# Title: Crop Yield Prediction

|  |  |  |
| --- | --- | --- |
| First name | Last Name | IIT Email |
| Amrutha | Gowda | [alakshmanegowda@hawk.iit.edu](mailto:alakshmanegowda@hawk.iit.edu) |
| Vishnu | Pajjuri | [vpajjuri1@hawk.iit.edu](mailto:vpajjuri1@hawk.iit.edu) |
| Aditya | Dubey | [adubey6@hawk.iit.edu](mailto:adubey6@hawk.iit.edu) |

1. **Introduction**

A new fast food chain is seeing rapid expansion over the past couple of years. They are now trying to optimize their supply chain to ensure that there are no shortages of ingredients. For this, they have tasked their data science team to come up with a model that could predict the output of each food processing farm over the next few years. These predictions could further increase the efficiency of their current supply chain management systems**.**

1. **Data Sets**

This problem has three datasets which contain farm data, train data, weather data and contain around 20000000 observations in it with 17 attributes and it is a mix between categorical and numeric values. Yield is dependent variable whereas rest are independent variables.

|  |  |
| --- | --- |
| Farm id | Unique farm ids |
| Date | Dates per hour from 2016 in train and from 2017 in test |
| Ingredient\_type | 4 types - w,x,y,z |
| Yield | Yield for each farm per hour |
| operations\_commencing\_year | Year the farm has started |
| num\_processing\_plants | processing plants present in the location/ farm |
| farm\_area | Area of the given farm |
| farming\_company | The company that owns the farm |
| deidentified\_location | Location of the farm |
| timestamp | Dates at which the weather was calculated at each hour |
| deidentified\_location | Location of the farm |
| temp\_obs | Temperature at that hour |
| cloudiness | Cloulds present in the sky at that hour |
| wind\_direction | The direction of the wind at the hour |
| dew\_temp | Dew temperature at the hour |
| pressure\_sea\_level | Pressure sea level at the hour |
| precipitation | Rainfall at the hour |
| wind\_speed | Wind speed at that hour |

1. **Research Problems**

* Explore the data and engineer new features
* Predict the yield for each farm crop.
* Given the forecasted demand for the next few months for a ingredient, device a strategy to source it

1. **Potential Solutions**

* Perform Pre-processing: Will inspect the data, perform data cleaning and necessary transformation as per the requirement
* By classification models – Decision Trees, Random Forest, XG Boost we can predict the yield of the crops and forecast the demands for upcoming months based on an ingredient.

1. **Evaluations**

Since our data set is large enough which is more than 5k hence we will go for Hold- Out evaluation.  The purpose of holdout evaluation is to test a model on different data to that from which it is  [learned](https://doi.org/10.1007/978-0-387-30164-8_830). This provides an unbiased estimate of learning performance, in contrast to  [in-sample evaluation](https://doi.org/10.1007/978-0-387-30164-8_405).

1. **Expected Outcomes**

Our expected outcome will be required crop for future months for each ingredient, which

Will help in increase the efficiency of current supply chain management system.