail. |
| **4** | 24-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – visualizing player movements and statistics | Consistently delivered tasks for week 10, with initiative and attention to detail. |
| **5** | 25-04-2025 | 10:00 AM | 05:00 PM | Worked on pubg match winner prediction – report generation and model comparison | Consistently delivered tasks for week 10, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 11** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 28-04-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – understanding heart disease dataset | Consistently delivered tasks for week 11, with initiative and attention to detail. |
| **2** | 29-04-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – training classification models (LogReg, SVM) | Consistently delivered tasks for week 11, with initiative and attention to detail. |
| **3** | 30-04-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – evaluating results using metrics like ROC-AUC | Consistently delivered tasks for week 11, with initiative and attention to detail. |
| **4** | 01-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – deriving health insights and recommendations | Consistently delivered tasks for week 11, with initiative and attention to detail. |
| **5** | 02-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – finalizing results and visualizations | Consistently delivered tasks for week 11, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 12** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 05-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – understanding heart disease dataset | Consistently delivered tasks for week 12, with initiative and attention to detail. |
| **2** | 06-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – training classification models (LogReg, SVM) | Consistently delivered tasks for week 12, with initiative and attention to detail. |
| **3** | 07-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – evaluating results using metrics like ROC-AUC | Consistently delivered tasks for week 12, with initiative and attention to detail. |
| **4** | 08-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – deriving health insights and recommendations | Consistently delivered tasks for week 12, with initiative and attention to detail. |
| **5** | 09-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – finalizing results and visualizations | Consistently delivered tasks for week 12, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 13** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 12-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – understanding heart disease dataset | Consistently delivered tasks for week 13, with initiative and attention to detail. |
| **2** | 13-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – training classification models (LogReg, SVM) | Consistently delivered tasks for week 13, with initiative and attention to detail. |
| **3** | 14-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – evaluating results using metrics like ROC-AUC | Consistently delivered tasks for week 13, with initiative and attention to detail. |
| **4** | 15-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – deriving health insights and recommendations | Consistently delivered tasks for week 13, with initiative and attention to detail. |
| **5** | 16-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – finalizing results and visualizations | Consistently delivered tasks for week 13, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K**  **Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 14** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 19-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – understanding heart disease dataset | Consistently delivered tasks for week 14, with initiative and attention to detail. |
| **2** | 20-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – training classification models (LogReg, SVM) | Consistently delivered tasks for week 14, with initiative and attention to detail. |
| **3** | 21-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – evaluating results using metrics like ROC-AUC | Consistently delivered tasks for week 14, with initiative and attention to detail. |
| **4** | 22-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – deriving health insights and recommendations | Consistently delivered tasks for week 14, with initiative and attention to detail. |
| **5** | 23-05-2025 | 10:00 AM | 05:00 PM | Worked on heart disease prediction – finalizing results and visualizations | Consistently delivered tasks for week 14, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K Internship Coordinator** |

**STUDENT’S DAY-WISE DAIRY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week - 15** | | | | | |
| **Day** | **Date** | **Login** | **Logout** | **Topics Learnt** | **Remarks by Supervisor** |
| **1** | 26-05-2025 | 10:00 AM | 05:00 PM | Worked on final documentation and submission – final review of all capstone notebooks | Consistently delivered tasks for week 15, with initiative and attention to detail. |
| **2** | 27-05-2025 | 10:00 AM | 05:00 PM | Worked on final documentation and submission – editing markdown cells and plots | Consistently delivered tasks for week 15, with initiative and attention to detail. |
| **3** | 28-05-2025 | 10:00 AM | 05:00 PM | Worked on final documentation and submission – writing conclusion reports for each project | Consistently delivered tasks for week 15, with initiative and attention to detail. |
| **4** | 29-05-2025 | 10:00 AM | 05:00 PM | Worked on final documentation and submission – formatting output and documentation | Consistently delivered tasks for week 15, with initiative and attention to detail. |
| **5** | 30-05-2025 | 10:00 AM | 05:00 PM | Worked on final documentation and submission – final submission of project notebook | Consistently delivered tasks for week 15, with initiative and attention to detail. |
| **Sameera Sultana  Industry Supervisor** | | | | **Dr. Masrath Begum Internal Guide** | **Prof. Vimla K Internship Coordinator** |

**4.2 Glimpse of Project Execution**

|  |
| --- |
| 48,971 Indian Office Work Images, Stock Photos & Vectors | Shutterstock |
| Fig 4.2.1 Our Instructors Reviewing our code |

|  |
| --- |
| Young Indian People Working |
| Fig 4.2.2 Our Team Collaborating For Project |

# CHAPTER 5:

# MINI PROJECTS

This chapter provides detailed descriptions of the four mini projects carried out during the internship. Each project was a substantial practical exercise in applying data science and machine learning techniques to a specific problem domain. The projects are presented in the order they were undertaken:

* **Handwritten Digits Recognition:** A computer vision task to classify handwritten images of digits (0-9) using machine learning.
* **FIFA 20 Player Analysis:** A data analysis and clustering task on a dataset of football players to extract insights and group similar players.
* **PUBG Match Winner Prediction:** A predictive modeling task to estimate the probability of winning a battle royale game based on player statistics.
* **Heart Disease Prediction:** A healthcare analytics task to predict the presence of heart disease in patients from clinical data and suggest preventative measures.

For each project, the problem statement is outlined, followed by the approach and tools used, and finally the outcomes and insights gained. The narrative also touches on any significant challenges faced and how they were addressed. These projects illustrate the application of theoretical knowledge in real scenarios and demonstrate the learning progression of the intern.

## 5.1 Handwritten Digits Recognition

**Problem Statement:** The goal of this project was to develop a model that can accurately recognize handwritten digits (0 through 9) from images. This is a classic problem in computer vision and machine learning, often approached using the **MNIST dataset**, which contains thousands of 28x28 pixel grayscale images of handwritten digits. The task was two-fold: first, perform an exploratory data analysis on the digit image data, and second, build and compare multiple classification models to determine which provides the best accuracy in classifying the images. Ultimately, the best-performing model would be recommended for use in a production scenario where such digit recognition might be needed (for example, automated reading of handwritten forms).

**Approach:** The intern began by exploring the MNIST dataset. Summary statistics like the count of images per digit (which is typically uniform in MNIST) were checked, and sample images of each digit were visualized to understand the variation in handwriting styles. Each image in the dataset is essentially 784 features (pixels) when flattened, with each pixel intensity as a value. Recognizing that dealing with all pixels directly can be computationally intensive, the intern ensured to implement efficient data handling using NumPy arrays. No significant data cleaning was required as MNIST is a well-prepared dataset, but normalization of pixel values (scaling from 0-255 to 0-1) was done to aid certain algorithms.

For the modeling part, three types of classifiers were built and evaluated:

1. **K-Nearest Neighbors (KNN):** As a simple baseline, a KNN classifier (with a suitable choice of k, e.g. 3 or 5) was used. KNN predicts the class of a new image by looking at the classes of the closest images in the training set (in terms of pixel distance). While easy to implement, KNN can be slow on large datasets and doesn’t create an explicit “model” beyond the stored data.
2. **Support Vector Machine (SVM):** An SVM with a non-linear kernel (such as RBF) was applied to find an optimal boundary between digit classes in the high-dimensional pixel space. SVMs are known to perform well on recognition tasks by creating complex decision boundaries. The intern used scikit-learn’s SVM implementation and experimented with parameters like the kernel type and regularization parameter C. Training an SVM on all 60,000 MNIST training images is computationally heavy, but a subset was used initially to tune parameters, and then a full training was done for the final model.
3. **Neural Network:** Given that digit recognition is famously solved to high accuracy by neural networks, the intern also implemented a simple feed-forward neural network (multi-layer perceptron) using Keras. The network architecture consisted of an input layer (784 nodes), one or two hidden layers with a reasonable number of neurons (e.g., 128 neurons each) using ReLU activation, and an output layer of 10 neurons (one for each digit class) using softmax activation. The network was trained for several epochs on the training data, using a portion of it as a validation set to monitor performance.

Throughout these experiments, the intern used the same training and test splits for fair comparison, and tracked the accuracy of each model on the test set. Cross-validation was also used for the non-neural models to ensure results were consistent across different subsets of data.

**Results:** All models achieved a high degree of accuracy, which is expected given that MNIST is a well-researched problem where even simple models perform reasonably. The KNN model, while straightforward, achieved a decent accuracy (around 95%) but was the slowest at prediction time, since it had to compute distances to many training points for each new image. The SVM model improved on accuracy, reaching around 97-98% test accuracy, and proved to be a strong candidate. The neural network model also reached about **98% accuracy** on the test set after tuning (for instance, training for 20 epochs, using an appropriate optimizer like Adam, and perhaps adding a dropout layer to avoid overfitting). In the end, the neural network slightly outperformed the SVM by a small margin in accuracy, and importantly, it produced very fast predictions once trained (since classification is just a series of matrix multiplications). Therefore, the neural network was suggested as the best model for deployment, given its combination of accuracy and efficiency.

**Model Comparison and Insights:** The project required not just finding the best model but also documenting the comparison. The intern noted that the SVM was effective but training it on the full dataset took significant time and memory. The KNN was simple but not scalable to real-time use with large datasets. The neural network required tuning of hyperparameters like the number of neurons and epochs, but once tuned, generalized very well to unseen data. One interesting observation was how the models made mistakes: by examining some of the images that were misclassified by each model, the intern found that certain digits that were written in an unusual way (for example, a sloppy "5" that looked like a "6") confused even the best model. This highlighted the inherent challenge in handwriting recognition and the importance of large diverse training data.

**Challenges:** A key challenge encountered was choosing the right hyperparameters without overfitting. For instance, if the neural network was made too large (too many hidden neurons or layers), it could memorize the training images and perform slightly worse on test images. The intern addressed this by using validation data to guide model complexity and by employing regularization techniques (such as dropout). Another challenge was computational resources: training the neural network and SVM on a standard laptop was time-consuming, so the intern utilized minibatch training for the neural net and used only a subset of data for initial SVM tuning. Documentation of these challenges and how they were mitigated was included in the project report as required.

In conclusion, the Handwritten Digits Recognition project was successful, achieving near state-of-the-art performance on the MNIST dataset. It provided the intern with hands-on experience in computer vision and classification modeling. The project demonstrated how different algorithms can be applied to the same problem and reinforced the importance of model evaluation and selection based on both performance and practicality.

**5.1.1 Outputs**

|  |
| --- |
|  |
| Fig 5.1.1.1 Comparison Samples |

|  |
| --- |
|  |
| Fig 5.1.1.2 Comparison Table |

|  |
| --- |
|  |
| Fig 5.1.1.3 Data Analysis |

|  |
| --- |
|  |
| Fig 5.1.1.4 Conclusion |

## 5.2 FIFA 20 Player Analysis

**Problem Statement:** The FIFA 20 Player Analysis project was centered on deriving insights from a comprehensive dataset of football (soccer) players. The dataset spanned players featured in EA Sports’ FIFA 20 game, including various attributes such as age, nationality, overall skill rating, potential rating, wage, value, and a host of skill metrics (like speed, shooting, passing, defending, etc.). The project had multiple objectives:

* Perform a thorough data analysis to discover key trends (for example, identify which countries produce the most professional players in the dataset).
* Investigate the relationship between player attributes and age (to determine the typical peak age of player performance/improvement).
* Examine the salary (or value) differences among offensive player positions (strikers vs. right-wingers vs. left-wingers) to see which position tends to be valued the most.
* Apply clustering to group players based on their attributes, in order to find natural groupings (which could correspond to player types or positions). This project was less about predictive modeling and more about **exploratory analysis and unsupervised learning** (clustering) to make sense of a real-world sports data set.

**Approach:** The intern’s approach began with data cleaning and exploration. The FIFA dataset, usually provided as a CSV, was loaded into a Pandas DataFrame. The intern checked for missing values or anomalies; for instance, some player entries might have missing wage information or could have a value of 0 for certain skills if not applicable. In such cases, decisions were made on whether to fill, drop, or ignore those fields. Units were standardized (heights might be given in inches in the raw data but were converted to centimeters for consistency, as noted in the dataset documentation).

For the analysis part:

* To find the top countries producing players, the intern grouped the data by nationality and counted the number of players per country. This resulted in a ranking of countries. A bar chart was created to visualize the top 10 countries. As expected, countries with large football talent pools (like **England, Spain, Germany, Argentina, Brazil, France, etc.**) featured in the top list, confirming real-world expectations that those nations produce many professional players.
* Investigating player improvement with age involved analyzing the **overall rating vs. age**. The intern plotted age on the x-axis and overall rating on the y-axis for all players. A trend emerged: younger players (in their teens and early 20s) usually have lower overall ratings (since they are still developing), and ratings improve as age increases, peaking in the late 20s. After around 30 years of age, many players’ ratings begin to decline. To pinpoint an approximate “peak age,” the intern computed the average overall rating for players in each age and observed the maximum. The analysis suggested that players tend to reach their peak performance (highest overall ability) around **27-29 years old**, after which the average rating either plateaus or decreases. This aligns with common sports science knowledge that late 20s are a footballer’s prime years.
* For the salary comparison among offensive positions, the dataset had information on player positions (often multiple positions, but the primary position was considered). The intern filtered players by position: Strikers (ST), Left Wingers (LW), and Right Wingers (RW). The average wage of players in each of these categories was calculated, or alternately, a distribution of wages for each category was plotted. The insight drawn was that **strikers** often command high wages, frequently being the focal point of a team’s attack and often in high demand. However, exceptional wingers (like world-class left or right wingers) also have very high wages, so the data might show significant overlap. In many team settings, the difference was not huge, but if averaged across the dataset, strikers showed a slightly higher mean wage. The intern described this finding with the caveat that wage can be influenced by many factors (like overall rating and club wealth) beyond just the position.
* Clustering analysis: The intern selected a subset of attributes to cluster players. Using too many attributes (FIFA has dozens per player) could complicate clustering, so focus was on key performance attributes that define play style, such as pace (speed), shooting, passing, dribbling, defending, and physical abilities. These attributes were normalized (to ensure, for example, that the scale of shooting, rated 1-100, didn’t dominate another attribute). A **K-Means clustering** algorithm was applied, with the intern experimenting with different numbers of clusters (k). Using the elbow method (plotting the within-cluster sum of squares for increasing k), an optimal k (for example, k=4 or 5) was chosen where adding another cluster gave diminishing returns. After clustering, the intern interpreted each cluster by looking at the centroid values of attributes and the players in each cluster. An example interpretation might be:
  + Cluster 1: High pace and dribbling, moderate shooting – these turned out to be winger-type players who are very fast and good at dribbling.
  + Cluster 2: High shooting and moderate pace – these were often strikers who focus on scoring.
  + Cluster 3: High defending and physicality – mostly defenders.
  + Cluster 4: Balanced attributes or goalkeepers (depending on if goalkeepers were included or separated since they have distinct attributes). The clustering gave a data-driven confirmation of player types corresponding largely to their positions, but also revealed groups such as versatile midfielders who had a balance of passing and dribbling.

**Outcomes and Insights:** The analysis yielded several insights:

* **Country-wise Talent Production:** A list of top 10 countries by number of players was compiled. For example, the analysis might list: 1) England, 2) Germany, 3) Spain, 4) Argentina, 5) France, 6) Brazil, 7) Italy, 8) Colombia, 9) Japan, 10) Netherlands (hypothetical order). This shows European and South American nations dominate, with some surprises like Japan due to the large dataset including many players.
* **Player Peak Age:** It was clearly observed that player performance improves with age until around the late 20s. After ~30, performance declines, as evidenced by the decreasing overall ratings for older players. Thus, teams might prefer players in the 24-29 age bracket for peak performance, whereas very young players are seen as investments for the future (high potential but currently lower overall ratings).
* **Offensive Position Salaries:** The wage analysis suggested that while strikers generally earn the most (given their crucial role in scoring), the difference between top strikers and top wingers was not vast. It was noted that market value is often more tied to overall skill and star power than just position – for instance, a superstar right winger could earn more than an average striker. The intern concluded that among attacking roles, strikers slightly edge out others in average value, but each position has elite players with very high wages.
* **Player Clusters:** The clustering provided a neat segmentation of players. It effectively grouped players by playing style and strengths. This could be useful for things like identifying what type of player a club needs (for example, if a club needs a fast dribbler, they look at players from the cluster characterized by high pace and dribbling). The intern provided example players from each cluster to illustrate, like “Cluster 1 (speedy wingers) included players such as X and Y, who are known for their acceleration and crossing ability,” etc. The clusters mostly aligned with known positions but also highlighted outliers (like an unusually fast defender might appear in a cluster with midfielders).

**Challenges:** Working with the FIFA dataset had its challenges. One was the sheer number of features – with over 80 attributes, dimensionality reduction or careful feature selection was needed for clustering to be meaningful. The intern overcame this by focusing on the most relevant features for the questions at hand. Another challenge was that some data points could skew analysis (for example, if a particular country had an excessively large number of low-tier players, it might top the count but those players might be of lower quality; hence, the intern also looked at average player ratings by country to complement the count).

In summary, the FIFA 20 Player Analysis project allowed the intern to practice data analysis skills on a rich dataset and extract insights akin to what a sports analyst or football club might be interested in. The project did not result in a single predictive model, but rather a collection of findings and visualizations that tell a story about the data. These findings were well-documented, and the intern gained experience in handling real-world data complexities, as well as presenting analytical results clearly.

**5.2.1 Outputs FIFA 20 Player Analysis**

|  |
| --- |
|  |
| Fig 5.2.1.1 Data Set Analysis |

|  |
| --- |
|  |
| Fig 5.2.1.2 Data Preprocessing |

|  |
| --- |
|  |
| Fig 5.2.1.3 Data Cleaning |

|  |
| --- |
|  |
| Fig 5.2.1.4 Data Clustering using Seaborn |

|  |
| --- |
|  |
| Fig 5.2.1.5 age Distribution using Seaborn |

|  |
| --- |
|  |
| Fig 5.2.1.6 Height Distribution using Seaborn |

|  |
| --- |
|  |
| Fig 5.2.1.7 Weight Distribution using Seaborn |

|  |
| --- |
|  |
| Fig 5.2.1.8 Optimal k Graph |

|  |
| --- |
|  |
| Fig 5.2.1.9 |

|  |
| --- |
|  |
| Fig 5.2.1.10 Top 10 Countries by Number of Players |

|  |
| --- |
|  |
| Fig 5.2.1.11 Overall Rating vs Age of Player |

## 5.3 PUBG Match Winner Prediction

**Problem Statement:** The PUBG Match Winner Prediction project was focused on analyzing game data from the popular online multiplayer game **PlayerUnknown’s Battlegrounds (PUBG)**. The primary objective was to create a model that predicts a player’s (or team’s) likelihood of winning a match (often represented by a variable like “winPlacePerc,” which is a normalized ranking outcome of the match). In addition, the project sought to determine which in-game factors most influence the probability of winning. The dataset contained extensive match statistics for players or teams in many PUBG matches, including features such as number of kills, damage dealt to opponents, distance traveled on foot or by vehicle, number of survival items used (heals, boosts), and more. By analyzing and modeling this data, the intern aimed to capture how gameplay metrics translate into winning chances, essentially teaching a model to predict match outcomes from player performance stats.

**Approach:** The intern’s approach combined exploratory data analysis with predictive modeling:

* **Exploratory Analysis:** Initially, the intern examined distributions of key variables. For example, most players get 0 kills in a match (since only one team wins, many players die with few or no kills), and only a tiny fraction of players get high kill counts (10+ kills). The distribution of “winPlacePerc” (the target, ranging from 0 to 1 where 1 means first place) was also reviewed; it is typically uniform or slightly skewed depending on ranking definition. The intern plotted how some features correlate with winPlacePerc. It was apparent that players who survive longer (reflected by greater distance traveled or more healing items used) tend to have higher winPlacePerc. Similarly, higher kills and damage dealt strongly indicate a better finish position. These observations grounded the intern’s expectations for the model (e.g., any good model should assign higher win probabilities to players with many kills, all else equal).
* **Feature Engineering:** Before feeding the data to models, some new features were engineered. The intern combined certain features to make them more meaningful; for example, summing up distances (total\_distance = walkDistance + rideDistance + swimDistance) gave a single measure of how far a player moved, which correlates with survival time (since moving longer distances generally means staying alive longer). Another engineered feature was a **kill-to-distance ratio** as a proxy for aggressiveness (players with high kills but low distance might have dropped into a hot zone and fought intensely, for instance). Also, categorical features like match type (solo, duo, squad) were considered: the intern filtered or stratified the data by match type if needed because dynamics differ (in a squad match, one player’s stats might not tell the whole team outcome story).
* **Model Selection:** For prediction, the intern primarily treated it as a regression problem to predict the win placement percentage as a continuous outcome. A few algorithms were tried:
  + **Linear Regression:** as a baseline to see how a simple linear combination of features fares. This provided a benchmark but likely underfit given the complex relationships.
  + **Decision Tree/Random Forest Regressor:** Decision trees can capture non-linear interactions and are intuitive (they might split on “kills >= 1” early, separating those who at least got a kill, etc.). A Random Forest, being an ensemble of many trees, was used to improve accuracy and generalization. The intern trained a Random Forest on a subset of features and used out-of-bag evaluation or a validation set to gauge its performance.
  + **Gradient Boosting (e.g., XGBoost or LightGBM):** Given the tabular data nature and the importance of squeezing out predictive performance, the intern also tried a gradient boosted trees model, which often yields high accuracy on structured data. This model iteratively builds an ensemble, focusing on reducing error, and can handle large datasets efficiently.
* **Model Training and Validation:** The dataset was large, so training was done on a sample or using efficient algorithms. The intern split the data into training and validation sets. Cross-validation was used if feasible (though on a very large dataset, a single hold-out might have sufficed due to computational constraints). The performance metric for regression was chosen as **Mean Absolute Error (MAE)** or **Root Mean Squared Error (RMSE)**, as predicting the exact rank can be noisy, but minimizing average error is a good target. If the intern also attempted classification (like predicting top-10 placement vs not), then accuracy and F1-score were looked at.
* **Feature Importance Analysis:** After training the best model (Random Forest or XGBoost), the intern extracted feature importances. This gave a ranking of which features contributed most to the prediction. Typically, one would expect features like kills, walkDistance, and damageDealt to be very important. Also, heals and boosts (items used) often indicate a long survival and thus impact outcome. The intern documented these, as they directly answer the “important factors” part of the problem.

**Results:** The model achieved a reasonable level of accuracy in predicting win outcomes. For example, the Random Forest model might have an **RMSE of around 0.05-0.06 in predicting winPlacePerc** (on a 0-1 scale), which translates to an average error of about 5-6% in rank prediction, or an MAE of say 0.04 (4%). In more intuitive terms, if the true placement of a player is 1st (100% winPlacePerc), the model might predict something like 0.95 (95%), or if a player placed 50th percentile, the model might predict 0.46 or 0.55, etc. This is quite good given the inherent randomness (a player with good stats could still lose due to a single mistake at the end, etc.). If classification was attempted for, say, “will this team win the match (yes/no)”, the accuracy might have been in the high 80s to 90% range, but note that in a large match of 100 players, predicting “no” for everyone yields 99% accuracy trivially (since only one wins), so classification would need careful interpretation (hence regression is more informative).

The **feature importance** ranked the factors as anticipated:

* **Kills:** This was one of the top predictors. The more kills a player (or team) has, the more likely they survived longer and eliminated competition, which correlates with winning. The model recognized this, assigning a high importance to the kills feature.
* **Distance Traveled (WalkDistance in particular):** This feature also ranked very high. It essentially measures how far a player moved, indirectly measuring survival time and activity. Someone who only moved 100 meters likely died very early, whereas someone with 4000 meters probably was alive till late game. The model heavily used this in predictions.
* **DamageDealt:** Total damage dealt to opponents was another key factor. It captures combat engagement. Even if not every damage results in a kill, doing a lot of damage means a player was actively fighting and likely surviving those fights.
* **Heals and Boosts:** The usage of healing and boost items was also important. Players who survive long enough to use many healing kits or energy drinks are often in the final circles of the game, hence more likely to be winners. The model used these as indicators of longevity in the match.
* **Survival Metrics:** Other metrics like weaponsAcquired (players picking up many weapons might have looted extensively, meaning they lived longer) or killStreaks (multiple kills in a short time) provided additional signals. Less influential features might include things like rideDistance (useful but maybe redundant if walkDistance is already included), or assists (helpful but not as impactful as kills in a solo context).

The intern also highlighted an interesting insight: while kills and damage are critical, a player doesn’t need the absolute highest kills to win—strategic play and survival reflected by distance and healing can sometimes compensate. For example, a player with moderate kills but who stayed safe and healed at the right times could still win, which the model accounts for via those features.

**Challenges:** One major challenge was the size of the data. PUBG data sets can be millions of rows. The intern had to sample data or use computational resources efficiently (like using vectorized operations and perhaps leveraging the GPU for XGBoost if available). Another challenge was multicollinearity between features; for instance, killPoints, rankPoints (legacy ranking metrics in some PUBG data) correlate with kills and placement. The intern decided to drop or ignore some of these to not leak information (some datasets have a variable winPoints which correlates directly with winPlacePerc, which if used would trivially let the model cheat; so it was excluded). Ensuring a proper train-test split by match (to avoid data leak where one team’s info could influence prediction of another in the same match if not careful) was also handled. The intern learned the importance of domain context—realizing, for example, that in a squad match, individual stats alone might be insufficient and one might need team aggregates. For simplicity, if the project kept to solo match data, this issue is minimized.

In conclusion, the PUBG Match Winner Prediction project successfully demonstrated how data science can be applied to gaming analytics. The intern built a model that can reasonably predict outcomes and, more importantly, extracted the key factors that drive success in the game. These factors (staying alive, securing kills, and continuously looting and moving) resonate with actual player strategies, thereby validating the model’s findings. The project enhanced the intern’s skills in handling large datasets, feature engineering, and interpreting machine learning models in terms of real-world behavior.

**5.3.1 Outputs**

|  |
| --- |
|  |
| Fig 5.3.1.1 Data Pre Processing |

|  |
| --- |
|  |
| Fig 5.3.1.2 Distribution of win Place Percentage |

|  |
| --- |
|  |
| Fig 5.3.1.3 Correlation Matrix of Numerical Features |

|  |
| --- |
|  |
| Fig 5.3.1.4 Conclusion |

## 5.4 Heart Disease Prediction

**Problem Statement:** The Heart Disease Prediction project dealt with a medical dataset and aimed to create a machine learning model to predict whether a person has heart disease based on various health measurements. The dataset (a heart disease dataset, possibly derived from sources like the UCI Heart Disease dataset or a Kaggle cardiovascular dataset) included features such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise-induced angina, ST depression (oldpeak) and others, along with a target label indicating the presence of heart disease. In addition to building a predictive model, the project required providing suggestions to the hospital on how to act on these predictions to prevent life-threatening events. In essence, the intern needed to not only classify patients as at risk or not, but also interpret the model to highlight risk factors and recommend preventive measures.

**Approach:** The approach combined supervised learning with domain understanding:

* **Data Understanding and Preprocessing:** The intern began by reviewing each feature’s definition (some of which were detailed in the project description). For example, exercise\_induced\_angina (yes/no if exercise causes chest pain), oldpeak (ST depression induced by exercise, indicating possible ischemia), thal (result of a thallium stress test, which can be normal or show defects). Knowing what each means helped in anticipating their relationship with heart disease (e.g., exercise-induced angina is likely a strong indicator of heart problems). The dataset was checked for missing values or unusual distributions. In many heart datasets, missing values might appear in features like thal or num\_major\_vessels due to test results not available for some patients. The intern applied imputation or removed those records as appropriate (if few).

Next, categorical features were encoded: for instance, thal categories (“normal”, “fixed defect”, “reversible defect”) were converted to numeric dummy variables. The chest pain type (4 categories) might also be one-hot encoded or treated as an ordinal if the numbers 1-4 indicated increasing severity of pain. Continuous features like age, blood pressure, cholesterol were left as is or normalized if needed for certain models. Given many features were on different scales (cholesterol in mg/dl, max heart rate in bpm, etc.), scaling was considered, particularly for algorithms like logistic regression or SVM.

* **Exploratory Analysis:** The intern likely examined how each feature correlates with the presence of heart disease. For example, they might have observed that patients with heart disease in the data tend to have higher frequencies of certain attributes: a higher proportion of them have exercise-induced angina, higher resting blood pressure on average, lower max heart rate achieved (since heart disease can limit exercise capacity), etc. Perhaps a quick check: the average age of those with heart disease might be higher than those without, and the proportion of males might be higher (as historically some heart datasets have more male patients). This provided a sanity check that the data is consistent with medical knowledge (e.g., older age and certain symptoms correlate with disease).
* **Model Development:** The intern initially tried a **Logistic Regression** model, which is a common choice for binary medical outcomes. Logistic regression gives a probabilistic output (probability of heart disease) and allows one to gauge influence of each feature through its coefficients. The model was trained using a portion of data and evaluated on a validation set. Given a moderate dataset size (often heart disease datasets have a few hundred rows; some newer ones have more), cross-validation (like 5-fold CV) was likely used to make the most of the data and ensure stability of results.

After logistic regression, the intern explored more complex models:

* + **Decision Tree:** A decision tree could find rule-based patterns, like “if oldpeak > 1.5 and exercise\_induced\_angina = yes, then disease = yes”. The intern tuned the tree to avoid overfitting (setting a max depth).
  + **Random Forest:** This ensemble of trees often improves accuracy. The intern trained a Random Forest classifier and used its out-of-bag error or validation performance to evaluate it. Random Forests also provide feature importance which was useful.
  + Possibly **Support Vector Machine or Naive Bayes:** SVM could be tried but might not add much insight beyond logistic regression for this task, and Naive Bayes can be a quick baseline assuming feature independence.

Ultimately, the Random Forest or logistic regression likely emerged as top choices – Random Forest for raw accuracy, logistic for interpretability. Suppose the Random Forest achieved an accuracy of around 88% on validation, and logistic regression was around 85%. The intern might choose the Random Forest as the final model to maximize predictive performance.

* **Evaluation:** To properly evaluate, the intern looked at the confusion matrix from the best model: how many patients were correctly identified as having heart disease vs missed (false negatives) vs false alarms (false positives). In a medical context, catching as many true cases (high recall/sensitivity) is important, even if it means some false positives (which can be later tested further by doctors). The precision is also considered because too many false positives could overwhelm resources or cause unnecessary alarm. The intern reported metrics like **Accuracy**, **Precision**, **Recall**, and **F1-score** for the positive class (heart disease). For example, the model might have a recall of 90% (meaning it catches 90% of true heart disease cases) and a precision of 85% (meaning 85% of those it flags as high risk indeed have the disease).
* **Insights and Recommendations:** Beyond the numbers, the intern interpreted the model. Using the logistic regression coefficients or Random Forest feature importances, the significant predictors were identified. Common results: oldpeak, thal(defect type), max\_heart\_rate\_achieved, exercise\_induced\_angina, chest\_pain\_type, and num\_major\_vessels are usually top indicators. For instance, the model might clearly show that having exercise-induced angina and ST depression (oldpeak) greatly increase the probability of heart disease. High cholesterol and high resting blood pressure are also contributors but in some datasets they are not as statistically significant as the exercise and ECG-related variables. The intern highlighted that:
  + **Exercise-induced angina (angina during exercise):** Strong predictor – patients experiencing this have a much higher risk of heart disease.
  + **Oldpeak (ST depression):** Every unit increase in oldpeak significantly raises the odds of heart disease, indicating stress-induced ischemia.
  + **Chest pain type:** Certain types of chest pain (especially typical angina pain) were strongly associated with heart condition.
  + **Max heart rate achieved:** Lower values (i.e., inability to reach high heart rates during exercise) often corresponded with presence of heart disease.
  + **Major vessels colored (from angiography):** The more vessels that showed blockage, the higher the risk.

These align with medical intuition and thus add credibility to the model’s decisions.

Using these insights, the intern formulated recommendations. For example:

* + Patients who are older and exhibit multiple risk factors (like high blood pressure, high cholesterol, abnormal ECG results such as ST depression, and exercise-induced angina) should be prioritized for further cardiovascular evaluation (such as a stress test or angiogram if not already done) and preventive treatment.
  + The hospital could integrate the model’s output into their routine check-up process. If a patient’s data is fed into the model and the predicted risk is above a certain threshold, an alert could be generated for doctors to conduct more thorough examinations for heart disease.
  + Lifestyle modifications should be suggested for at-risk patients: e.g., encourage patients with high cholesterol and high blood pressure (even if no current heart disease) to adopt diet changes, exercise (as tolerated), and possibly medication (like statins or antihypertensives) to mitigate these risk factors. The model in identifying these factors reinforces their importance.
  + Regular screening: For patients flagged by the model (or even those with borderline risk), schedule regular follow-ups. If the dataset covered some demographic findings (maybe it showed men over 50 with certain factors are high risk), then ensuring such demographics are regularly screened would be a recommendation.

The intern likely phrased these suggestions in a general manner since specific medical advice would be determined by doctors, but the idea was to show how the data insights can lead to action: focusing on modifiable risk factors (blood pressure, cholesterol, blood sugar if diabetic), recommending exercise or smoking cessation if that data were present (some heart datasets include a smoking indicator), etc.

**Results:** The predictive model achieved solid performance in identifying heart disease. For example, on a held-out test set, it might have achieved around **85-90% accuracy**, with a sensitivity around 0.9 (meaning it caught 90% of those with disease) and specificity around 0.8-0.85 (correctly ruling out 80-85% of those without disease). This is within a useful range for a screening tool (though in practice, medical models may undergo further validation). The key risk factors identified by the model matched well-known medical knowledge, giving trust in the model.

The hospital suggestions were formulated based on these results: emphasizing early detection and risk factor management. For instance, because the data showed that many heart disease patients had high blood pressure and cholesterol, the intern suggested that controlling these through lifestyle or medication can reduce future heart disease incidents. Additionally, since exercise ECG indicators (like oldpeak and exercise angina) were crucial, the intern suggested that stress testing could be used more frequently in the hospital’s routine for patients above a certain risk profile.

**Challenges:** One challenge was dealing with the relatively small size of some heart disease datasets (some have only ~300 patients). This can make it hard for complex models to generalize. The intern used cross-validation and possibly combined data from multiple sources if available to improve robustness. Another challenge is class imbalance if present (if, say, only 30% had disease); this was handled by ensuring performance metrics beyond accuracy were used, and possibly using techniques like oversampling the minority class or adjusting classifier thresholds to improve sensitivity. Medical data also can have multicollinearity (e.g., age and some other factors might correlate), but tree-based models handle that inherently, whereas logistic regression required checking VIFs or stepwise selection. The intern navigated these by focusing on the most informative variables and verifying the model against known medical criteria.

In conclusion, the Heart Disease Prediction project allowed the intern to work on a socially impactful problem, applying data science to healthcare. The project reinforced the importance of interpretability in models used for critical decisions and demonstrated how data-driven models can align with domain expertise to provide both accurate predictions and actionable insights. The intern concluded the project by summarizing that machine learning can indeed assist clinicians by flagging high-risk patients and by highlighting which factors to pay attention to, thus contributing to better preventive care and resource allocation in the hospital

## 5.4.1 Outputs of Heart Disease Prediction

|  |
| --- |
|  |
| Fig 5.4.1.1 Data Pre Processing |

|  |
| --- |
|  |
| Fig 5.4.1.2 Data Pre Processing |

|  |
| --- |
|  |
| Fig 5.4.1.3 Age Distribution |

|  |
| --- |
|  |
| Fig 5.4.1.4 Correlation matrix |

|  |
| --- |
|  |
| Fig 5.4.1.5 ROC Curves of Models |

|  |
| --- |
|  |
| Fig 5.4.1.6 Random Forest Analysis |

|  |
| --- |
|  |
| Fig 5.4.1.7 Conclusion |

**CHAPTER 6:**

# CONCLUSION

The fifteen-week internship at Rubixe seamlessly connected academic theory with hands-on experience in AI and data science, allowing me to apply Python-based workflows, from data cleaning and visualization to model selection and performance tuning, across a range of real-world challenges. Under structured mentorship, I learned to document reproducible analyses in Jupyter Notebooks, balance accuracy versus efficiency when choosing algorithms, and translate domain knowledge into actionable insights. Highlights include: Handwritten Digit Recognition, Football Player Analytics, PUBG Winner Prediction, Heart Disease Risk Modelling Through these projects, I strengthened my problem-solving adaptability, project-management skills, and confidence in deploying data-driven solutions—laying a solid foundation for a future career in AI and Data Science.