**BASIC CERTIFICATION COURSE IN ARTIFICIAL INTELLIGENCE**

List of Projects

*[Project completion certificate is attached towards the end ]*

**Case Study 1:**

**Analysis of Treatment Quality and Efficiency of Hospitals**

**Source:** Journal of Health and Medical Sciences, Authored by Viju Raghupathi and Wullianallur Raghupathi, School of Business, Brooklyn College, City University of Newyork, USA.

**Problem Statement:** The Objective of this research is to **escalate the health care expenditure and the number of hospitalizations.**

**Data Source:** The health data for the study was extracted from the Statewide Planning and Research Cooperative System (SPARCS) of the New York State Department of Health (NYSDOH). We extracted 200,000 patient records for the period 2009 to 2012. Data from such a large dataset is naturally characterized by a degree of incompleteness and fuzziness. The data includes the hospitalization indicators such as facility ID, procedure description, type of admission, patient disposition upon discharge, APR severity of illness, source of payment, and age group. The indicators for treatment result include length of stay and APR risk.

****

**Questions to answer:**

1. **State Machine Learning Algorithms that can be used to train the data.**

**Answer:** To analyse the healthcare data extracted from the SPARC dataset we can select algorithms based on the specific task and goal we want to achieve with our analysis.

a. **Logistic Regression:** Suitable for binary classification tasks, such as predicting hospitalizations or health outcomes.

b. **Random Forest:** Effective for both classification and regression tasks. It can handle complex interactions in the data.

c. **Gradient Boosting:** Algorithms like XGBoost, LightGBM, or CatBoost can be used for predictive modelling, especially when we want to optimize accuracy.

d. **Decision Trees:** Useful for interpretable models and feature selection.

e. **Support Vector Machines (SVM):** Good for binary classification tasks with a clear margin of separation.

**The choice of algorithm should be driven by the specific research questions and the nature of your dataset.**

1. **State the Hardware Requirements for your proposed model.**

**Answer:** The hardware requirements for training and running machine learning models can vary depending on the complexity of the model and the size of the dataset.

**a.** **CPU/GPU:** Most machine learning tasks can be run on a CPU, but for deep learning models, having access to a GPU (Graphics Processing Unit) can significantly speed up training. **High-performance GPUs from NVIDIA (e.g., GeForce or Tesla series) are commonly used.**

**b.** **RAM:** The amount of RAM required depends on the dataset size and model complexity. We might need several **gigabytes or even terabytes of RAM** for large datasets.

**c.** **Storage:** We need sufficient storage to store the dataset and model weights. **SSDs are preferred for faster data access.**

**d.** **Cloud Resources:** Cloud computing platforms like **AWS, Google Cloud, and Microsoft Azure offer** scalable resources for machine learning.

**e. Specialized Hardware:** For very large datasets and deep learning models, we might consider using specialized hardware such as **TPUs (Tensor Processing Units).**

1. **Can this Machine Learning Algorithm be transformed into AI? Justify.**

**Answer:** Machine learning algorithms can be considered a subset of artificial intelligence (AI). AI encompasses a broader field that aims to create systems and technologies that can perform tasks that typically require human intelligence. Machine learning is one of the key techniques used in AI to enable systems to learn and make predictions based on data.

We are using machine learning algorithms to **analyse health care data and make predictions about health care expenditure and hospitalizations.** ***This is a form of AI, as it involves the use of algorithms to make intelligent decisions based on data.*** The machine learning algorithms are a fundamental part of the AI system that we are building.

To summarize, the machine learning algorithms we are using are already a part of the AI framework. However, to create a more comprehensive AI system, we might consider integrating these machine learning models into a larger framework that includes other AI techniques, such as natural language processing (NLP), computer vision, or expert systems, to provide a more holistic solution for healthcare analytics and decision-making.

**Case Study 2:**

**Review Comments about Financial Product**

The data displayed is a collection of complaints about consumer financial products and services that we sent to companies for response. Complaints are published after the company responds, confirming a commercial relationship with the consumer, or after 15 days, whichever comes first.

**Data Description:**

The dataset comprises of Consumer Complaints on Financial products and we’ll see how to classify consumer complaints text into these categories: Debt collection, Consumer Loan, Mortgage, Credit card, Credit reporting, Student loan, Bank account or service, Payday loan, Money transfers, Other financial service, Prepaid card. Also we will try to identify the companies from the dataset.

**Questions to Answer:**

1. **What is the right algorithm to be used to predict product and issue using the complaints issue.**

**Answer:** To predict the product and issue categories based on consumer complaints text, we can use natural language processing (NLP) techniques and machine learning algorithms.

**a. Text Classification:** We can use text classification algorithms, such as **Multinomial Naive Bayes, Support Vector Machines (SVM), or deep learning models (e.g., recurrent neural networks, convolutional neural networks)** to classify the complaints into the appropriate product and issue categories.

**b. Pre-processing:** Pre-processing the text data by tokenizing, removing stop words, stemming or lemmatizing words, and converting text into numerical features (e.g., TF-IDF or word embedding like Word2Vec or FastText).

**c. Label Encoding:** Encoding the product and issue categories as numerical labels.

**d. Model Evaluation:** Using metrics like accuracy, precision, recall, and F1-score to evaluate the model's performance.

**e. Hyperparameter Tuning:** Fine-tuning the model parameters for better results.

**f. Cross-validation:** Implementing cross-validation to ensure model robustness.

The choice of the specific algorithm and approach can be determined through experimentation and performance evaluation on the dataset.

**2. Predict the complaints which cannot be resolved.**

**Answer:** Predicting whether a complaint can be resolved or not is a binary classification task. To achieve this, you can follow a similar approach as above.

**a. Data Labeling:** We need a labelled dataset where each complaint is marked as either "resolved" or "unresolved."

**b. Feature Extraction:** Using NLP techniques to extract features from the complaint text. These features might include the text content, the number of interactions with customer support, time taken to resolve, etc.

**c. Binary Classification Algorithm:** Train a binary classification algorithm like Logistic Regression, Random Forest, or a neural network to predict whether a complaint can be resolved or not.

**d. Model Evaluation:** Use standard binary classification metrics like accuracy, precision, recall, and F1-score to assess the model's performance.

**e. Deployment:** Implement the model into your workflow to predict whether incoming complaints are likely to be resolved or not.

**3. Represent the requirements for developing the AI?**

**Answer: Developing AI solutions for this task would require the following key components:**

**a. Data Collection:** Gathering a dataset of consumer complaints, including text data and labels for product, issue categories, and complaint resolution status.

**b. Data Pre-processing**: Cleaning and pre-processing the text data, including tokenization, stop-word removal, and encoding categorical labels.

**c. Feature Engineering:** Extracting relevant features from the text data and other metadata.

**d. Machine Learning Models:** Implementing text classification models for predicting product and issue categories and binary classification models for predicting complaint resolution.

**e. Model Evaluation:** Using appropriate metrics and cross-validation techniques to assess the models' performance.

**f. Hyperparameter Tuning:** Fine-tuning model parameters for optimal results.

g**. Deployment:** Integrating the trained models into a system or application where they can automatically categorize and assess incoming complaints.

**h. Continuous Monitoring:** Continuously update and retrain the models as new data becomes available to ensure their accuracy and relevance.

**i. User Interface:** Developing a user-friendly interface for users to input complaints and receive predictions.

**j. Legal and Ethical Considerations:** Ensuring compliance with data privacy regulations and ethical guidelines in handling consumer complaints.

**k. Data Security:** Implementing robust data security measures to protect sensitive consumer information.

**l. Scalability:** Designing the system to handle a large volume of complaints efficiently.

**m. Documentation:** Maintaining clear and comprehensive documentation for the AI system, including its functionality, architecture, and operation.

**This process would involve collaboration between data scientists, machine learning engineers, and domain experts to ensure the AI system meets the requirements and serves its intended purpose effectively and ethically.**

**Case Study 3**

**Stamp Verification**

An automatic system for stamp segmentation and further verification is needed especially for environments like insurance companies where a huge volume of documents is processed daily. However, detection of a general stamp is not a trivial task as it can have different shapes and colors and, moreover, it can be imprinted with a variable quality and rotation. This dataset was collected to help researchers build such a system.

**Dataset Description:**

This dataset contains 400 scanned document images. The documents are automatically generated invoices that were printed, stamped and scanned with 200 dpi resolution. They include color logos and color texts which makes the evaluation results more realistic. There are stamps of many different shapes and colors including black ones in the data set, sometimes the stamps are overlapped with signatures or a text. In some documents there are multiple stamps or none at all. The ground truth consists of binary images with masks of the stamp strokes which allows for accurate pixel-wise evaluation. This dataset contains the following folders, each with 400 items (one for each image):

**scans:** scans of the stamped genuine documents

**ground-truth-maps:** maps defining the region of the stamp(s)

**ground-truth-pixel:** pixel-level ground truth

**info:** contains text files with the info for each file. Each info file contains the following information:

**signature [0|1]:** signature present [0] or not [1]

**textOverlap [0|1]:** stamps overlap with printed text [1]

**numStamps [0|…|n]:** number of stamps on the page

**bwStamp[1|…|n]: stamp[1|…|n] is black stamp [1] or colored [1]**

**Source:** Micenkov, B., & van Beusekom, J. (2011, September). Stamp detection in color document images. In Document Analysis and Recognition (ICDAR), 2011 International Conference on(pp. 1125-1129). IEEE.

**Questions to Answer:**

1. **Which algorithm best suits to classify the stamps?**

To classify stamps in scanned document images with varying shapes, colors, and rotations, as described in the dataset, we can use a combination of computer vision and machine learning techniques.

**Convolutional Neural Networks (CNNs):**

Convolutional Neural Networks are particularly suited for image classification tasks.

**Steps:**

* 1. **Pre-processing:** Loading and pre-processing the scanned document images, including resizing, normalizing, and enhancing image quality. Preparing the binary masks (ground truth) for stamps, which provide pixel-level segmentation of stamp regions.
  2. **Data Augmentation:** Performing data augmentating to increase the diversity of the dataset. Techniques like rotation, scaling, and flipping can be used to handle variations in stamp orientation and positioning.
  3. **Model Architecture:** Designing a CNN architecture that can effectively extract features from the document images.
  4. **Training:** Training the CNN using the scanned images as input and their corresponding stamp masks as target labels. Using a loss function that measures pixel-wise accuracy or other relevant metrics if we have a pre-trained model (e.g., VGG, ResNet) consider transfer learning to improve classification accuracy.
  5. **Post-processing:** After classification, post-process the results to eliminate false positives or combine multiple stamps when necessary.
  6. **Evaluation:** Evaluate the model's performance using metrics such as pixel accuracy, intersection over union (IoU), and F1-score.
  7. **Deployment:** Implement the model in a system that can automatically segment and classify stamps in scanned documents.

1. **Give the Analytical Flow of implementation of the algorithm.**

**Answer:** The implementation flow for classifying stamps in scanned document images can be summarized as follows.

* 1. **Data Pre- processing:** Loading the scanned document images and their corresponding stamp masks. Performing image pre-processing, including resizing, normalization, and data augmentation.
  2. **Model Design:** Designing a Convolutional Neural Network (CNN) architecture for stamp classification. Consider using pre-trained models for feature extraction.
  3. **Data Split:** Split the dataset into training and testing sets.
  4. **Model Training:** Training the CNN using the training set, where input images are the scanned documents, and target labels are the stamp masks. Defining an appropriate loss function for pixel-wise classification.
  5. **Model Evaluation:** Using the testing set to evaluate the model's performance. Compute metrics such as pixel accuracy, IoU, and F1-score to assess classification accuracy.
  6. **Post-processing:** Implement post-processing steps to refine the classification results, such as eliminating false positives and merging overlapping stamps.
  7. **Model Deployment:** Implementing the trained model in an application or system that can process scanned documents, classify stamps, and create stamp segmentation masks.
  8. **Continuous Improvement:** Monitoring the system's performance and make necessary updates and improvements to the model and post-processing steps as new data becomes available.

The entire workflow is iterative and may require fine-tuning based on the dataset and real-world performance. Regular updates and maintenance are essential to ensure accurate stamp classification in a dynamic environment like an insurance company's document processing pipeline.

**Certificate of Project Completion**

This is to certify that

**Ajinkya Manohar Pawar**

has successfully completed the projects for BASIC CERTIFICATION COURSE IN ARTIFICIAL INTELLIGENCE.

This certificate is awarded in recognition of the dedication, hard work, and commitment demonstrated throughout the project.

Authorized Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

[Authority Name]

[CDAC-Pune]