Data Science and Machine Learning Projects

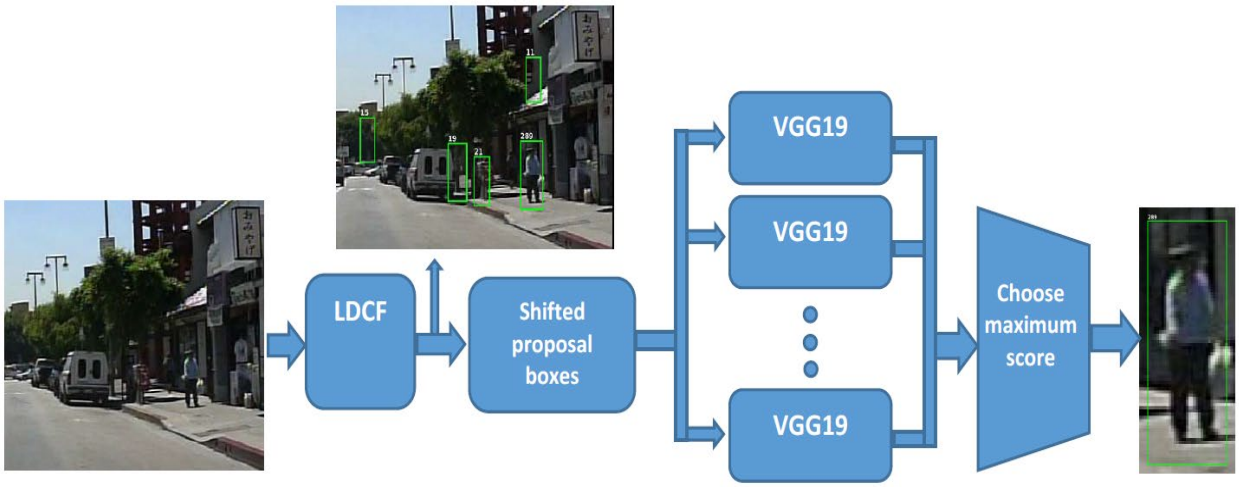
Outline:

- **Pedestrian detection in images**
- **Analyzing Hotel Reviews**
- **Heart disease prediction**
- **Finding Similar Items using LSH**
- **Evaluation of three methods for linear regression**

1. Pedestrian detection in images

- Developed a Convolutional Neural Network model based on VGG-19 architecture for Pedestrian detection in images.
- Abstract:

- Human-computer interaction has been widely studied in different applications such as driving assistant systems, human behavior analysis, and intelligent surveillance in recent years. Pedestrian detection is known as one of the practical fields in computer vision. Over the last decade, a variety of methods have been proposed to improve pedestrian detection systems. In recent years, many of these methods have used convolutional neural networks and deep learning to detect pedestrians. This thesis is aiming to improve pedestrian detection by presenting novel ideas in this area.
- The proposed system includes three parts. The first is an LDCF detector which is used to extract proposal windows from images. The first novelty of this thesis is to use extra windows around the proposed ones to improve matching between of Miss Rate over FPPI, which is well known in pedestrian detection evaluations, improves 2.71 percent. The second part of the presented system is a deep convolutional neural network based on VGG19 architecture. The second novelty of this thesis is the usage of grayscale images instead of color images to train the network. Since pedestrian detection system should not concern the color of body and clothes of people, by removing the redundant information, Miss Rate is improved 1.25 percent. The third novelty of this thesis is the use of three parallel networks, while the second and the third ones are trained on the data which are not classified correctly or their classification confidence is low in its previous network. By summing the results of these three networks, Miss Rate is improved 2.84%. On the whole, the proposed method in this thesis reduces the Miss Rate of basic system by 6.8%, achieving the Miss Rate over FPPI criterion of 12.63 percent.



- Green boxes: ground truth, Red boxes: improved box, Yellow box: initial LCDF proposal.

- **Publication**: S. Ghaffari and A. A. Raie, "Pedestrian detection using improved proposal boxes and grayscale convolutional learning," *2017 10th Iranian Conference on Machine Vision and Image Processing (MVIP)*, 2017, pp. 70-75, doi: 10.1109/IranianMVIP.2017.8342371.

2. Analyzing Hotel Reviews

- Analyzed hotel reviews (from Booking.com) and summarized them based on Frequent Words

- Preprocessed the review text by tokenizing and removing stop words. Extracted features using TF-IDF method and Natural Language ToolKit (NLTK) library.
- Applied K-means and BIRCH clustering algorithms on the data. Determined the frequent items in each cluster.
    - Tools: Python, NLTK
    - Link to the complete report : Link
    - Link to the github repository : Link
    - Which feature is more important to visitors?

3. Heart disease prediction

- Analysis of given data and 3 supervised classification methods, namely Logistic Regression, KNN, and Random Forest, for predicting heart disease in near future using given features based on Python using Pandas, NumPy, Scikit-learn, and Matplotlib libraries. In this project Accuracy, Precision, and Recall are used as the evaluation metrics.
- Tools: Python, Pandas, NumPy, Scikit-learn, Matplotlib, Jupyter Notebook
- Link to the github repository: link
- More information to be added.

4. Finding Similar Items using LSH

- Implemented locality sensitive hashing for finding the similarities in 290k questions from Quora efficiently in limited time.
- Tools: Python
- Link to the github repository: link
- More information to be added.

5. Evaluation of three methods for linear regression

- Three methods are coded in Python
- Tools: Python
- 1. Solving normal equation
- 2. Gradient descent (GD)
- 3. Stochastic gradient descent (SGD)
- Link to the github repository: link
- More information to be added.