LaTeX Template for SBE304 Project Proposal

Mohammed El-Sayed
mohammed.elraoof98@eng-st.cu.edu.eg

Esraa Sayed
esraa.sayed98@eng-st.cu.edu.eg

Galal Hossam
galalhossam555@gmail.com

Zeinab Walid
zeinab.anwer97@eng-st.cu.edu.eg

1. Motivation

We are eager to enter the field of Bioinformatics as it is the only connection between Biomedical Engineering and the future of Machine Learning, so we searched for a topic that relates the Cancer disease with the gene expression. as this topic is very important for patients suffering from this dangerous disease. We are looking forward to detecting the type of cancer disease depending on the spectrum of genomic alterations that promote oncogenesis, origin of cancer cell and location.

2. Project objectives

Our main objective in this project is to reach to the best performance out of this dataset as we face a great challenge in preprocessing stage as we have too much data and we tried to select the best of it to reach to the best output in classifying different types of cancer disease. and we have to choose the best model that visualize this data , so we will use different models and select the model that obtain the best performance.

3. Data

We use dataset for our project depending on this paper:The Cancer Genome Atlas Pan-Cancer analysis project [1]

4. preprocessing

1- Feature Selection : we will use a method like T testing to select the best features with the highest significant level $p_value <= 0.05$, as T testing works on two classes only , so we search for another method that fit with 5 classes like Anova filter based approach that select the features that give the best performance. Our features can not be selected simply without a model dependeing on a method like p-value.
2- Feature normalization : the values of our dataset are near from each other so we don't need to make feature scaling, as feature normalization is important to make data near to each other to get the best performance.
3- Data imputation: we have no missing data.

5. Exploratory data analysis (EDA)

we work on large data with more than 20,000 which is a challenge for us now to visualize these massive data , so we trying to find the best way to select the best features ,then use the selected features to visualize our data. Each model has its own way of visualization ,so we will visualize upon the models we chose. There are different methods to visualize data depending on different methods.

6. Methodology

we will try to use these 4 methods and make a comparison between them to select the best compatible one to our data set:

- Decision Trees: it is easy to interpret and explain which is suitable for our dataset.
- Logistic regression: it is a useful method as you don't have to worry as much about features being correlated.
- K-nearest neighbors (KNN) model:it is very simple in implementation , and classes don't have to be linearly separable.
- Naive Bayes (NB) Classier (or Gaussian NB Classifier): it is a simple model that perform very well.

7. Timetable

We want that all of us study all the models to get more familiar with these models and to get the most benefit from the project.

- Week 1: Preprocessing and split data into training set and testing set.
- Week 2 : Decision tree and Logistic regression models - Updating our websites.

- Week 3 : K-nearest neighbors (KNN) model and Naive Bayes (NB) Classier (or Gaussian NB Classifier)- Updating our websites
- Week 4 :Select the best model.

8. Personal Websites

- Mohammed El-Sayed
- Esraa Sayed
- Galal Hossam
- Zeinab Walid

References

[1] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, p. 1113, 2013.