EAS 508 Exam 1

Name: Venkata Satya Surya Sai Vineet Atyam

UB Person number: 50419767

UBIT Name: vatyam

PART A:

1.) Assumptions for Linear Regression

Linearity of the data:
The relationship between the predictor (x) and the outcome (y) is assumed to be linear.
Normality of residuals:
The residual errors are assumed to be normally distributed.
Homogeneity of residuals variance:
The residuals are assumed to have a constant variance (homoscedasticity).
Independence of residuals error terms:
There should be no relation between consecutive residuals.

2.) Multicollinearity is the presence of high correlations between two or more descriptors in a data. It results in a less reliable statistical inference.

Ways to deal with multicollinearity:

- Identify collinear variables and then remove all the variables except one.
- It can also be dealt with by combining two or more collinear variable into one.
- Perform an analysis designed for highly correlated variables, such as principal components analysis or lasso and ridge regression techniques to eliminate unwanted features.

3.) When a model is made from few features and is too simple, it may be underfitted and would have high variance and low bias as it is more generalized, whereas when more features are included, since it has more parameters, there may be situation where the model is over fitted for the training data, and would not work as expected for the unseen data, which is low variance and high bias.

This is the bias-variance trade-off, and hence we need a model that balances both the bias and variance for it to work even in unseen data as an optimal balance of bias and variance would never overfit or underfit the model.

4.) Bagging: Bagging (Bootstrap Aggregation) is the application of the Bootstrap procedure to a high-variance machine learning algorithm, typically decision trees. The objective is to create several subsets of data from training sample chosen randomly with replacement. Each collection of the created subsets is used to train the model to get a generalized result.

An average over all the predictions from the model trees are used, to provide a more robust model than using a single decision tree.

Boosting: Boosting is a sequential process, where each subsequent model, tries to correct the errors of the previous model. The successive models are dependent on the previous models. It is done to improve the accuracy of the next tree based on the prior tree errors. Each of the trees can be small, with few terminal nodes. The shrinkage parameter also helps in slowing the process down even further to create many different shaped trees to solve the residuals, and provide a more accurate model. It is very useful when there is huge data present and the decisions trees are expected to be complex.

5.) Overfitting: Overfitting refers to a model fitting the training data too well. This occurs when a model learns the noise and the details of training set to the extent that it negatively affects the performance of the model on the unseen data. This leads to the model not working for the test data or any unseen data as the concepts of the training model do not apply to the test data and hence cannot be a generalized model. Overfitting can be prevented by limiting the features or parameters or penalising the model when it includes more features to make it a more robust model. Low bias and high variance are good indicators of overfitting.

Underfitting: A statistical model is said to be underfit when it cannot capture the underlying trend of the data and hence destroys the accuracy of the model. It means that the model does not fit the data well enough. This may happen when there are less descriptors or when there is very less training data due to which the model will not be able to identify the dominant trend and hence is a very bad generalization. High bias and less Variance are good indicators of underfitting.

6.) Support Vectors are the data points that are closer to the hyperplane and influence the position and the orientation of the hyperplane. The support vectors are used to maximize the margin of the classifier, these are the points that help us build the support vector machine and are the most critical elements of the training set.

7.) Non-linear data is difficult to separate using a linear hyperplane, hence we need to apply a function that transforms the data such that it becomes linearly separable. Kernel trick allows us to operate in the original feature space without computing the coordinates in a higher dimensional space. As in a higher dimension we can find a space which can linearly separate the data by a given hyperplane.

Usage of the Kernel Trick:

It is used in case of Gauss Process Regression or also in case of Support Vector machine to find the linear separator clearly in a higher dimension.

In case of Gauss Process, the Gauss Kernel takes into account the probability distribution, by having the weighting function of all possible observations. By applying the kernel in a higher dimension, we can efficiently capture a separation of data and convert the random connections into an ordered connection.

Kernel trick can also be used in ridge regression (kernel ridge regression) so that it helps in dealing with non-linear data.

SVR defines the kernel and then puts in a margin of variance off of that for prediction, with the accuracy determined by how far off from that margin, this is done in a higher dimension to define a separable hyperplane based on it.

8.) Need to reduce the number of descriptors:

There may be lot of descriptors with high correlation between them and hence gives us incorrect results while interpreting the trained model.

It is computationally expensive to include all the features, it is better to rather remove the less important features and retaining the most important features. It is also difficult to explain the models interpretability with the increase in number of descriptors.

The presence of right descriptors helps in building a more robust and accurate model as compared to a model with more features which may result in overfitting.

To explain as much as data variation as possible while discarding the highly correlated variables.

Ways to minimize the number of descriptors

1) VIP Analysis and the PCA:

Important features for the target prediction can be found by calculating the correlation between the weights and the target property to identify the most import features and reduce the dimensions. Since eigen vectors are calculated in the Principal Component Analysis, new axes are defined which capture more information and thus we can use fewer dimensions while losing minimal amounts of information.

2) Lasso Regression can be used for feature selection as it provides the exact features used for regression and can be perfectly used for minimizing the descriptors as the coefficients of non-selected features turn out to be 0. Ridge regression on the other hand provides coefficients for all descriptors, and hence it cannot be used for minimizing the descriptors.

PART B:

For Property1 Using Regression Analysis:

MODEL	TRAIN R-SQUARED	TEST R-SQUARED	RMSE TRAIN	RMSE TEST
GAUSSIAN PROCESS REGRESSION	0.8486822	0.8402881	0.2038774	0.2162202
SUPPORT VECTOR REGRESSION	0.8921201	0.2125141	0.158546	0.8340677
RANDOM FOREST REGRESSION	0.9145271	0.9554783	0.108974	0.0899742

Accuracy:

The model accuracy is good for regression in all the models used which are Gaussian Process Regression, Support Vector Regression, Random Forest Regression for the training set.

However, the models Gaussian Process and Random Forest fared well in predicting the test set, whereas the Support vector model did not, leading to think that the Support vector has been overfit to the training data and hence cannot be used for unseen data analysis.

Specificity:

The models had nearly the same R-squared values and RMSE values when data was split in different proportions and also in random sampling, indicating that the models are pretty generalized and can help in predicting the unseen tests to a certain confidence. This holds true for Gaussian and Random Forest, but not for Support Vector

Sensitivity:

Feature1 13.447025, Feature2 5.737678, Feature3 2.587802,

Feature4 5.525100, Feature5 1.164147, Feature6 4.518099,

Feature7 3.875395, Feature8 6.286060, Feature9 17.831905,

Feature10 1.414930, Feature11 12.545221, Feature12 17.673573

It seems that feature12, feature11 and feature9 have higher importance and hence when these values change the model results vary more, hence the model is sensitive to these features the most (From Random Forest Regression).

I believe the best model to be deployed for regression in for the provided data would be the Random Forest Regression as it is more robust and has better metrics.

Predicted Values for the first 10 rows of Property1 using Random Forest:

1	2	3	4	5	6
0.6531179,	0.3040662,	0.1391103,	4.3285543,	1.7359782,	0.2063053,

7	8	9	10
0.1502779,	0.3067506,	0.4071374,	1.2278224

For Property1 using Classification Algorithms (Bagging, Boosting, Random Forest and Logistic Regression)

MODEL	TRAIN ACCURACY	TEST ACCURACY
BAGGING	0.954416	0.9393939
RANDOM FOREST	0.9547009	0.9545455
BOOSTING	0.9581197	0.969697
LOGISTIC REGRESSION	0.8409091	0.6060606

The values in property1 have been converted to categorical values based on the median data(median = 0.7) where value > 0.7 is set as ‘Yes’ and else it is a ‘No’

Accuracy:

All the models can be used for predicting the unseen values as each model has high training and test accuracy. Since the logistic regression computes the probabilities, there is less accuracy in case of this.

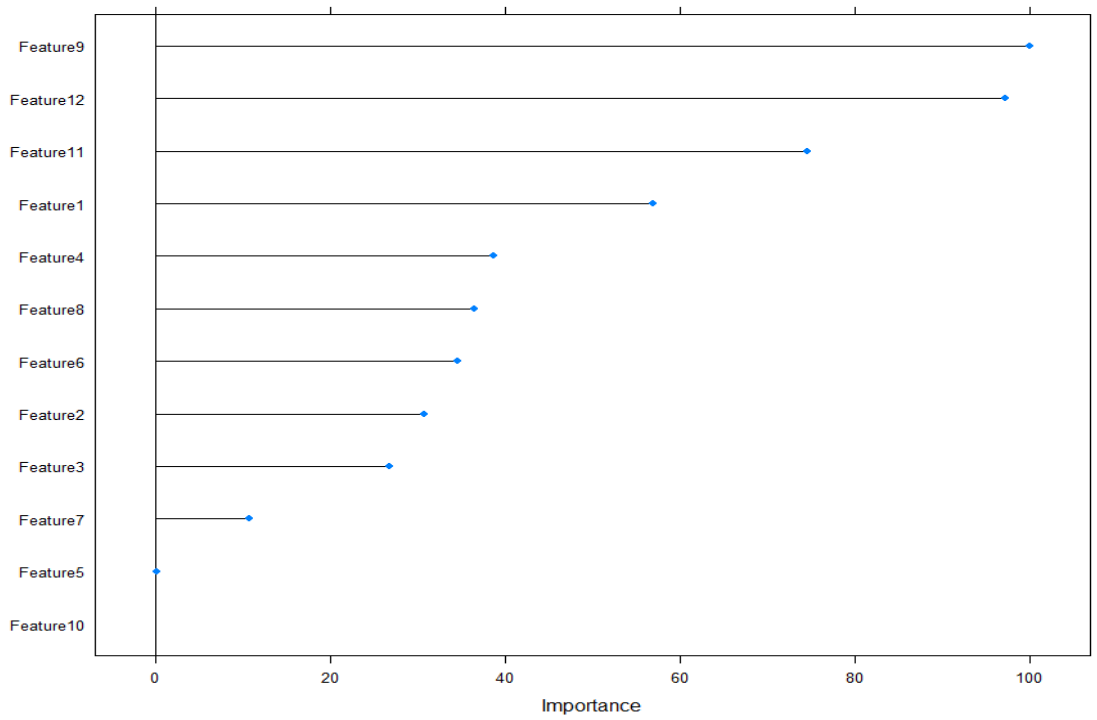
Specificity:

Since the models have high test and training accuracies, it seems that the models are robust and are having similar accuracies with different data splits and cross validations, hence the models are generalized to work for very different unseen datas.

Sensitivity:

Different algorithms have different importance of features and hence depends on the algorithm used. Here is the Sensitivity analysis for each model:

Bagging:



Treebag variable importance

Feature9 100.0000, Feature12 97.2526, Feature11 74.5644, Feature1 56.8760,

Feature4 38.6394, Feature8 36.4842, Feature6 34.4968, Feature2 30.7151,

Feature3 26.7837, Feature7 10.7638, Feature5 0.1666, Feature10 0.0000

For this model, it seems that Feature9 and Feature12 are more important and hence with change in these features would change the predicting ability of the model.

Random Forest:

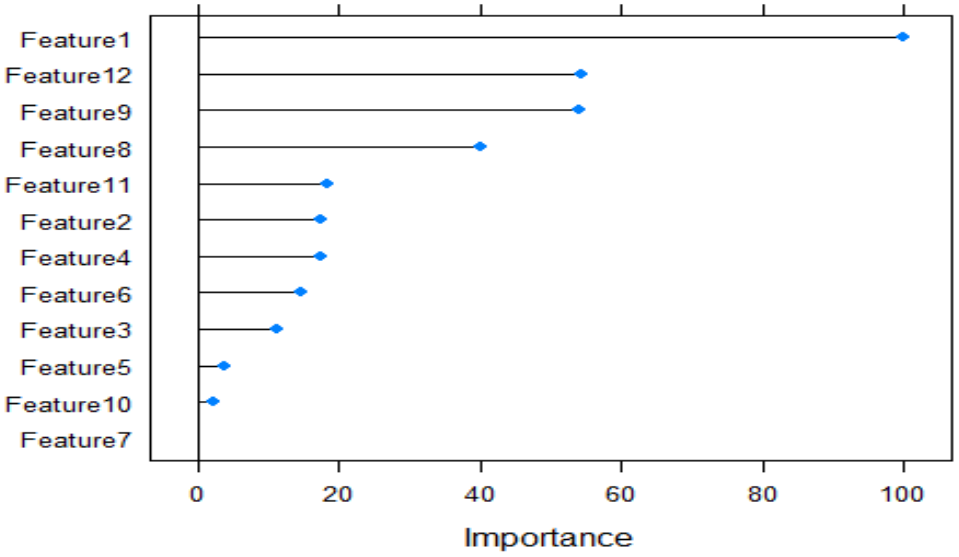
rf variable importance

Feature1 100.000, Feature12 54.362, Feature9 54.175, Feature8 39.932,

Feature11 18.284, Feature2 17.459, Feature4 17.455, Feature6 14.628,

Feature3 11.003, Feature5 3.576, Feature10 2.164, Feature7 0.000

The features 1, 12 and 9 are more important for the Random Forest Classification, here which will change.



Boosting:

Boosting variable importance

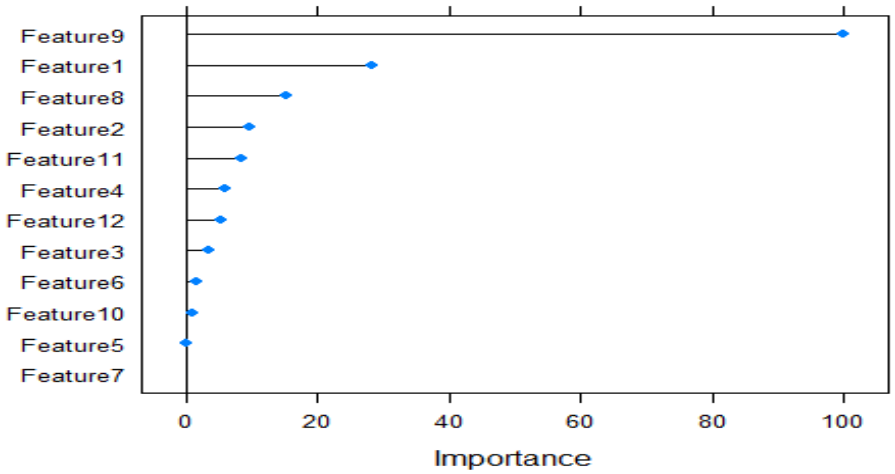
Feature9 100.00000, Feature1 28.38948, Feature8 15.26305,

Feature2 9.72458, Feature11 8.40516, Feature4 5.97589, Feature12 5.28241,

Feature3 3.27030, Feature6 1.63359, Feature10 0.75236, Feature5 0.06303,

Feature7 0.00000

Features 9 and 1 are important in this case and hence varying them would change the results by a large amount.



Logistic Regression:

Coefficients:

(Intercept)	Feature1	Feature2	Feature3	Feature4	Feature5	Feature6
-492.1707	5.8105	-41.2337	-7.3370	-107.2182	7.0811	-3.3713

Feature7	Feature8	Feature9	Feature10	Feature11	Feature12
-15.1498	218.2834	1148.1043	-0.2631	-2.1616	-1209.3202

The features 9 , 12 and 1 have very high coefficients which majorly influence the result of the prediction.

The Model which is recommended for the predictions of Property1 would be Boosting as it has higher training and test accuracy as compared to other models and also boosting is a robust scheme with minimized errors.

The predicted values for the first 10 rows as per boosting algorithm are:

No No No Yes Yes No No No No Yes

Seeing the accuracies of the trained models, it can be said that the predictions of the unseen 10 rows can be predicted with good level of confidence.

For Property2 using Classification Algorithms (Bagging, Boosting, Random Forest and Logistic Regression)

MODEL	TRAIN ACCURACY	TEST ACCURACY
BAGGING	0.9504274	0.9848485
RANDOM FOREST	0.9470085	1
BOOSTING	0.9547009	1
LOGISTIC REGRESSION	0.8977273	0.6666667

Accuracy:

All the models can be used for predicting the unseen values as each model has high training and test accuracy. Since the logistic regression computes the probabilities, there is less accuracy in case of this. Random Forest and Boosting can be used for predicting the models effectively.

Specificity:

As the models have high test and training accuracies, the models are robust and are having similar accuracies with different data proportions and cross validations, hence the models are generalized to work for very different unseen datas and are suitable for unseen data predictions.

Sensitivity:

Different algorithms have different importance of features and hence depends on the algorithm used. Here is the Sensitivity analysis for each model:

Bagging:

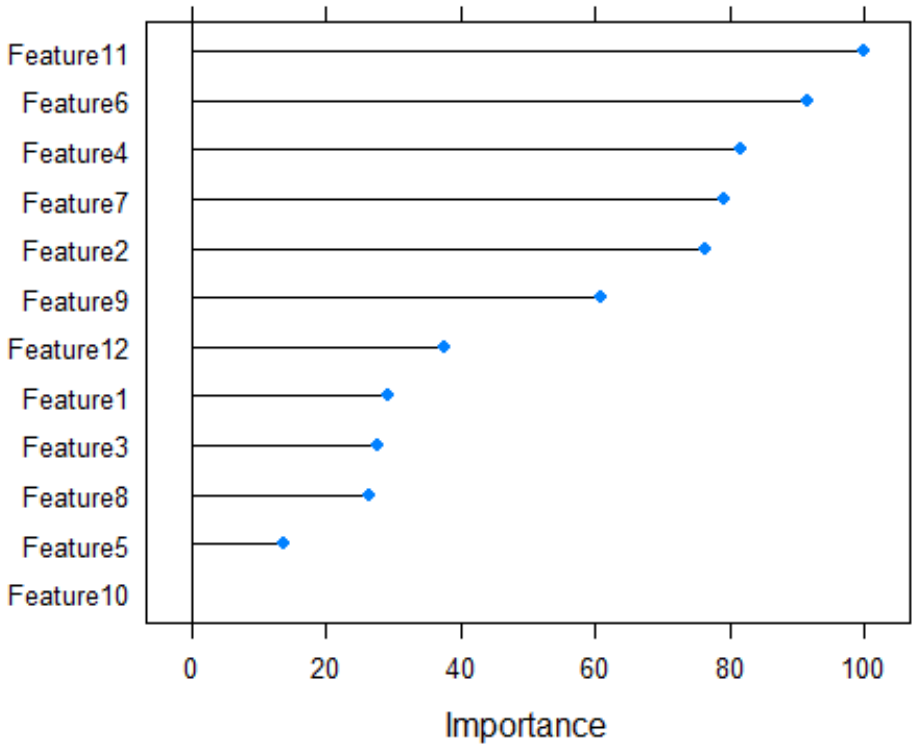
Bagging variable importance

Feature11 100.00, Feature6 91.72, Feature4 81.81, Feature7 79.21,

Feature2 76.39, Feature9 60.93, Feature12 37.66, Feature1 29.28,

Feature3 27.59, Feature8 26.42, Feature5 13.70, Feature10 0.00

The features 11, 6 and 4 influence the property2 classification result.



Random Forest:

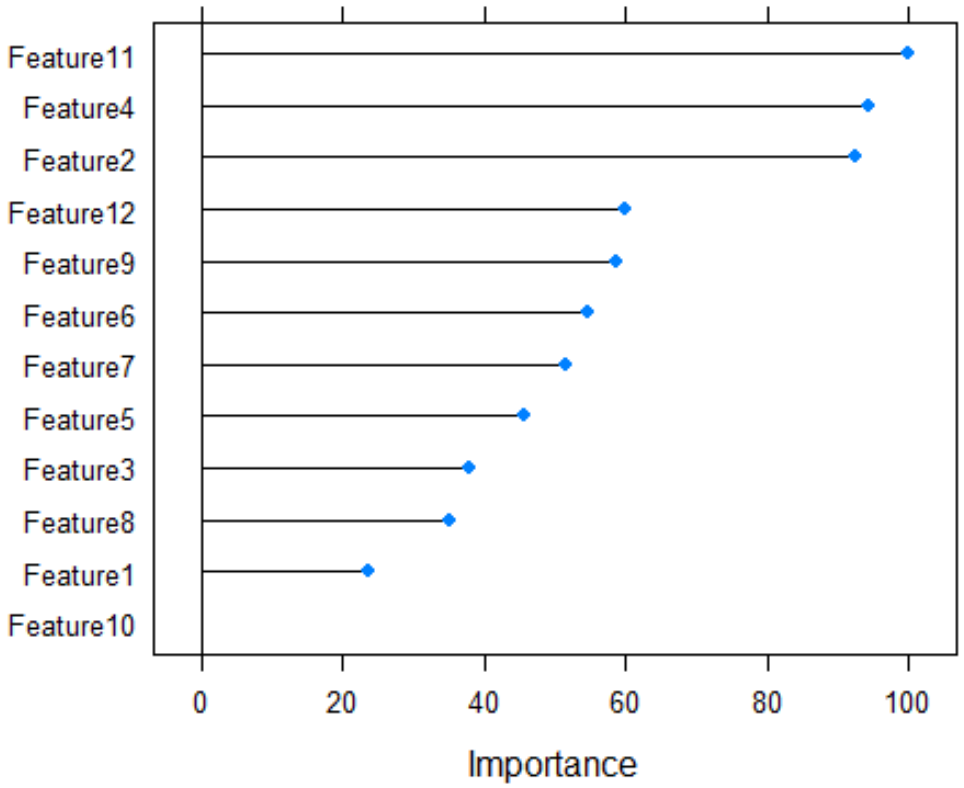
rf variable importance

Feature11 100.00, Feature4 94.35, Feature2 92.56, Feature12 59.96,

Feature9 58.78, Feature6 54.72, Feature7 51.54, Feature5 45.62,

Feature3 38.00, Feature8 35.13, Feature1 23.43, Feature10 0.00

Features 11, 4 and 2 influence the model for Property2



Boosting:

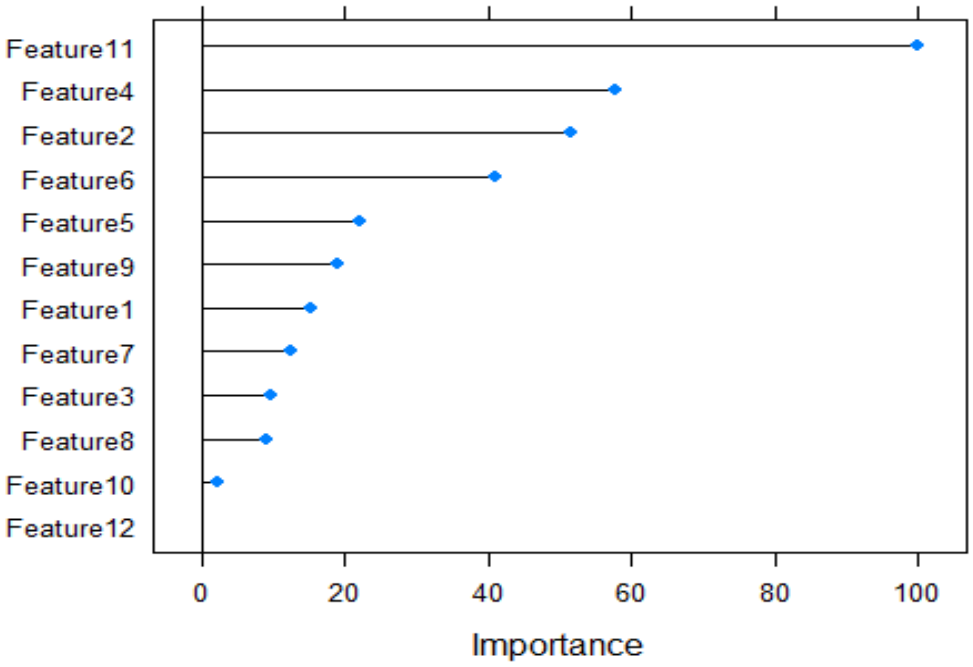
Boosting variable importance

Feature11 100.000, Feature4 57.778, Feature2 51.546, Feature6 40.992,

Feature5 22.068, Feature9 18.973, Feature1 15.305, Feature7 12.458,

Feature3 9.743, Feature8 9.017, Feature10 2.098, Feature12 0.000

Features 11, 4 and 2 influence the model for Property2 predictions



Logistic Regression:

Coefficients:

(Intercept)	Feature1	Feature2	Feature3	Feature4
6.836e+01	1.954e+01	4.291e+01	-1.803e+01	6.186e+01
Feature5	Feature6	Feature7	Feature8	Feature9
8.717e+00	-9.464e+00	-1.274e+01	-5.935e+01	7.536e+02
Feature10	Feature11	Feature12		
8.742e-02	-4.025e+00	-1.225e+03		

The features 12, 10 and 9 are important according to the Logistic Regression model for predicting the Property2.

The Boosting model can be used for predicting the property2 with better accuracies as compared to the other models, as it provided 100 percent accuracy for a large test data set.

Prediction Values for Property2 using the Boosting model 10 rows:

no no yes no no yes yes yes no no

The results are predicted with high confidence as the models are robust and worked well in finding the test dataset predictions.

The Property1 and Property2 seem to be more in the negative region with both negative more than both positive cases, this can be changed if more of the dominant features are increased further so as to bring a positive result in the output.

PART C:

The paper discusses the most critical questions that need to be answered in the field of Big Data where the questions include the steps from data collection to that of the final part of data inference to produce insights. These questions include the motive for collection of data, how the access for data collection is given, how biased is the analysis, how logical is the inference etc. These questions not only bring the notion of how things are working in the current society but also bring about the hierarchy and the permissions given to a certain section of the society to effectively collect data than some other groups which have limited access to public data.

Big Data reframes fundamental concerns regarding the nature of knowledge, research methods, how we should interact with information, and the structure and classification of reality. It also questions whether data can provide an 'objective truth,' or is any interpretation unavoidably influenced by some subjective filter or the manner in which the data is cleaned?

Main conceptual ideas in the paper include that the importance or the use of large amounts of data or whether small samples of data are good enough for the research to be done. This depends on the scale of the questions and the answers needed from the research, many a times, even the smallest of data can provide huge insights as compared to handling large volumes of data, which in turn create more problems than solutions. It also brings light to the situation where collecting data from an api is good enough to span over the entire sample space or is it a selected section of people whose ideas are being analysed. It also questions the ethics and rights of the people when their data is being accessed without their concern by third parties due to which there is large scale privacy and security issues being raised.

There are situations where some unwanted or illogical insights are found which are not possible in the real life, which means there’s a need to understand the data collected and the relation between the insights drawn, which in most cases is not done by the researchers due to their ignorance or reluctance to admit their mistakes.

The follow up questions that need to be pursued in such problems are:

How would these big data tools help in defining the world for the future generations? What are the relations we find helping us to do for the benefit of the society? And how can we ascertain that the research is unbiased without any social implication when brought out for public viewing.

When these questions are answered, it will help in creating a better analysis, and also in providing a better and a more workable solution and a place to live in.