18.12.24 – Laying down the overall plan

Finally after finishing my last exams for this semester, I have some time for myself. I have wanted to get into Data Science and Machine Learning for a while now, so I will devote this holiday exactly for this purpose. Currently I have basic understanding of the python language, where I can construct small programs, however I have little to no experience with handling datasets. I do however have an understanding of the theoretical side of statistics and databases through University courses. The plan for this holiday is as follows:

- learn the required python libraries for data analysis (1)
- Revise the SQL language (2)
- Learn the required machine learning concepts (3)

20.12.24 – The more precise plan (The materials)

I do require a solid theoretical foundation to even know what I am doing in the first place, so books are going to be my primary source of learning. However all the 3 areas also require programming and active engagement with the materials, so this is my plan to address each of the 3 areas:

1) Python Data Analysis
- O'Reilly 'Python Data Science Handbook' Jake VanderPlas
- Enrolling myself into Data Analysis with Python course

2) SQL language
- 'Fundamentals of Database Systems' Elmasri Navathe
- Implementation of the code in PostgreSQL built in dataset

I will only cover Chapter 6 and 7 from the book as these chapters primarily deal with SQL queries

3) ML concepts

- Machine Learning CS229 Stanford University
- Implementing them from Scratch as well learning to use the built-in library function for it

4) Implement a project for each area (for portfolio)

I will do all three of these side by side, instead of just focusing on one. This will allow me to simply switch to another area if I get bored or stuck in any particular one.

28.12. 24 - Python libraries and SQL

So far I have almost completed the python course through the aid of the book and other online sources. I now am familiar with the numpy, pandas and matplotlib (seaborn) libraries from python. I can now:

- Load the dataset
- Clean the data
- Handle missing values
- Plot the correlation between different variables (through correlation matrix)
- Visualize the relationships using different types of plots
- Do time series analysis
- Image processing (mirroring, dimming, finding clusters etc)

For the Data Analysis course now it's time to move to the Scikit-Learn library. However, first I'll need to familiarize myself with some ML concepts, before learning to use the libraries. The SQL queries did look simple, both because of the structure of the queries but also because I already had some familiarity with it. Here I feel like I can directly move onto the project part.

03.01.25 – Progress report II

I have fully completed the Data Analysis with python course including the basics of the Scikit Learn library. More importantly I feel like I have understood how things work at a technical level. I also learned some basic machine learning concepts such as linear/logistic regression, Naive Bayes classifier, basic neural network (perceptron), KNN (k-nearest neighbour), K-means clustering, some of these I learnt by implementing them from scratch in python.

I will attach the link to those below in the Project section below. All that is remaining now is the projects to showcase my understanding. I also need to learn how to upload the projects into my GitHub profile, which currently I am not too familiar with. Whilst I was learning machine learning, I felt like it opened up an entire new area for me to explore. There are many more ML algorithms that I will need to familiarize myself with. However, before that I will implement projects from what I have learnt so far.

Projects:

For the SQL project, I will just explore the dvdrental dataset, that is commonly used for learning purposes. For the data exploration using python, I will use the palaeontology dataset from Helsinki, further details are mentioned in the project.

Python Data Exploration project (Fossil NOW dataset)

This is by far my biggest project up until today, as it encompasses everything that I have learnt so far on the practical side. I did it using the palaeontology dataset curated in Helsinki. I had to load the dataset, analyze it using pandas, and also do many visualizations on the dataset including geospatial visualization, where I became particularly interested afterwards. I also did a small machine learning model on the dataset.

SQL project

This is a fairly small project, which taught me more about putting the projects out on GitHub. The project is titled SQL.md, where I practiced using the dvd rental datasets to formulate SQL queries for different purposes.

Distinguishing between number 5 and 9 using the simplest neural network setup (or any two numbers)

KNN from scratch (This is more of a small exercise than a project)

Reflection: 10.01.25

These are the last days before the start of the new semester. I have completed almost everything I have set out to do, including the projects in data analysis. The plan for the holiday was to 1) Revise the basics of SQL and construct a small project to showcase those. 2) Learn the basic python libraries for data analysis 3) Learn the basic machine learning concepts.

I have completed all three of these goals. I had previous knowledge in SQL, and I did end up spending very little time revising these but my focus laid primarily on learning the python libraries. I had taken a course for this, which was fairly long and now I feel like I understand numpy, pandas, matplotlib, seaborn, scikit-learn. The project in this area was very important, where I could evaluate what I have learnt. The project also turned out to be something that I am fairly proud of, where I feel that it showcases the quality of what I have learnt so far.

In area 3, I wasn't able to make the progress that I would've liked. I was already somewhat familiar with a perceptron and had programmed it from scratch to distinguish between two

handwritten digits. I did end up learning techniques like Linear/Logistic Regression, KNN, K-means clustering algorithm as well as the basics of decision trees. I, however, couldn't go into too much depth here in terms of programming these from scratch in python. I did however get introduced to using Linear Regression, KNN, K-means clustering through the scikit-learn library.

Area of development

- Whilst doing the projects, although I made it strict for myself not to use LLMs for answers (as this was for learning purposes), I did have to consult reading materials and examples quite a lot, so I will have to crystallize these concepts. This can only be done by more projects in the future, which I will do more of.

Plan for the next few months:

- Interactive data visualizations
- Geospatial data analysis
- Machine Learning

My original plan was to get into the depths of machine learning and deep learning after this but I developed a new interest whilst learning this course. I want to get more into geographical and geospatial data analysis. The project that I was doing briefly introduced me to this idea, and playing around with world maps and creating different visualizations seemed quite fun to do. I already learned some basic static visualizations during the course, however I want to move on to active data visualization using python libraries.

So, the next step is to learn geospatial data-analysis, and visualize world maps as per different criterias. The deadline I am setting myself for this is by the end of February.

Everything is on this GitHub page:
Github: https://github.com/ghimire-aayush/The-Start.git