Liis Siigur
Mia Männi
Robert Sarnet

Project D10: KAGGLE - PLANT DISEASES

Business Understanding

1. Identifying Your Business Goals

Background
Agriculture plays a vital role in sustaining life and supporting economies worldwide. It provides essential food and beverages while also contributing significantly to the GDP of many nations. However, plant diseases significantly affect agricultural productivity and food security worldwide by drastically reducing crop yields, causing economic losses, and leading to food shortages. This is particularly damaging in poorer countries where agriculture constitutes a larger share of GDP or where food insecurity is prevalent.

Timely detection of plant diseases is crucial to prevent their spread and minimize losses, however farmers and agricultural professionals often struggle to detect and diagnose diseases early, leading to reduced yields and increased costs. Moreover, small/scale growers and hobby gardeners, who may lack the expertise and resources to identify plant diseases, often find it challenging to manage their gardens effectively when plants exhibit signs of distress.

With advancements in machine learning and computer vision, it is now feasible to automate plant disease detection using image classification techniques to make advanced diagnostics more accessible and practical for everybody. This project aims to leverage such technologies to address this pressing issue.

Business Goals
The primary goal of this project is to develop an image classification model that accurately identifies plant diseases across multiple categories, including different plant species and specific diseases. This model will provide a scalable, efficient, and user-friendly tool for diagnosing plant diseases, helping farmers, especially individual and small-scale growers, make timely and informed decisions about disease management.

Business Success Criteria
The project will be considered successful if the following objectives are met:

1. Achieve a high accuracy (85% or more) in classifying plant diseases.
2. Provide an intuitive interface for end-users to upload images and receive disease diagnostics.
3. Reducing the time needed for diagnosis compared to traditional methods.

Liis Siigur
Mia Männi
Robert Sarnet

2. Assessing Your Situation

Inventory of Resources

- **Data:** A dataset containing labeled images of plant leaves, both healthy and with a disease, with categories such as Apple___Apple_scab, Blueberry___healthy, Corn___Cercospora_leaf_spot, and others. Dataset contains approximately 87,000 RGB images categorized into 38 classes, with predefined training and validation datasets, and a separate test set consisting of 33 different images.
- **Infrastructure:** Access to cloud-based computing resources for model training (e.g., GPUs).
- **Tools:** Python programming language, libraries like TensorFlow/Keras, and data analysis tools such as Pandas and NumPy.
- **Expertise:** Knowledge of deep learning techniques, computer vision, and plant pathology (one team member is majoring in biology).

Requirements, Assumptions, and Constraints

- **Requirements:** The model must support real-time or near-real-time predictions with high accuracy and low latency.
- **Assumptions:** The dataset is representative of the target use cases, with sufficient diversity in lighting, angles, and image quality.
- **Constraints:** Limited computational resources may affect the size and complexity of the model. Additionally, field implementation may face challenges like inconsistent internet connectivity.

Risks and Contingencies

- **Data Quality Risks:** Poor-quality images or imbalanced classes could reduce model performance. To mitigate this, data augmentation and balancing techniques will be employed.
- **Implementation Risks:** End-user acceptance may depend on the tool's accuracy and ease of use. Conducting user testing and iterative improvements can address these risks.
- **Time Risks**: Time constraints due to competing priorities from other projects and exams in different subjects. To mitigate the risk, it is good to organize and plan the week in advance, allocating dedicated time slots for project work to ensure consistent progress without neglecting other responsibilities.

Terminology

- **Image Classification:** Assigning labels to images based on their content.
- **Plant Diseases:** Conditions caused by pathogens like fungi, bacteria, viruses or insects that affect plant health.
- **Deep Learning:** A subset of machine learning focused on neural networks with multiple layers.
- **Augmentation**: Techniques to artificially increase the dataset by modifying images.

Liis Siigur
Mia Männi
Robert Sarnet

Costs and Benefits

- **Costs:** Initial time investment in model development, data acquisition, and infrastructure setup.
- **Benefits:** The project aims to significantly reduce the time and cost involved in detecting plant diseases. Improved crop yields, reduced economic losses, and enhanced decision-making for farmers, hobby gardeners and agricultural businesses.

3. Defining Your Data-Mining Goals

Data-Mining Goals
The core objective is to create a robust image classification model capable of identifying plant species and diseases across the provided categories with high accuracy. The model should generalize well to unseen data and provide confidence scores for its predictions.

Key deliverables include:

1. A trained and validated deep learning model.
2. Performance metrics, including accuracy, precision, recall, and F1-score.
3. A deployment-ready solution (e.g., web or mobile application).

Data-Mining Success Criteria
The success of the data-mining effort will be evaluated based on:

1. Model Accuracy: Achieving at least 85% accuracy across all classes.
2. Generalizability: The model performs well on a separate test set and real-world images.

Liis Siigur
Mia Männi
Robert Sarnet

Data Understanding

1. Gathering Data

Outline Data Requirements
The goal of this project is to develop a neural network-based image classifier for plant disease detection. The data must meet the following criteria:

- **Content:** High-quality labeled images of plant leaves, both healthy and diseased.
- **Representation:** A balanced dataset that includes diverse plant species and disease categories.
- **Format:** Images should be in a standard format (e.g., JPEG or PNG) suitable for preprocessing and neural network input.

Verify Data Availability
The dataset for this project has been sourced from Kaggle. It contains images of plant leaves categorized by species and health status, such as *Apple___Apple_scab*, *Corn___Cercospora_leaf_spot*, *Potato___Early_blight*, and others. Each image is labeled according to its class, making it ideal for supervised learning. However, some classes lack both healthy and diseased examples, as detailed below.

Define Selection Criteria
To ensure balanced representation of healthy and diseased plant categories, the following criteria were applied:

- Exclude categories with only healthy samples (Blueberry, Raspberry, Soybean).
- Exclude categories with only diseased samples (Orange, Squash).
- Include only categories that have both healthy and diseased samples.
- The dataset is limited to images of leaves, ensuring consistency in feature representation and eliminating extraneous variables like plant stems or fruits.

2. Describing Data

The dataset consists of the following key attributes:

- **Image Name:** A unique identifier for each image file.
- **Class Label:** The health status and species of the plant, e.g., *Apple___healthy*, *Tomato___Late_blight*.
- **Image Data:** Pixel data representing the visual content of the plant leaves.

Selected Categories:
The selected dataset includes the following plant species and their health statuses:

- Apple (*Apple___Apple_scab*, *Apple___Black_rot*, *Apple___Cedar_apple_rust*, *Apple___healthy*).
- Cherry (*Cherry_(including_sour)___Powdery_mildew*, *Cherry_(including_sour)___healthy*).

- Corn (*Corn_(maize)___Cercospora_leaf_spot*, *Corn_(maize)__Common_rust*, *Corn_(maize)___Northern_Leaf_Blight*, *Corn_(maize)___healthy*).
- Grape (*Grape___Black_rot*, *Grape___Esca_(Black_Measles)*, *Grape___Leaf_blight_(Isariopsis_Leaf_Spot)*, *Grape___healthy*).
- Peach (*Peach___Bacterial_spot*, *Peach___healthy*).
- Pepper (*Pepper,_bell___Bacterial_spot*, *Pepper,_bell___healthy*).
- Potato (*Potato___Early_blight*, *Potato___Late_blight*, *Potato___healthy*).
- Strawberry (*Strawberry___Leaf_scorch*, *Strawberry___healthy*).
- Tomato (*Tomato___Bacterial_spot*, *Tomato___Early_blight*, *Tomato___Late_blight*, *Tomato___healthy*).

3. Exploring Data

Class Distribution

A preliminary exploration revealed the number of images per class to be quite even, around 1600-2000 images of all classes. However, there are much more classes of different diseases of tomatoes than of other species - there are a total of 10 classes of tomato leaves, while other species have 2-4 classes. This might cause the model to classify plants as tomatoes too frequently.

Image Characteristics

- Image Resolution: The images have uniform resolution.
- Color Channels: All images are in RGB format, making them suitable for convolutional neural networks (CNNs).
- Background Noise: Most images seem to not include extraneous background elements.

Feature Analysis

The primary features in the images are the shape, texture, and color of the leaf, along with disease-specific symptoms such as spots, discoloration, or mildew.

Augmented data

Most categories in the dataset include augmented images. Augmentation involves techniques such as rotation and flipping along the X/Y axis.

Test set

Although the dataset contains a test set, the test set does not contain images to all categories.

4. Verifying Data Quality

Completeness

The dataset includes labeled images for all selected categories. All categories have approximately the same number of images.

Liis Siigur
Mia Männi
Robert Sarnet

Accuracy
The class labels appear accurate based on the dataset description provided on Kaggle. Random manual inspection of the images aligns with their respective labels.

Consistency
The dataset is consistent in terms of the image format and labeling schema, with no discrepancies observed during initial exploration.

Missing Data
There are no missing images or labels in the selected subset. Each image has a corresponding label.

Liis Siigur
Mia Männi
Robert Sarnet

Planning the Project

1. Task

- Discard images from the dataset that belong to unwanted categories.
- Choose suitable images from either the training or validation set, to replace missing images in the test set.
- Add data to the GitHub Repository.
- Task done by Robert. Estimated time to complete the task is 1-2h.

2. Task

- Connect folder names with images as classification labels.
- Task done by all team members. Estimated time to complete the task is 1h.

3. Task

- Develop and train the model.
- Task done by all team members. Estimated time for each team member is around 25h.

4. Task

- Evaluate and optimize the model
- Task done by all team members. Estimated time for each team member is around 5h.

5. Task

- Gather the results and create the final poster.
- Task done by all team members. Estimated time for each team member is around 3h.

Repository

https://github.com/robertsarnet/IDS-project