Agriculture Data Analytics

1 Background

Agriculture has been the key sector in all major economies and especially most important one for a country like India. Over a past couple of decades, technology has been revolutionized and much of the focus was shifted on to the technological advancement without much contribution to the agricultural domain. With the increasing population, the need for an increased amount of crop production can be observed. The annual food production of India was around 50 million tons in 1950-51 and touched 240 million tons by 2010-11. The demand for food grains is expected to reach 280 million tons by 2020. Various factors like newly emerging crop diseases, pests, untimely and excessive use of pesticides and fertilizers by farmers and extreme climatic conditions have made it difficult to meet the required needs. Considering all these problems, exploiting technological advancement to solve agriculture related problems seems to be of high importance. However, much of the technological advancement has not been translated successfully to agricultural domain. Nevertheless, there have been many successful attempts to deliver technology-enabled solutions to our farmers. During last decade, several research efforts were being made to build IT-based systems to analyze the agricultural information and provide necessary solutions for various problems that arise in agricultural domain. In 2004, an agricultural advisory system called eSagu was developed which aims at providing farm-specific agricultural expert advice to the farmers at regular. The esagu system requires several inputs including the images of the crop, weather details, soil conditions etc and by analyzing this information, the expert generates the advice. In this project, we are going to use the eSagu advice data and weather data as input to generate the required farm-specific advice.

2 Summary of the problem

This project proposes to analyze the data collected by eSagu. In particular, the idea is to analyze the advice data coming from farms along with weather conditioning and attribute set by domain experts. Eventually, the target is to learn the mapping between visual and non-visual data by employing existing mathematical models like Neural Network (NN) and topic modeling for developing a prescriptive analytic solution in agriculture domain.

3 Related work

In India, there were about 120 million farm holdings by 2005 and the number has been growing year by year. Estimates indicated that 60 per cent of farmers do not access any source of information for advanced agricultural technologies resulting in huge adoption gap. During last decade, several organizations and government departments have built web portals and there were research efforts to build agriculture information delivery systems like aAQUA, e-Krishi, AGRISNET, Kisan Call Centres, and Digital Green. "Esagu" was made at IIIT, Hyderabad.

In Esagu system, the agricultural experts receive latest information about the crop situation in the form of both image and non image data. Image data includes several pictures capturing several views of the farm and non image data includes details like the weather information, soil conditions, location of the farm etc. There will be a coordinator assigned to a specific region and he collects the above mentioned information from various farms and passes this information, along with observation form signed by him, to the agricultural expert. By analyzing the non image data, the agricultural expert identifies the problem, suggests a solution and passes the advice to a higher level expert, who in turn, verifies the solution, makes changes if necessary and finally approves the advice which is delivered to the farmer.

The idea of this project is to minimize the human effort in generating the required advice i.e, to automatize the process to the possible extent. The known information about the crops and various problems that may arise under specific conditions depending upon the crop species, season, locality of the farm, type of the soil, rainfall etc.. is used to predict the possible cause of the problem as soon as the information regarding the farm is received. If the database is unable to find a matching cause i.e, the information about the problem is not in our database yet, the database will be updated and this new cause will be added to the list. After each advice, feedback is collected from the respective farms and will be used to measure the accuracy of the system and modify it whenever required.

Whenever the data about a farm is provided to the esagu system, it suggests the most possible cause of the problem to the Agricultural expert who in turn verifies it and passes it on to the higher level expert.

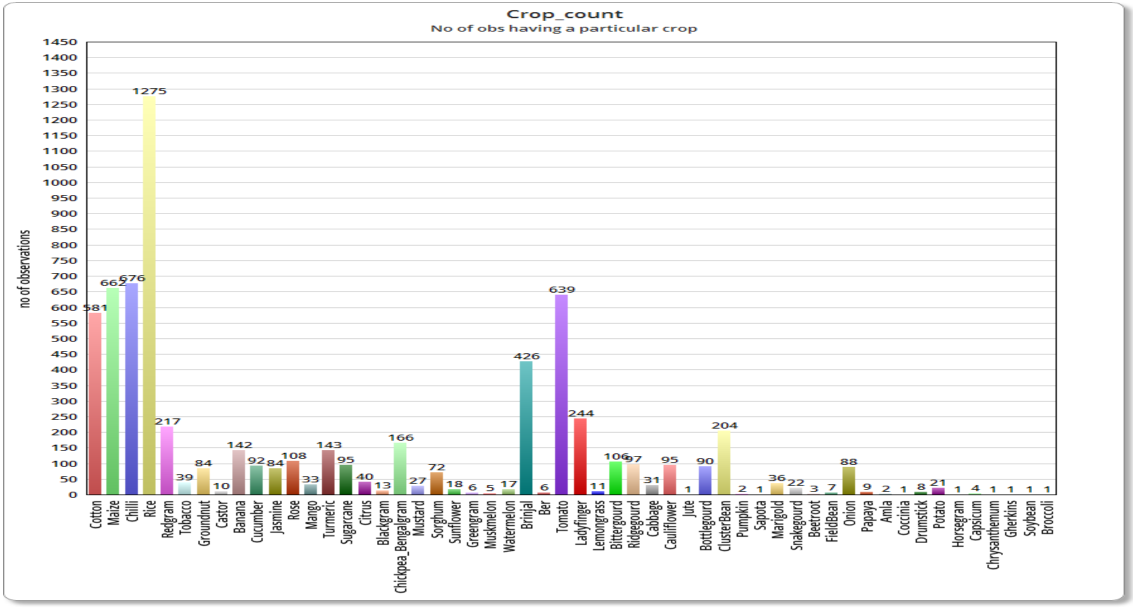
4 Esagu advice data

The advice data generated by eSagu contains four fields. First one is the date of advice. The next field contains the farm id, which gives us all the details regarding the crop including its location, crop id and date of sowing. The third field contains the species of the crop and the final field contains the main advices provided by the agricultural experts. Our project uses this advice data along with the weather details of the farm locations and tries to find the correlation between weather and the crop diseases.

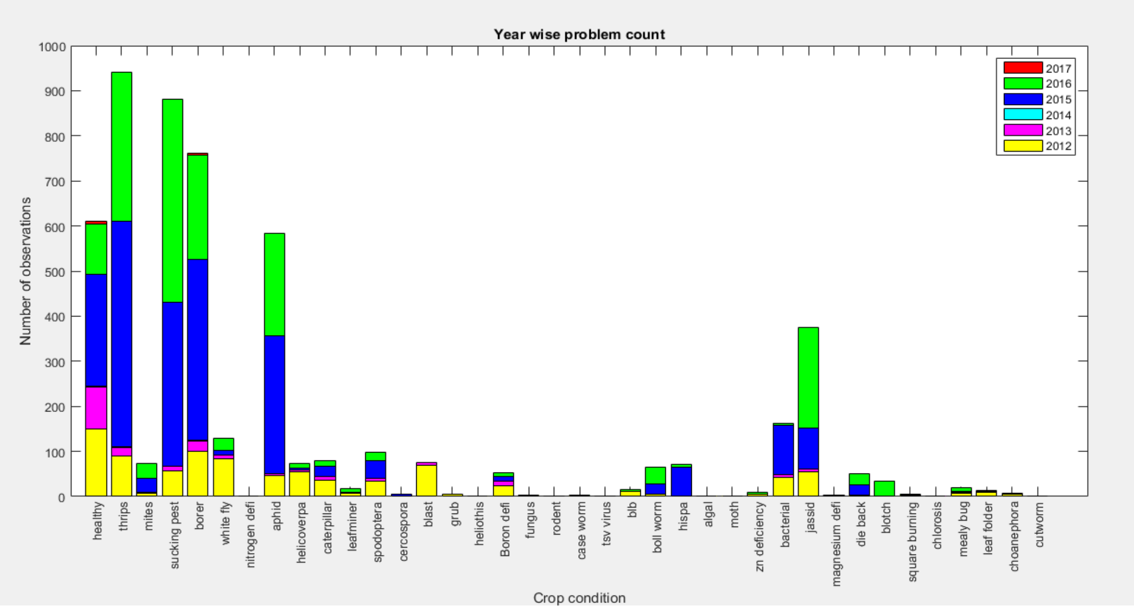
4.1 Data preprocessing

Since the advice data generated by eSagu contains the disease name along with the advice, we used this advice data to identify the possible diseases. Keywords corresponding to diseases and pesticides/insecticides were manually identified and their occurrences were counted. The data we received corresponds to 5 years(2012-16). The years 2015 and 2016 account for 90% of the total observations while the year 2014 had almost zero observations. Also, since the data is not uniform across the crops, a few of the crop problems account for most of the observations. Totally, there are 39 different crop problems found in the advice data. Out of these 39 different problems, only 19 problems had the total count more than 5 over the 5 years. Hence, we only use these 19 problems as the output classes for our prediction model.

Figure 1: Total crop count

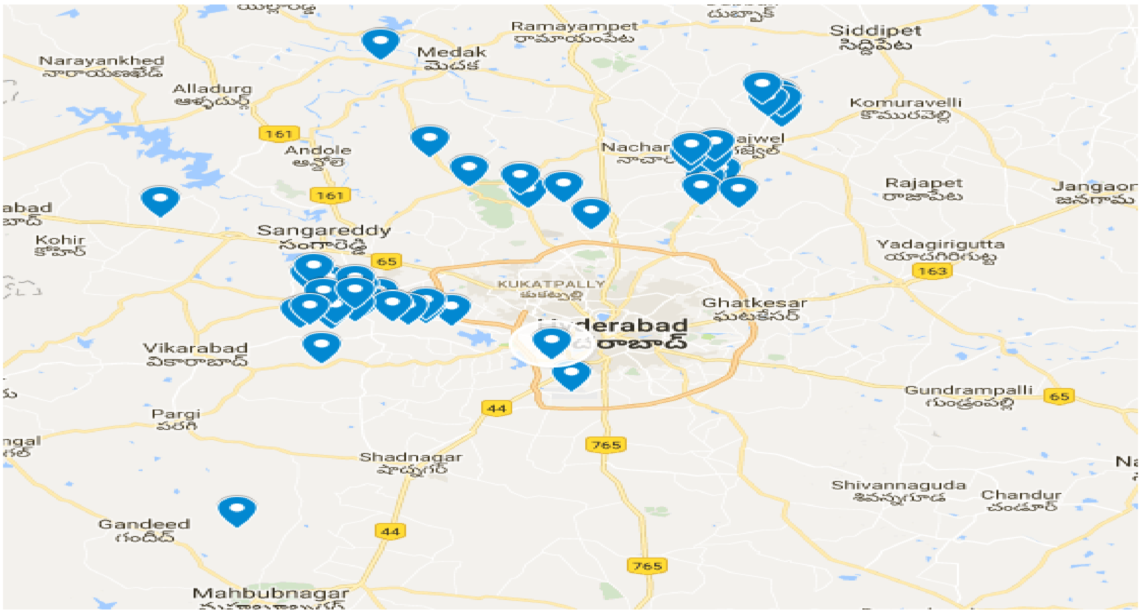


4.1.1 Yearwise disease count



The advice data we processed is sparse and that alone is not enough to model the dynamics of crop problems. We also need some information regarding weather and soil condition of the crops. We were able to identify the geo-coordinates of the farm locations based on the village names from the farm id. We have found that all the eSagu registered village names we received are concentrated around Rajendranagar mandal. So we are using Rajendranagar weather data for our prediction model, assuming identical weather conditions across all the farms.

4.2 Distribution of eSagu registered farms



5 Correlation with weather

Several plots were made in order to identify any significant correlation with weather parameters. It is observed there is a considerable dependency between each of the weather parameters and the crop diseases. The season in which the crop is grown and the age of the crop are also found to have an effect on the crop condition. Based on the season they are grown, the crops in India have been divided into Kharif, Rabi and Summer crops. The Kharif cropping season is from July to October

Parameter vector

Obtained from crop-id			Weather Parameters							
Species of the crop	Month	Age of the crop	Max. T(C)	Min. T(C)	RH1(%)	RH2(%)	WS(kmph)	RF(mm)	SS(hours)	EVP(mm)

and the Rabi season is from November to February. Crops grown from March to June are Summer crops. Since certain species of crops are grown in certain seasons, they are prone to the respective seasonal diseases/pests. Also, there will be a significant change in the weather conditions as the year progresses and the varying weather factors have a significant impact on the pattern of crop diseases . Hence, all these parameters are to be taken into account while building a prediction model.

The weather data we received contains day wise average values for the 8 weather parameters - Maximum temperature(C), Minimum temperature(C), Morning Relative Humidity(RH1 %), Evening Relative Humidity(RH2 %), Wind speed(Kmph), Rainfall(mm), Sunshine hours(hours) and Evaporation(mm). Along with these 8 weather parameters, three more parameters corresponding to the details of age of the plant, its species and the seasonal information(the month of reporting) are also provided as input to the model. The month will be known already and the type of crop will be provided in the crop-id from advisory data generated by eSagu. The age of the plant can be calculated once we know its date of sowing, which is included in the crop-id. Thus, information about 11 different parameters, that tend to affect the dynamics of crop problems, is available and hence, a 11 length vector containing the aforementioned parameters is generated. We will be considering this 11 length vector as the input for our model through out the rest of the project.

Based on the aforementioned 11 input parameters, the following classifiers are applied to find any correlation between the crop diseases and the parameters.

5.1 Bayes classifier

We have built a Bayes estimator that predicts the output problem based on the input crop details and weather information. Since the season in which the crop is grown plays a major role in determining the potential diseases/pests that may affect the crop, the probabilities of occurrence of different crop problems are calculated for each of the 3 cropping seasons. And also, as certain diseases are limited to certain crop species, the probabilities of occurrence of different problems per crop are also calculated. Once the season and the crop species is fixed, our Bayes estimator predicts the output problem, given the input parameters(weather details and crop information), by maximizing the posterior probability P(problem/factors). The following steps describe how our Bayes estimator was built.

- The given advices are divided into 3 parts of equal size. One of the three parts is used for evaluating the performance and the remaining two parts are used for calculating the required conditional probabilities. The following steps are repeated thrice using each of the 3 parts for performance evaluation.
- For each of the 3 seasons and for each of the crop species, we find the conditional probabilities p(factor/problem) for all the 11 input factors, for all the 19 different problems. This is done for 3 seasons and we get 3 different sets of probabilities for each season.
- Given the input, we first fix the season and the crop species. Then for each of the 19 output classes, we find the posterior probabilities p(problem/factor) using the formula p(problem/factor) = p(factor/problem)*p(problem)/p(factor).
- While finding the conditional probabilities for weather parameters, we first gave a count of 1, instead of zero, for all different possible values. Then we did interpolation across the values axis thereby filling the missing entries. Finally, we normalized the counts across the values axis to get the probabilities.

The accuracies obtained for each of the 3 folds are 53.21%,52.90% and 57.65% respectively. The overall 3fold cross validation accuracy obtained is 54.58%. When any one of the top 3 predictions is compared with the actual output, the accuracy obtained is 80%.

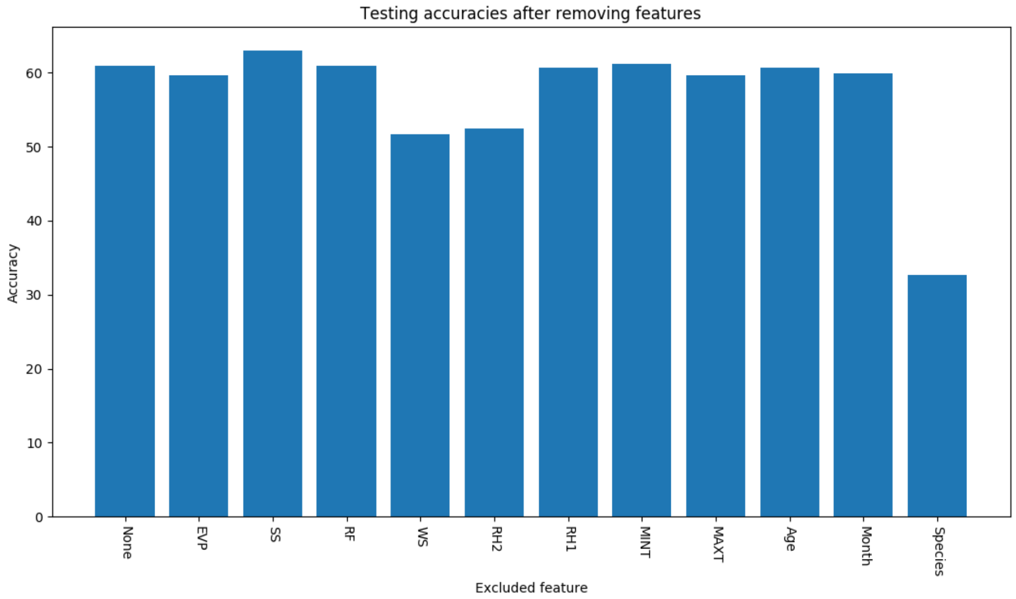
5.2 Neural networks

The problem with our Bayesian classifier is that it assumes independence between all the input factors. But in reality, the weather parameters are not independent. The prediction accuracy obtained using Bayes Classifier is not improving beyond 80%. So, we considered neural networks as our next classification option. Since there are only 19 problems that have count greater than or equal to 5. We have considered only these 19 problems and implemented neural networks with output number of classes as 19. There are 8 weather parameters and 11 input parameters in total when the crop species, month of observation and the age of the crop are included. Hence, the number of input neurons is 11 and the number of output neurons is 19. Several neural network architectures are considered with varying number of hidden layers, number of neurons in each hidden layer, learning rate and with different activation functions.

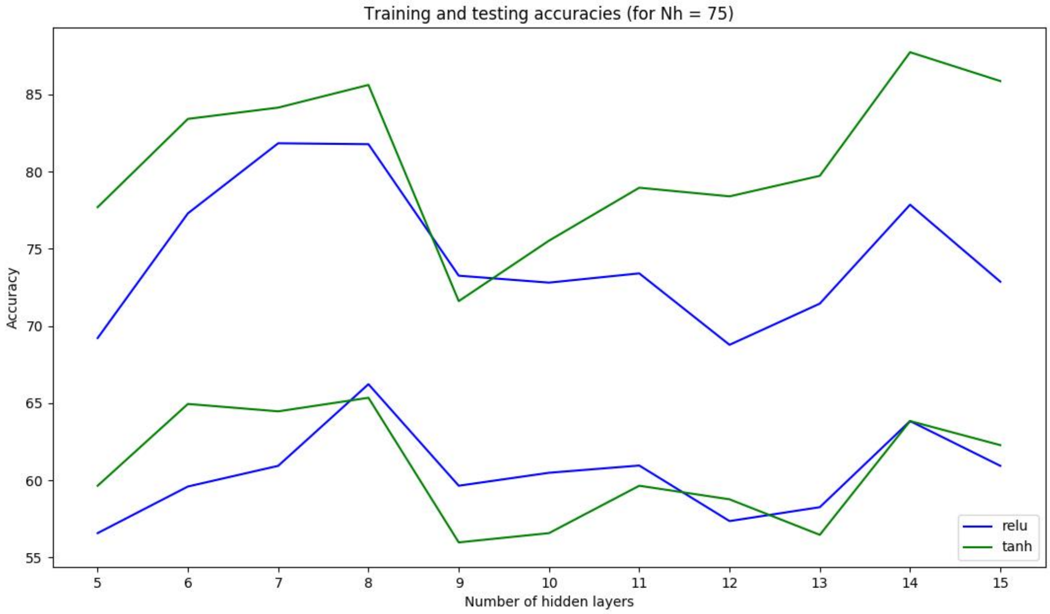
5.3 Dependency on the input parameters

We have also tried removing each feature among the 11 input features and found out that the testing accuracy drops significantly if the crop species parameters is excluded. It implies that the dynamics of crop problems have a significant dependency on the type of crop, as there will be some diseases which occur in only some types of crops. Though there is no change in accuracy if one of the remaining features is excluded, it doesn't mean that they do not have any impact on the crop condition. Since there will be a considerable dependency among the weather parameters, exclusion of one parameter may not have much impact on model prediction. As there is no significant improvement in the testing accuracy when other features are excluded, we have considered all the 11 input parameters for our model.

Figure 2: Testing accuracies obtained after removing each feature



Maximum accuracy achieved with neural networks is 66%.The corresponding number of hidden layers is 8, each layer having 75 neurons. We used 70% of input data for training the model and 30% of input data for testing. Since our model predicts the top 3 possible problems, we have used another evaluation metric which counts as a success if the actual problem is among the top 3 predicted problems. When this type of evaluation is considered, the accuracy of the system is 85%.



5.4 Meta classifier

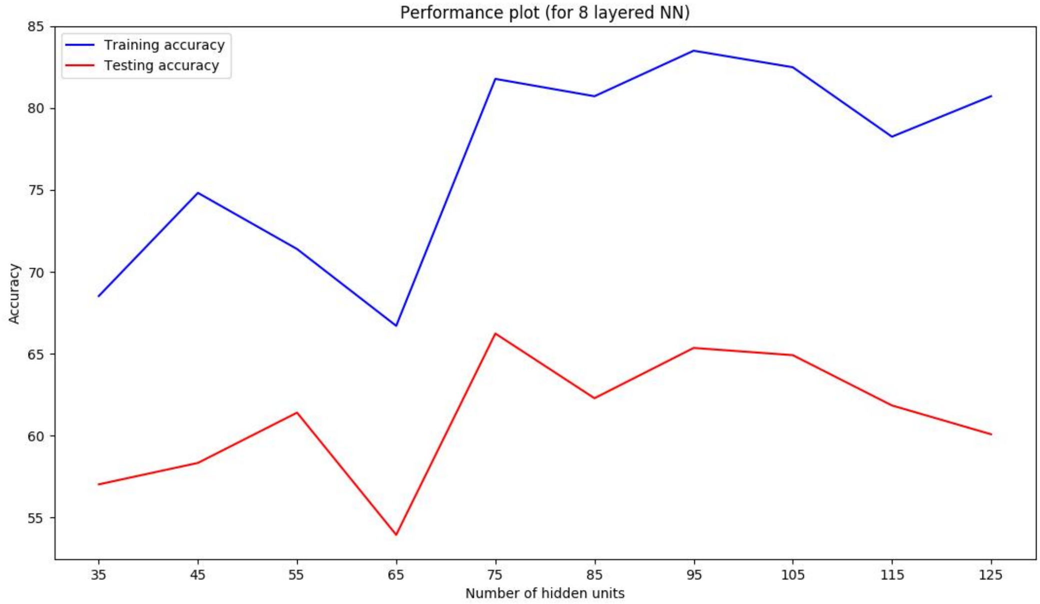
When we observed the corresponding training and testing accuracies for each of the 19 classes, we found that 12 of them had very high accuracies while the remaining 7 had very low accuracies. Then we applied neural networks for these 2 groups separately and obtained testing accuracies of 72% for both the groups, which rose to 93% when top 3 predicted outputs are used for evaluation. Based on this, we first divided the 19 classes into 2 groups and applied a meta classifier that uses 2 different sets of weights for the 2 classes. 2 different 7 layered neural networks are trained for these 2 classes separately and another classifier is built for classifying the input across these 2 samples. We have tried this for varying number of layers and activation functions. The final 3 fold cross validation accuracy obtained using the meta classifier is 77.64%. The 3 fold accuracy when top 3 outputs are considered is 87.92%.

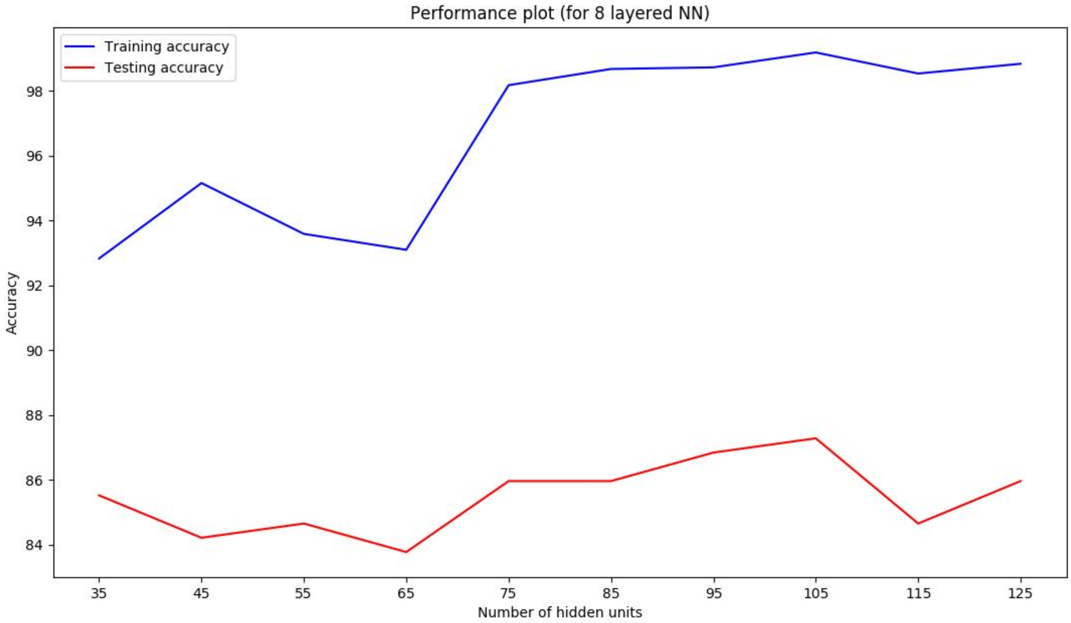
5.5 SVM

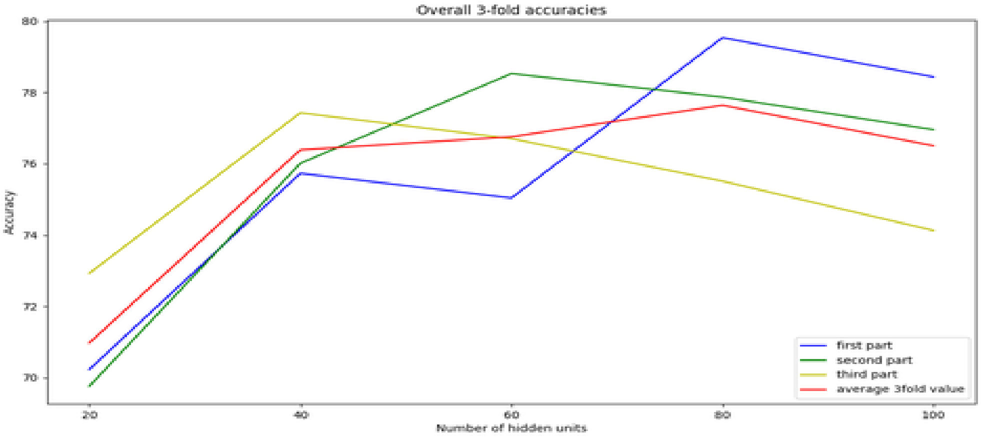
Since it is a classification problem we have also built a SVM on this data. First we have fit a multiclass SVM on the entire data. The best parameters are found to be as follows: Kernel - polynomial,degree - 5, penalty term - 1, max iterations - 9000 and this multiclass classifiers was built on the basis one versus one strategy i.e. for every pair of class a binary classifier is built. On the initial data and with top 3 strategy the maximum accuracy achieved was 92% both using 3 fold and 4 fold cross validation. The accuracy with top 2 strategy was 86% using both 3 fold and 4 fold. After this to further improvise the model individual class wise accuracy was found. For those classes with low testing accuracies as compared to that of training accuracy a new model has been built.

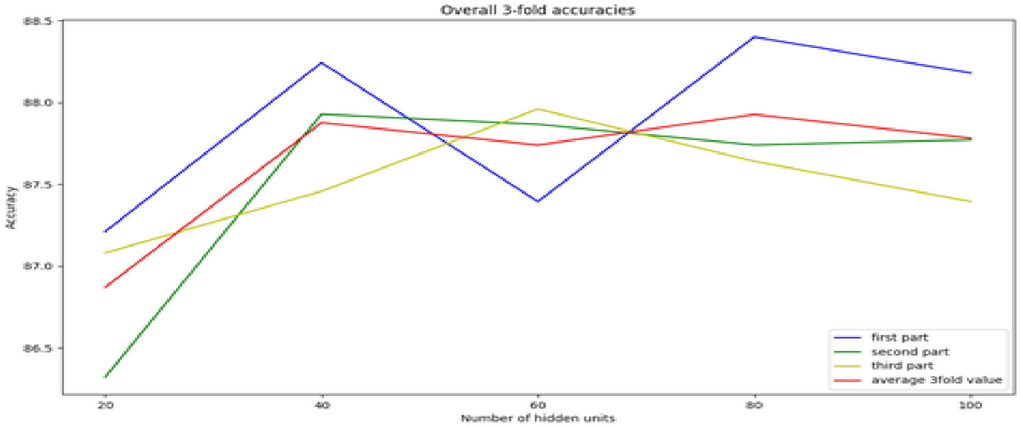
So this new classifier is comprised of mutiple SVMs. The a binary classifier segregates the data into two parts. These two parts contain data corresponding to more than one class and these two data parts are fed into two mutliclass SVMs. The two parts are segregated in such a way in order to maximize accuracies of individual multiclass SVMs. So the accuracy for such a model was 93% only. So it was not a great increase from 92% as we expect more increase as you are using a new SVM for some classes. So we have built a model on new strategy i.e. ensure that both the classes get approximately equal amount of data and also classes with more data points than 100 have to be equally balanced and also the one with less than 100 in such a way that approximately same amount of data is feeded into both the SVMs. So this way the testing accuracy reached around 97%. The logicality of this approach can be justified by the fact that neither of the models are getting overfit or underfit and the class distribution has also been done. All the above accuracies are for 2 fold cross validation.

6 Results


Performance plot (for 8 layered NN)
Training accuracy
Testing accuracy
Accuracy
85
80
75
70
65
60
55
35
45
55
65
75
85
95
105
115
125
Number of hidden units


Performance plot (for 8 layered NN)
Training accuracy
Testing accuracy
Accuracy
98
96
94
92
90
88
86
84
35
45
55
65
75
85
95
105
115
125
Number of hidden units


Overall 3-fold accuracies
Accuracy
80
78
76
74
72
70
20
40
60
80
100
Number of hidden units
first part
second part
third part
average 3fold value



7 Conclusion

In this project, we have reported the results related to the dynamics of crop problems obtained after applying 3 different classification techniques by combining the advice data set generated under eSagu prototype and weather data corresponding to Rajendranagar mandal, around which, the majority of farms are distributed. Although it is difficult to identify the crop diseases solely based on the weather data, a significant correlation can be observed once the other factors like the seasonal information, species of the crop and the age of the plant are taken into account along with the weather information. Also, this whole project assumed same weather conditions across all the farms and tried to correlate the pattern of crop problems with weather parameters. If it is possible to obtain the weather data corresponding to all the villages containing the farms and is considered for respective farms instead of the same weather data for all the farms, the prediction model will be more accurate and it would provide solution to a major problem in agriculture domain.

8 References

- http://onlinelibrary.wiley.com/doi/10.1002/wea.6080570507/pdf
- http://insait.in/AIPA2012/articles/005.pdf
- http://insait.in/AIPA2012/articles/009.pdf
- http://www.cst.ecnu.edu.cn/~slsun/pubs/MVML_final.pdf
- https://www.cia.gov/library/readingroom/docs/DOC_0000380263.pdf