Personalized Medicine

Nakul Sai.A

As we need to classify the data points into cancer type files . This is classification problem and we have xgboost ,neural networks , Random forests etc for classification

Key columns

Variation , Text, Gene these are the columns we have. Based up on the data We understood that when an variation occurs to particular Gene in our body .. Body sends some cells to repair that gene but when it fails it leads to cancer… Type of cancer depends up on the type of variation in gene ..

We will get variation and gene from the columns Variation and gene . but the reason for the failure is described in text.

Feature Engineering:

Assumption1 : There is a pattern in Gene and variation ie. The first letter and last letter defines the class of the gene and variation

Adv1: 3321 train genes and variations reduced to around 50 combinations

So now we can map between gene class , variant class to cancer type by getting first letter and last letter from the words

Now we come to Text… As it contains huge paragraph describing the cause of cancer. We can also find there is almost one to one relation between type of cancer and text…

We cant load entire Paragraph to model .. so we try find the key words in the paragraph .. This process includes 1.remove stop words, lemmatization then count vectorizer. We use Tfidf vectorizer for the above

Now we got the key words for each row in text….These key words will be our features

So obtained features are keywords, Gene first letter , Gene last letter , Variant first letter , variant last letter.

Target is class

So now train the model either using xgboost or neural networks.

I feel this is not tree based algorithm or not top down approach .. so I choose neural networks implemented using keras.

Now test the model and rearrange the number of hidden layers to get best accuracy.

Make sure your model is not overfitted