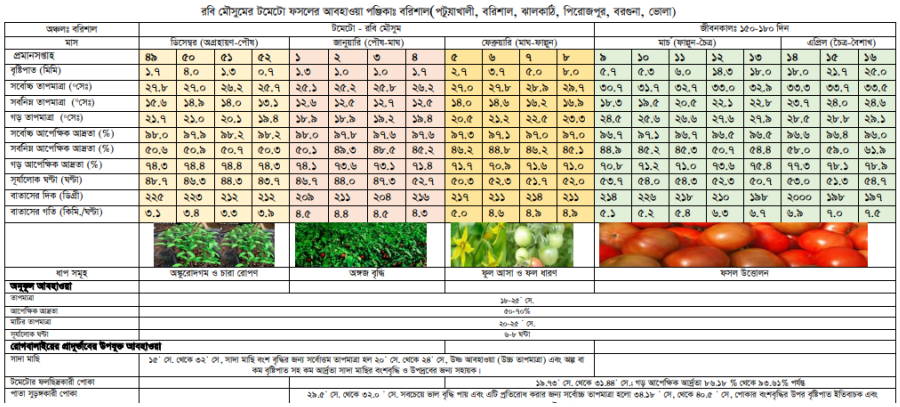
A Bangladesh-Specific Crop Recommendation System: Dataset Development and Insights

Dataset Creation:

- Why to create a new dataset?
  While reviewing the existing literature, I discovered that nearly all the research in this domain has been conducted using datasets tailored to Indian weather conditions [1]. Only one paper was identified that focused on Sri Lanka, leaving a significant gap in research specific to Bangladeshi agricultural conditions. This lack of localized datasets is a critical limitation, as climatic, soil, and environmental factors vary greatly between regions.
  To address this gap, I developed a new dataset specifically for Bangladesh. The goal is to ensure that the model recommendations are not only accurate but also contextually relevant to local farmers.

- Notable Works on the Indian Dataset
  One of the most significant works achieved an **accuracy of 99.55%** [2]. Other notable works can be explored through its citations.

- How the new dataset[3] has been made?
  Source of Data: Website of Bangladesh Agro-Meteorological Information Service (BAMIS) [4]
  BAMIS provides individual crop calendars for 41 crops across different regions. However, the data is presented in a discrete format rather than a consolidated form.
  Here is a sample crop calendar –



রবি মৌসুমের টমেটো ফসলের আবহাওয়া পঞ্জিকাঃ বরিশাল(পটুয়াখালী, বরিশাল, ঝালকাঠি, পিরোজপুর, বরগুনা, ভোলা)

অঞ্চলঃ বরিশাল	টমেটো - রবি মৌসুম																			
মাস	ডিসেম্বর (অগ্রহায়ণ-পৌষ)				জানুয়ারি (পৌষ-মাঘ)				ফেব্রুয়ারি (মাঘ-ফাল্গুন)				মার্চ (ফাল্গুন-চৈত্র)					এপ্রিল (চৈত্র-বৈশাখ)		
প্রমানসপ্তাহ	৪৯	৫০	৫১	৫২	১	২	৩	৪	৫	৬	৭	৮	৯	১০	১১	১২	১৩	১৪	১৫	১৬
বৃষ্টিপাত (মিমি)	১.৭	৪.০	১.৩	০.৭	১.৩	১.০	১.০	১.৭	২.৭	৩.৭	৫.০	৮.০	৫.৭	৫.৩	৬.০	১৪.৩	১৮.০	১৮.০	২১.৭	২৫.০
সর্বোচ্চ তাপমাত্রা (°সেঃ)	২৭.৮	২৭.০	২৬.২	২৫.৭	২৫.১	২৫.২	২৫.৮	২৬.২	২৭.০	২৭.৮	২৮.৯	২৯.৭	৩০.৭	৩১.৭	৩২.৭	৩৩.০	৩২.৯	৩৩.৩	৩৩.৭	৩৩.৫
সর্বনিম্ন তাপমাত্রা (°সেঃ)	১৫.৬	১৪.৯	১৪.০	১৩.১	১২.৬	১২.৫	১২.৭	১২.৫	১৪.০	১৪.৬	১৬.২	১৬.৯	১৮.৩	১৯.৫	২০.৫	২২.১	২২.৮	২৩.৭	২৪.০	২৪.৬
গড় তাপমাত্রা (°সেঃ)	২১.৭	২১.০	২০.১	১৯.৪	১৮.৯	১৮.৯	১৯.২	১৯.৪	২০.৫	২১.২	২২.৫	২৩.৩	২৪.৫	২৫.৬	২৬.৬	২৭.৬	২৭.৯	২৮.৫	২৮.৮	২৯.১
সর্বোচ্চ আপেক্ষিক আর্দ্রতা (%)	৯৮.০	৯৭.৯	৯৮.২	৯৮.২	৯৮.০	৯৭.৮	৯৭.৬	৯৭.৬	৯৭.৩	৯৭.১	৯৭.০	৯৭.০	৯৬.৭	৯৭.১	৯৬.৭	৯৬.৫	৯৬.৫	৯৬.৬	৯৬.৪	৯৬.০
সর্বনিম্ন আপেক্ষিক আর্দ্রতা (%)	৫০.৬	৫০.৯	৫০.৭	৫০.৩	৫০.১	৪৯.৩	৪৮.৫	৪৫.২	৪৬.২	৪৪.৮	৪৬.২	৪৫.১	৪৪.৯	৪৫.২	৪৫.৩	৫০.৭	৫৪.৪	৫৮.০	৫৯.০	৬১.৯
গড় আপেক্ষিক আর্দ্রতা (%)	৭৪.৩	৭৪.৪	৭৪.৪	৭৪.৩	৭৪.১	৭৩.৬	৭৩.১	৭১.৪	৭১.৭	৭০.৯	৭১.৬	৭১.০	৭০.৮	৭১.২	৭১.০	৭৩.৬	৭৫.৪	৭৭.৩	৭৮.১	৭৮.৯
সূর্যালোক ঘন্টা (ঘন্টা)	৪৮.৭	৪৬.৩	৪৪.৩	৪৩.৭	৪৬.৭	৪৪.০	৪৭.৩	৫২.৭	৫০.৩	৫২.৩	৫১.৭	৫২.০	৫৩.৭	৫৪.০	৫৪.৩	৫২.৩	৫০.৭	৫৩.০	৫১.৩	৫৪.৭
বাতাসের দিক (ডিগ্রী)	২২৫	২২৩	২১২	২১২	২০৯	২১১	২০৪	২১৬	২১৭	২১১	২১৪	২১১	২১৪	২২৬	২১৮	২১০	১৯৮	২০০০	১৯৮	১৯৭
বাতাসের গতি (কিমি./ঘন্টা)	৩.১	৩.৪	৩.৩	৩.৯	৪.৫	৪.৪	৪.৫	৪.৩	৫.০	৪.৬	৪.৯	৪.৯	৫.১	৫.২	৫.৪	৬.৩	৬.৭	৬.৯	৭.০	৭.৫
ধাপ সমূহ	অঙ্কুরোদগম ও চারা রোপণ				অঙ্গজ বৃদ্ধি				ফুল আসা ও ফল ধারণ				ফসল উত্তোলন							
অনুকূল আবহাওয়া																				
তাপমাত্রা	১৮-২৫° সে.																			
আপেক্ষিক আর্দ্রতা	৫০-৭০%																			
মাটির তাপমাত্রা	২০-২৫ ° সে.																			
সূর্যালোক ঘন্টা	৬-৮ ঘন্টা																			
রোপণপরবর্তীকের প্রাথমিক উপযুক্ত আবহাওয়া																				
সাদা মাটি	১৫° সে. থেকে ৩২° সে. সাদা মাটি বংশ বৃদ্ধির জন্য সর্বোত্তম তাপমাত্রা হল ২৯° সে. থেকে ২৪ সে., উচ্চ আবহাওয়া (উচ্চ তাপমাত্রা) এবং অল্প বা কম বৃষ্টিপাত সহ কম আর্দ্রতা সাদা মাছির বংশবৃদ্ধি ও উপদ্রবের জন্য সহায়ক।																			
টমেটোর ফলচিত্রকারী পোকা	১৯.৭৩° সে. থেকে ৩১.৪৪° সে., গড় আপেক্ষিক আর্দ্রতা ৮৬.১৮ % থেকে ৯৩.৩১% পর্যন্ত																			
পাতা সুড়ঙ্গকারী পোকা	২৯.৫° সে. থেকে ৩২.০ ° সে. সবচেয়ে ভাল বৃদ্ধি পায় এবং এটি প্রতিরোধ করার জন্য সর্বোচ্চ তাপমাত্রা হলো ৩৪.১৮ ° সে. থেকে ৪০.৫ ° সে., পোকার বংশবৃদ্ধির উপর বৃষ্টিপাত ইতিবাচক এবং																			

  To compile the dataset, each crop calendar was accessed on the BAMIS website. Key environmental parameters were identified and recorded, focusing on the planting stages of each crop. The collected fields include rainfall, mean temperature, average humidity (RH), sunshine hours, wind direction, and wind speed.

- Key Points Related to the dataset
  Rows: 374, each representing a unique crop-environment combination.
  Features:
  i. Agricultural Zone: 14 distinct zones covers 64 districts.
  ii. Month (Chara): Indicates the sowing month for each crop, accounting for seasonality.
  iii. Rainfall (mm): The required rainfall during the planting stage.
  iv. Temperature (Avg): The average temperature conducive to crop growth.
  v. Humidity (Avg): The average relative humidity suitable for the crop.

vi. Sunlight (Hours): The total hours of sunlight needed during early growth.
vii. Wind Direction (°): The dominant wind direction in degrees.
viii. Wind Velocity (km/h): The speed of the wind, relevant to crop stability and pollination.
ix. Label: The recommended crop under the specified environmental conditions.

Observations: Class-wise box plotting of features revealed very high variance.

Model Training

- Initial Preprocessing[Code1]

  Data Augmentation: To train the model effectively for all crop-environment combinations, the original dataset with 374 unique entries was augmented. For each existing combination, 20 new rows were generated by introducing small random variations to key environmental features.

  Rainfall (mm): ±3 mm variation using a uniform distribution.
  Temperature (Avg): ±2°C variation.
  Humidity (Avg): ±10% relative humidity variation.
  Sunlight (Hours): Kept unchanged as it's less prone to short-term fluctuations.
  Wind Direction (°): ±40° variation.
  Wind Velocity (km/h): ±5 km/h variation.
  Label: Retained from the original row to ensure consistency.

  This augmentation, guided by BAMIS-specified variations, significantly expanded the dataset and ensured the model could generalize effectively to diverse real-world conditions.

  Label Encoding(label), One Hot Encoding(Agricultural zone) and others

  Class wise Robust Scaling:
  Here is the implementation,

```
for cls in y.unique():
    scaler = RobustScaler()
    class_data = X[y == cls]
    scaled_class_data = scaler.fit_transform(class_data)
    scaled_class_df=pd.DataFrame(scaled_class_data,columns=X.columns,
index=class_data.index)
    scaled_data = pd.concat([scaled_data, scaled_class_df])
    scalers[cls] = scaler
```

  As you can see, there will be 41 distinct scalers for each class. This class wise scaling significantly boosts the model's accuracy.

- Algorithms

  Mostly Decision Tree based algorithms perform here with this dataset. A voting classifier, combining multiple tree-based models, achieved an impressive **99.55% accuracy** on the test set.

	Decision Tree	Random Forest	Gradient Boosting	HistGradientBoosting	Logistic Regression	SVM	XGBoost	KNN	CatBoost	LightGBM
Accuracy	0.977666	0.98995	0.985483	0.089336	0.21608	0.298157	0.989391	0.750419	0.997767	0.033501

- Challenges

  While taking a new set of data as input, which robust scaler will be used, as there are 41 scalers. I also can't find resources on internet to deal with it. I've implemented an idea but which actually

haven’t worked out[Code2].

Web Implementation

- As it was my 2-1 project, I’ve implemented the very initial model with 67% accuracy. The web actually take agricultural zone and district as input and then fetch weather data through API and recommend the class. [web]

[1] https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

[2] https://ieeexplore.ieee.org/document/10534479

[3] https://github.com/kefaet03/agriqo/blob/master/Data%20set/Agriqo(slider1)%20%20Sheet1.csv

[4] https://www.bamis.gov.bd/calendar/

[Code1] https://colab.research.google.com/drive/1qYSYa51DZgmhQTjm9jpP6ggvOGsm7K3C#scrollTo=WBMrF1QNI17_

[Code2] https://colab.research.google.com/drive/1D1LytB8dpJpEVcq8rzunpcBMvm2Nal37

[web] https://kefaet03.github.io/UniqoXAgriqo/index.html