Deepak Thakur

Data Scientist

9971678995 deepak2009thakur@gmail.com Github Linkedin

Experience

3 Pillar Global **Remote**

Data Assurance Engineer *Jan 2021 – Present*

Project - Detecting Doctor Specialties using Patient Health Data

- Developed and implemented machine learning models to **predict doctor specialties based on patient health data**. Apply QA principles to ensure data and model accuracy. Develop robust testing strategies for ML algorithms and data pipelines to ensure reliable performance and accuracy.
- Perform advanced data analysis including handling missing values, and outliers. Identify anomalies and data inconsistencies using statistical methods and visualization. Collaborate with cross-functional teams to ensure high data quality for analysis.
- Increased accuracy by 2 % and performance of the predictive models through rigorous feature engineering and optimization techniques.

Projects

Mice Protein Expression | *Python, flask, Kmeans, MySql* **August 2022**

- Classification of Mice based on the value of 77 proteins, Genotype(control or trisomy), Treatment type, and behavior.
- Developed an automated system to train the model and got Auc score of **98 %** for random forest.
- This system performs data validation, preprocessing, clustering, and model selection using hyperparameter tuning.
- Used MYSQL database to store training and prediction batch files.

Garment Recommendation System | *Python, K-nearest neighbors, trasnfer-learning, Resnet50, streamlit* **Jan 2023**

- Developed a **deep learning based web-app** to improve the user experience and to recommend various types of fashion products with respect to the choices.
- **ResNet50** pre-trained model is being used to take out the embedding for all the 44k images.
- K-nearest neighbors algorithm is used in the project to take out a similar product for the recommendation.

Sign Generation Language using Yolov5 | *Python, cnn, Yolov5* **Jan 2023**

- Designed a Sign Generation language model using **Yolov5**, to detect and classify signs in real-time.
- Improved model performance and reduced overfitting by implementing techniques such as data augmentation, transfer-learning, and hyperparameter tuning.
- Pre-processed **custom dataset of 6 different classes** with consistent size and quality inputted into yolov5 model.

Python Question Classification: ML-based StackOverflow Text Analysis | *DVC, mkdocs, NLP* **Feb 2023**

- binary classification model to classify StackOverflow questions as Python-related or not
- Achieved high accuracy and F1-score by preprocessing the data, and converting text data to numerical format using **TF-IDF vectorization**.
- Ensured reproducibility and efficient tracking of changes using **DVC**, with clear project documentation using **MKDocs**
- Demonstrated proficiency in end-to-end machine learning, including data preprocessing, model selection, project organization, and documentation.

Technical Skills

Languages: Python, Java, SQL
Libraries: Numpy, Pandas, Matplotlib, scikit-learn, keras, Tensorflow, Transformers
Technologies/Frameworks: GitHub, AWS, MongoDB, Flask, Streamlit,
Machine Learning: Linear & Logistic Regression, Decision Tree, Ensemble Techniques, PCA, Clustering algorithms, supervised, unsupervised, Exploratory data analysis, ANN, statistics, NLP

Blogs

Contributing to the technical community by writing informative blog posts on different topics of Data Science/Machine Learning, NLP highlighting their advantages, applications, architecture, and implementation details.

Education

Maharishi Dayanand University **Palwal, Haryana**

B.Tech(ECE) *2010 – 2014*