LE DUY KHANG

Data Analyst Intern

PERSONAL INFORMATION

30/05/2004

0961125575

ld.khang305204@gmail.com

An Phu Dong, District 12, Ho Chi Minh City

https://github.com/LDKhang35

SKILLS

PROGRAMMING LANGUAGE

Python, SQL

DATA PROCESSING

Pandas, Numpy

DATABASE MANAGEMENT SYSTEM

MySQL, SQLite, MongoDB

BIG DATA & ETL

Hadoop, Spark (basic)

DATA VISUALIZATION

Power BI

DEVELOPMENT TOOLS & PLATFORMS

GitHub, Docker, Google Colab, Linux.

SOFT SKILLS

Strong teamwork abilities.
Quick to learn new technologies.
Proactive and highly responsible.

CAREER GOALS

I am a third year student majoring in Data Technology, I am looking for an internship opportunity in an Information Technology (IT) position. My desire is to experience a real working environment, learn operational procedures and apply the knowledge I have learned to specific projects.

EDUCATION

DATA TECHNOLOGY MAJOR — 2022 - Present

Van Lang University

COURSE PROJECT

ANIMAL IMAGE CLASSIFICATION USING CNN — 01/2025 - 03/2025

Members: 6 | Role: Build CNN model from scratch

Description:

- •Objective: Build an image classi cation system for animals using convolutional neural networks (CNN).
- •Technology used: Python, TensorFlow, Numpy, Pandas, Matplotlib.
- •Works performed:
  - ◦ Collected 26,200 images (~614 MB) from Kaggle and performed preprocessing (resizing, augmentation, etc.) to standardize the data for machine learning models.
  - ◦ Divide the data set (train 80%, validation 10%, test 10%).
  - ◦ Build and train a CNN model from scratch using TensorFlow.
  - ◦ Evaluate the model based on the accuracy of the test set.
  - ◦ Visualize the training process (loss/accuracy by epoch).
  - ◦ Developed a user-friendly interface that allows users to upload images and view classification results.
- •Results achieved:
  - ◦ The model achieved 92.41% accuracy on the test set.
  - ◦ User-friendly interface, easy to use.

GitHub link: https://github.com/LDKhang35/CNN.git

LUNG CANCER PATIENT RECORDS MANAGEMENT USING HADOOP & SPARK — 02/2025 - 04/2025

Members: 5 | Role: Data processing and analysis using Spark, storing on HDFS

Description:

- •Objective: Build a lung cancer patient management system.
- •Technology used: Python, Hadoop (HDFS), Spark, Pandas, RandomForest.
- •Works performed:
  - ◦ Collected dataset from Kaggle (CSV file, ~10 MB) and performed data cleaning and preprocessing using Pandas
  - ◦ Stored data using Hadoop Distributed File System (HDFS).
  - ◦ Build and train a Random Forest model to predict the risk of disease.
  - ◦ Use Spark to count common disease characteristics of patients by age.
  - ◦ Deploy a web app interface using Flask to display analysis results and support uploading CSV les for prediction.
- •Results achieved:
  - ◦ Successfully analyzed common characteristics among patients.
  - ◦ User-friendly interface, easy to use.

GitHub link: https://github.com/LDKhang35/Lung-cancer-Hadoop-Spark.git