Comorbidity of diabetics with helminth infection

Naga venkata sai kumar(CS18S003), Rahul Biswas(CS18S008)

Abstract

Helminths are multicellular organisms that develops a wide range of strategies to manipulate the host immune system. Immunity to helminths involves profound changes in both the innate and adaptive immune compartments,which can have a protective effect in inflammation and autoimmunity. Finding the features which are important in predicting the double disease cases exclusively can be used for manufacturing drug for union of the two diseases.

Index Terms

Comorbidity, Decision tree, RandomForest, Bagging

I. Introduction

Helminths are multicellular organisms that have developed a wide range of strategies to manipulate the host immune system to survive and complete their reproductive cycles successfully.Improved sanitation, together with infection control, has removed immunoregulatory mechanisms on which our immune system may depend. Recently,Helminth-derived antigens and molecules have been tested in vitro and in vivo to explore possible applications in the treatment of inflammatory and autoimmune diseases, including T1D.The project focuses on finding the features which are exclusively important in identifying the double disease cases and building a multi-class classification model using the bagging based classification model random forests.

II. Methods

A. Gini impurity, entropy and information gain

Entropy calculates the amount of randomness in the system:

$$H = -\sum_i p_i \left(\log_2 p_i\right)$$

Gini impurity is similar to entropy but calculating it is computationally efficient compared to entropy hence we use the gini impurity.

$$I_G(p) = 1 - \sum_{i=1}^{N} p_i^2$$

Information gain is the amount of change in impurity when we split a dataset based upon the attribute X

$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$

when we split a dataset T into smaller datasets T1,T2,....Tn we calculate the wieghted entorpies for these datasets which sum up to Entropy(T,X).

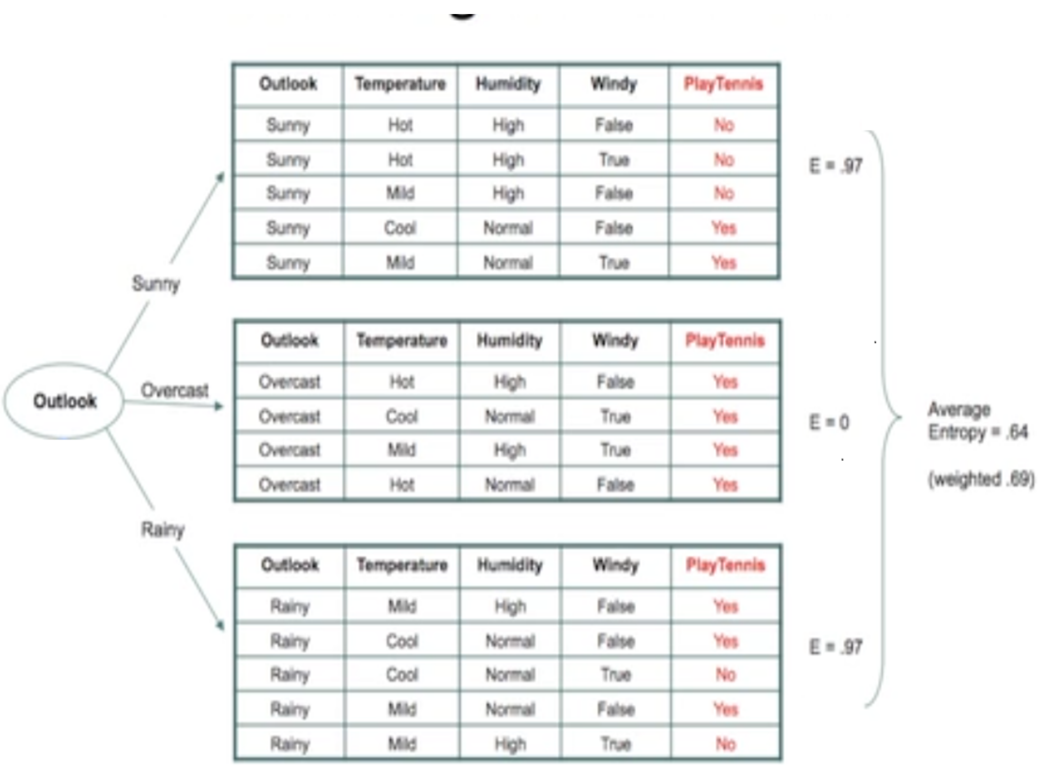
B. Decision trees

Decision trees are classification algorithms which are easily interpretable. decision trees are built by finding the information gain for each feature and then splitting the dataset based upon the feature with maximum information gain and applying this recursively down the decision tree until we reach a specific depth.

M. Shell was with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

J. Doe and J. Doe are with Anonymous University.

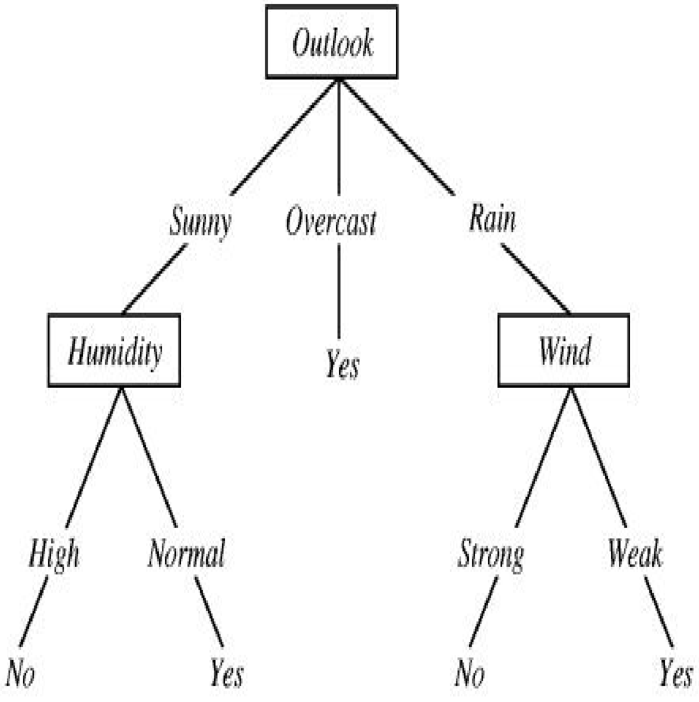
Manuscript received April 19, 2005; revised August 26, 2015.



Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Sunny	Mild	Normal	True	Yes

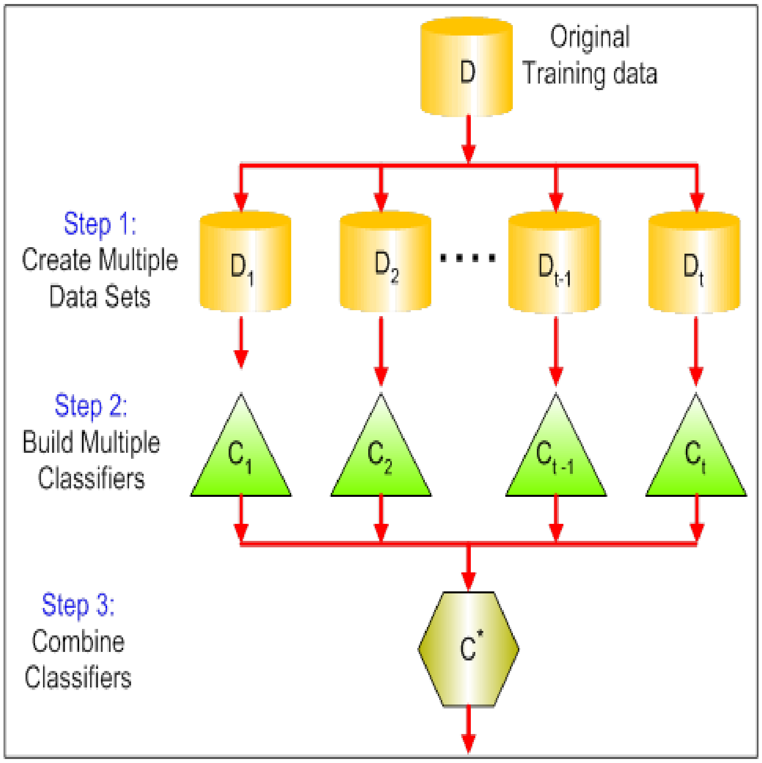
Outlook	Temperature	Humidity	Windy	PlayTennis
Overcast	Hot	High	False	Yes
Overcast	Cool	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes

Outlook	Temperature	Humidity	Windy	PlayTennis
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Rainy	Mild	Normal	False	Yes
Rainy	Mild	High	True	No



C. *Random forests*

Random forests is a bagging technique which contains number of decision trees (base learners) trained by sampling subsets of the training data and result for an input query is obtained by aggregating the results of all these decision trees by using a majority vote.Bagging is a combination of bootstrapping(random sampling)+aggregation.


D
Original Training data
Step 1: Create Multiple Data Sets
D_1
D_2
....
D_{t-1}
D_t
Step 2: Build Multiple Classifiers
C_1
C_2
C_{t-1}
C_t
Step 3: Combine Classifiers
C^*

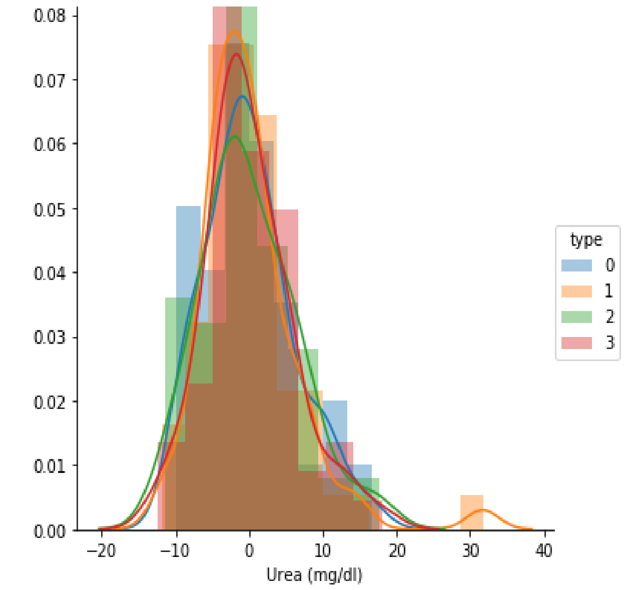
III. Experimental Result

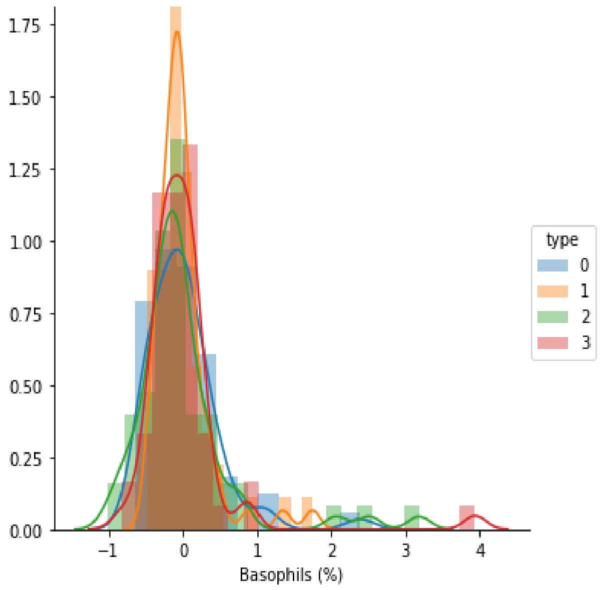
We used four datasets each of 60 datapoints.These are as follows: (1)'Hel+DM+' - subjects with both helminth and diabetes, (2) 'Hel-DM+' - subjects with diabetes only, (3)'Hel+DM-' - subjects with helminth only. (4) 'Hel-DM-' subjects with both diseases free. Each datapoints have 50 features(factors) which are used for clinical test of diabetes and helminths. for example **IL-2**s an interleukin, a type of cytokine signaling molecule in the immune system; **Mean corpuscular volume** MCV blood test measures the average size of your red blood cells; **WBC** white blood cells;**Neutrophils** a type of WBC that protect us from infections etc

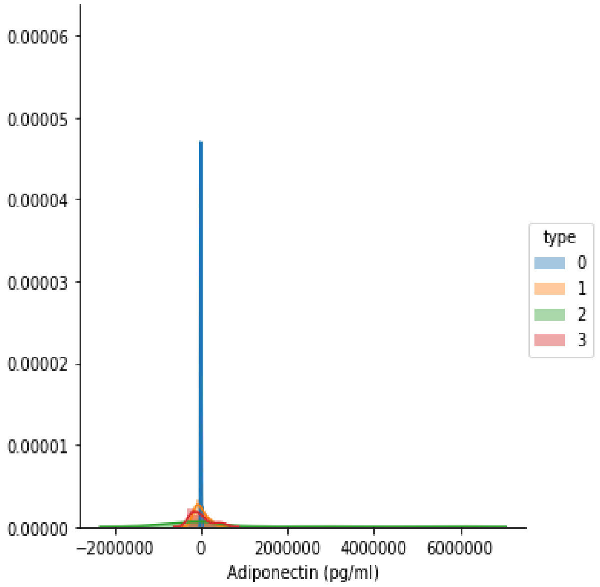
We found that out of 50 features, some of the features scored comparable high score for predicting double disease case than single disease case. The below barchart shows the result. Feature score of **urea, IL-2, MCV** are high in double disease case than single disease cases.

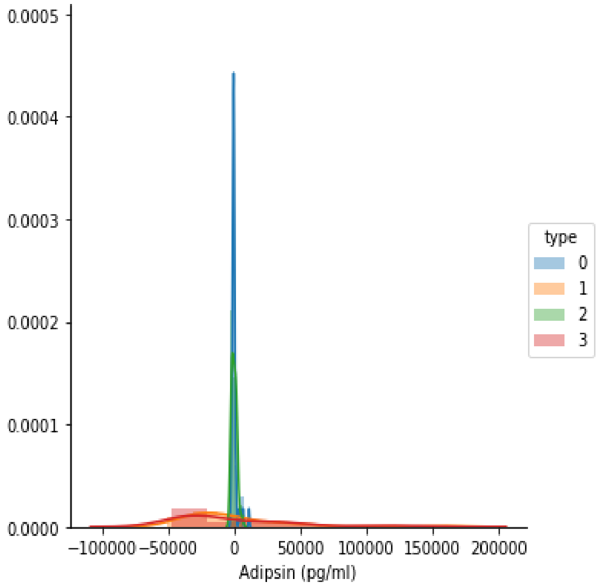
We have performed a univariate analysis on all the features and obtained 50 such plots from which we couldnt infer much about which features can be helpful in succesfully differentiating the double disease classes from the other three classes.The images below are the univariate analysis plots of 4 of such features urea,basophils,adipocetonin,adipsin.

We have made a random forest classifier for the multi class classification and have got a classification accuracy of 95+% and also found the important features which are required for classifying the double disease cases exclusively by using the same model based on the information gain parameter. We found a subset of features important for identifying the double disease cases exclusively and subset of features are **Urea (mg/dl),IL-5 (pg/ml),MCH (pg),MCV (fL),Eosinophils (%)** and we have pruned the 50 feature set to a set of 5 features and found out the classification acuracy of double disease cases vs all the rest of the 3 classes and have obtained an accuracy upto 81.6%. We have obtained different sets of feature importance values each time of running the classifier and we have selected the set of 5 features which appeared commonly among each time we run the classifier .



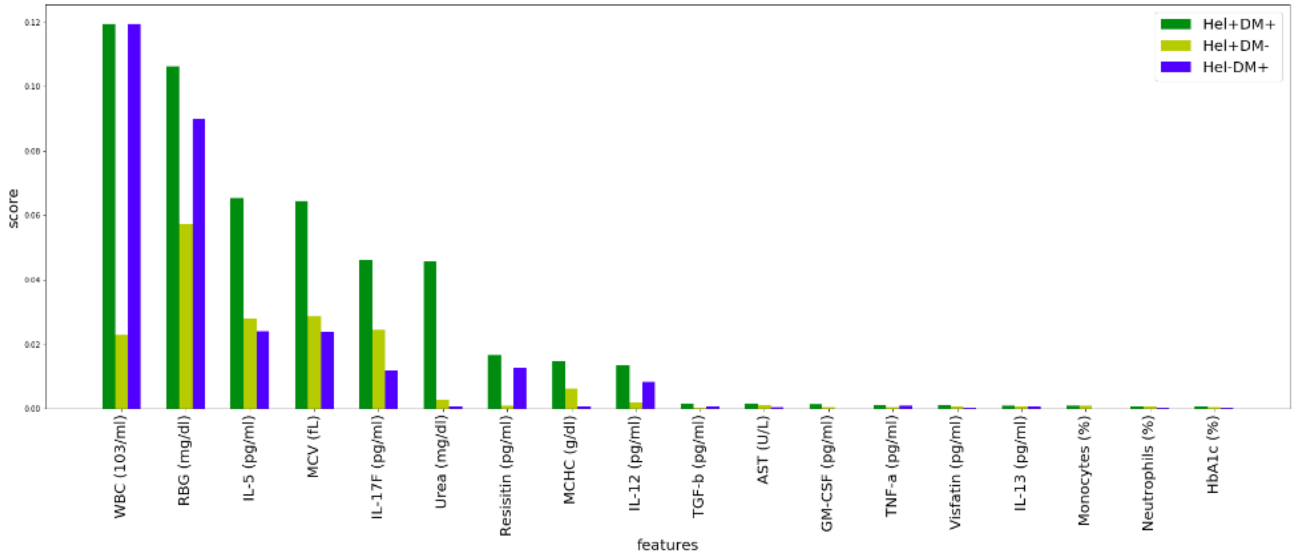


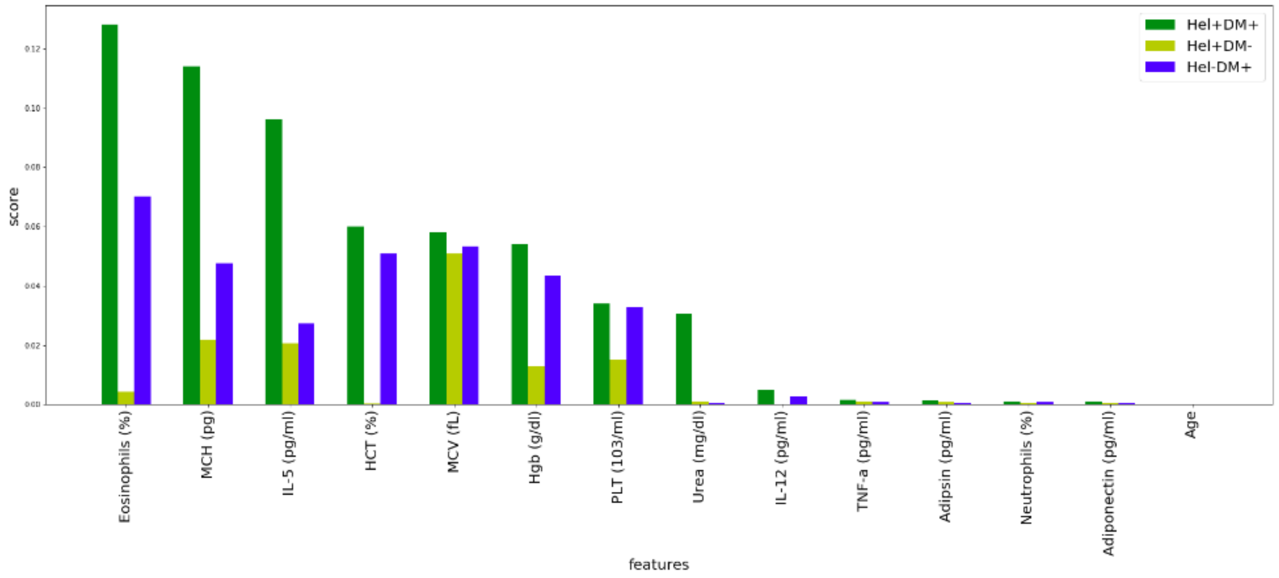




0- Hel+DM-
1- Hel+DM+
2- Hel-DM-
3- Hel-DM+

.





IV. Conclusion

We found a subset of features important for identifying the double disease cases exclusively.The subset of features are **Urea (mg/dl),IL-5 (pg/ml),MCH (pg),MCV (fL),Eosinophils (%)** and we have pruned the 50 feature set to a set of 5 features and found out the classification acuracy of double disease cases vs all the rest of the 3 classes and have obtained an accuracy upto 81.6%.