Case Studies – Fortnightly Task 1

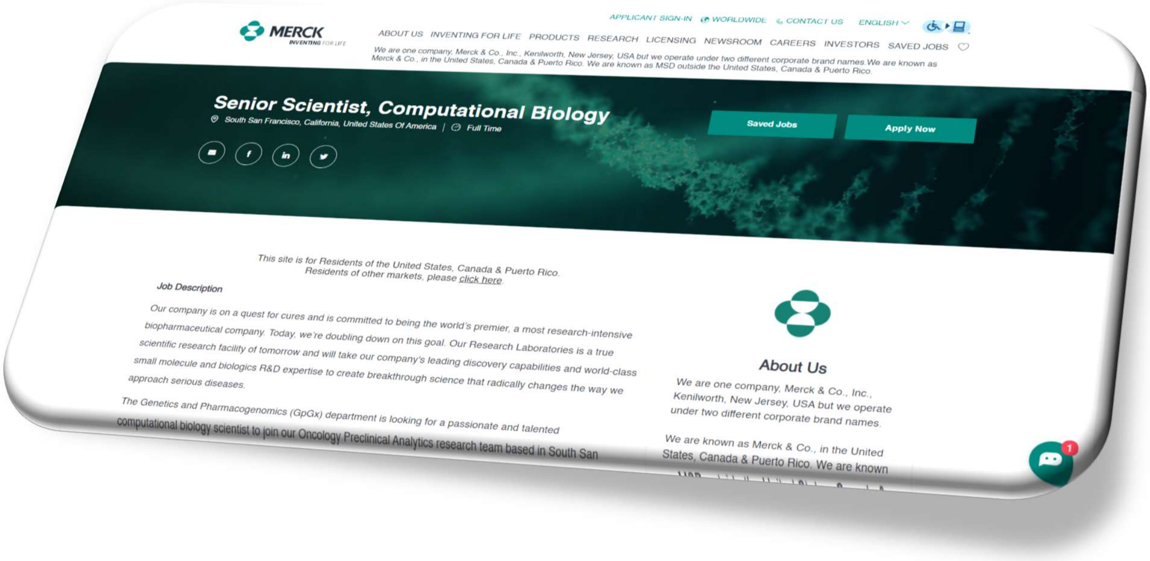
Vishal Patnaik Damodarapatruni - s3811521

03/07/2020

#**#

1. Part 1: Job Role

a. Senior Scientist, Computational Biology. [1]



We are one company, Merck & Co., Inc., Kenilworth, New Jersey, USA but we operate under two different corporate brand names.We are known as Merck & Co., in the United States, Canada & Puerto Rico. We are known as MSD outside the United States, Canada & Puerto Rico.

Senior Scientist, Computational Biology

South San Francisco, California, United States Of America | Full Time

This site is for Residents of the United States, Canada & Puerto Rico.
Residents of other markets, please click here.

Job Description

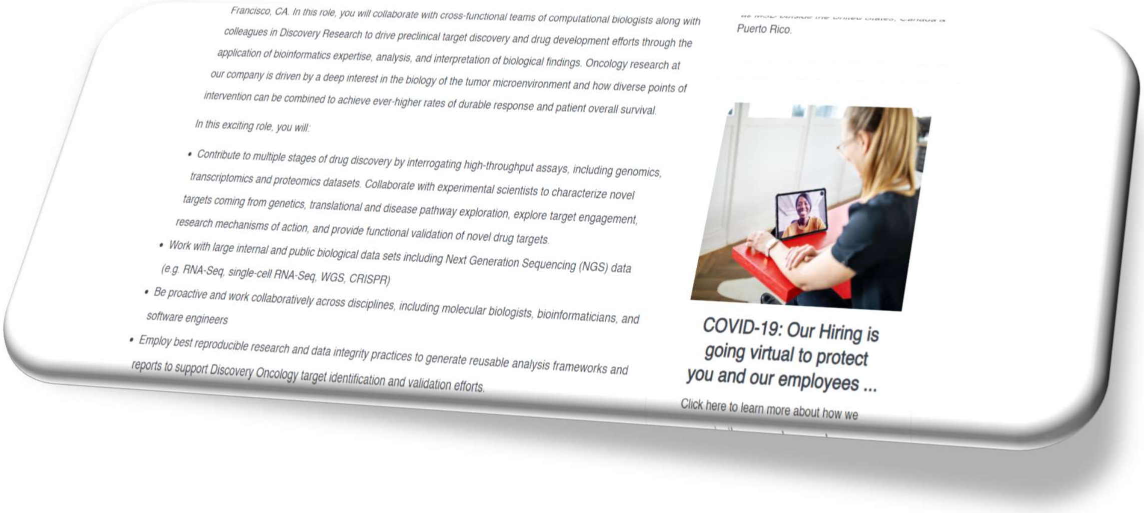
Our company is on a quest for cures and is committed to being the world's premier, a most research-intensive biopharmaceutical company. Today, we're doubling down on this goal. Our Research Laboratories is a true scientific research facility of tomorrow and will take our company's leading discovery capabilities and world-class small molecule and biologics R&D expertise to create breakthrough science that radically changes the way we approach serious diseases.

The Genetics and Pharmacogenomics (GpGx) department is looking for a passionate and talented computational biology scientist to join our Oncology Preclinical Analytics research team based in South San

About Us

We are one company, Merck & Co., Inc., Kenilworth, New Jersey, USA but we operate under two different corporate brand names.

We are known as Merck & Co., in the United States, Canada & Puerto Rico. We are known



Francisco, CA. In this role, you will collaborate with cross-functional teams of computational biologists along with colleagues in Discovery Research to drive preclinical target discovery and drug development efforts through the application of bioinformatics expertise, analysis, and interpretation of biological findings. Oncology research at our company is driven by a deep interest in the biology of the tumor microenvironment and how diverse points of intervention can be combined to achieve ever-higher rates of durable response and patient overall survival.

In this exciting role, you will:

- Contribute to multiple stages of drug discovery by interrogating high-throughput assays, including genomics, transcriptomics and proteomics datasets. Collaborate with experimental scientists to characterize novel targets coming from genetics, translational and disease pathway exploration, explore target engagement, research mechanisms of action, and provide functional validation of novel drug targets.
- Work with large internal and public biological data sets including Next Generation Sequencing (NGS) data (e.g. RNA-Seq, single-cell RNA-Seq, WGS, CRISPR)
- Be proactive and work collaboratively across disciplines, including molecular biologists, bioinformaticians, and software engineers
- Employ best reproducible research and data integrity practices to generate reusable analysis frameworks and reports to support Discovery Oncology target identification and validation efforts.

Puerto Rico.

COVID-19: Our Hiring is going virtual to protect you and our employees ...

Click here to learn more about how we

b. Field.

Computational Biology

c. Job role description.

- The Job role looks for the one who can look insights into the data and make predictions through the data as well.
- As the role itself states that we should understand and analyze the genomics, transcriptomics, and proteomics datasets by collaborating with the experimental scientists (Looking for data insights).
- It demands machine learning algorithms to predict and solve biological problems (Predictions).

*#***#*

2. Part 2: Data Set

a) Data [2]

Source : https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression

The data set taken is,

- ***Data_Cortex_Nuclear.xls***

  This data collection comprises of the expression levels of 77 proteins/protein changes that delivered perceptible signs in the atomic division of cortex. There are 38 control mice and 34 trisomic mice (Down disorder), for a sum of 72 mice. In the analyses, 15 estimations were enrolled of every protein per test/mouse. Along these lines, for control mice, there are 38x15, or 570 estimations, and for trisomic mice, there are 34x15, or 510 estimations. The dataset contains a sum of 1080 estimations for each protein. Every estimation can be considered as a free example/mouse.

  The eight classes of mice are portrayed dependent on highlights, for example, genotype, conduct and treatment. As indicated by genotype, mice can be control or trisomic. As per conduct, a few mice have been invigorated to learn (setting stun) and others have not (stun setting) and so as to survey the impact of the medication memantine in recuperating *the capacity to learn in trisomic mice, a few mice have been infused with the medication and others have not.*

- ***Data Set Characteristics : Multivariate***
- ***Attribute Characteristics : Real***
- ***Associated Tasks : Classification, Clustering***
- ***Field / Domain : Life***
- ***Instances / Observations / Rows : 1080***
- ***Attributes / Columns : 82***
    - ***1 : Mouse ID***
      Type -> Object

    - ***2 – 78 : Values of expression levels of 77 proteins.***
      Type -> Double

    - ***79 : Genotype***
      Type -> Factor
      They are 2.
      [control (c) or trisomy (t)]

    - ***80 : Treatment***
      Type -> Factor
      They are 2.
      [memantine (m) or saline (s)]

    - ***81 : Behavior***
      Type -> Factor
      They are 2.
      [context-shock (CS) or shock-context (SC)]

    - ***82 : Class***
      c-CS-s: control mice, stimulated to learn, injected with saline (9 mice)
      c-CS-m: control mice, stimulated to learn, injected with memantine (10 mice)
      c-SC-s: control mice, not stimulated to learn, injected with saline (9 mice)
      c-SC-m: control mice, not stimulated to learn, injected with memantine (10 mice)

t-CS-s: trisomy mice, stimulated to learn, injected with saline (7 mice)
t-CS-m: trisomy mice, stimulated to learn, injected with memantine (9 mice)
t-SC-s: trisomy mice, not stimulated to learn, injected with saline (9 mice)
t-SC-m: trisomy mice, not stimulated to learn, injected with memantine (9 mice)

b) Data Set and Job relativity. [3]

The Field Computational science includes the turn of events and utilization of information diagnostic and hypothetical strategies, mathematical modeling, and computational simulation techniques to the study of biological, ecological, behavioral, and social systems. The field is extensively characterized and remembers establishments for biology, applied mathematics, statistics, biochemistry, chemistry, biophysics, molecular biology, genetics, genomics, computer science, and evolution.

This has the following subfields –

- ***Computational anatomy***
- ***Computational biomodelling***
- ***Computational genomics***
- ***Computational neuroscience***
- ***Computational pharmacology***
- ***Computational evolutionary biology***
- ***Cancer computational biology***
- ***Computational neuropsychiatry***

Clearly our job falls under the subfield ***Computational genomics****. Coming to the data set, it gives the information about the mice proteins based on chromosomes which is related to molecular biology involving the studies based on DNAs and RNAs. This in turn relates to genomes and can be termed as genomic data. So, I chosen this data set to perform modeling and predict based on the analysis made which is obvious to the job role.*

#**#

3. Part 3 : Experiment

Modelling.

Model Background

- *Two Supervised Machine learning algorithms, Random Forest and Support Vector Machines are chosen and build for modelling to foresee the order of mice classes or basic classes itself.*
- *Also, to anticipate the significant or crucial proteins of each class.*
- *Based on input factors the model ought to anticipate the estimations of target variable.*

1. ***Random Forest –*** [4]
   - *Random forest is a supervised learning algorithm which is utilized for both classification as well as regression.*
   - *This algorithm makes decision trees on data and thereby getting the prediction from each tree.*
   - *Lastly it chooses the best arrangement by methods for casting a ballot.*
   - *It is an assembling technique which is better than a single decision tree because it diminishes the over-fitting by averaging the outcome.*
2. ***Support Vector machine (SVM) –*** [5]
   - *Support vector machines (SVMs) are incredible yet adaptable supervised machine learning algorithms which are utilized both for classification and regression. But mostly for Classification Problems.*
   - *SVMs have their remarkable method of execution than the other Machine Learning algorithms.*
   - *They are very popular due to their capacity to deal with various continuous and categorical variables.*

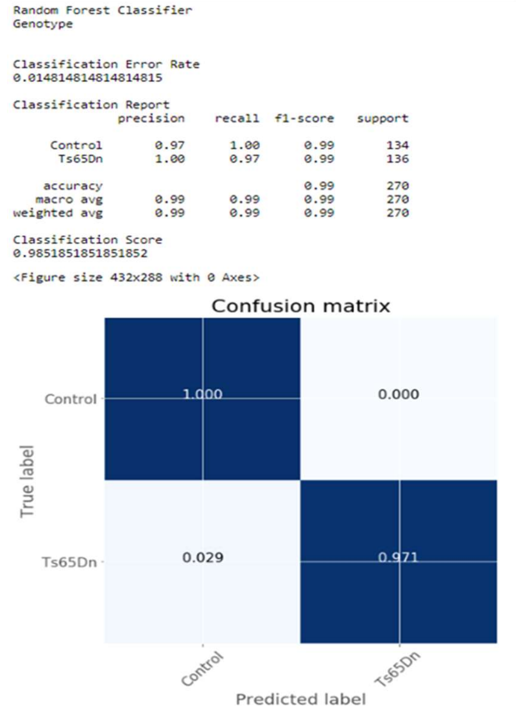
c) Model Results

➢ *Here 2 models are built,*

- *Random Forest Classifier with all factor or binary variables and reported individually.*
- *SVM Classifier using same process as that of Random Forest.*

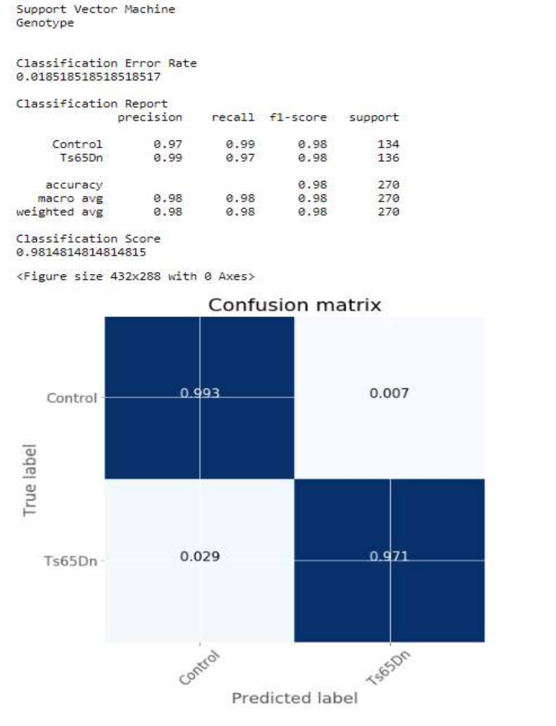
➢ GENOTYPE

RANDOM FOREST



```
Random Forest Classifier
Genotype


Classification Error Rate
0.014814814814814815

Classification Report
              precision    recall  f1-score   support

     Control       0.97      1.00      0.99       134
      Ts65Dn       1.00      0.97      0.99       136

    accuracy                           0.99       270
   macro avg       0.99      0.99      0.99       270
weighted avg       0.99      0.99      0.99       270

Classification Score
0.9851851851851852

<Figure size 432x288 with 0 Axes>
```

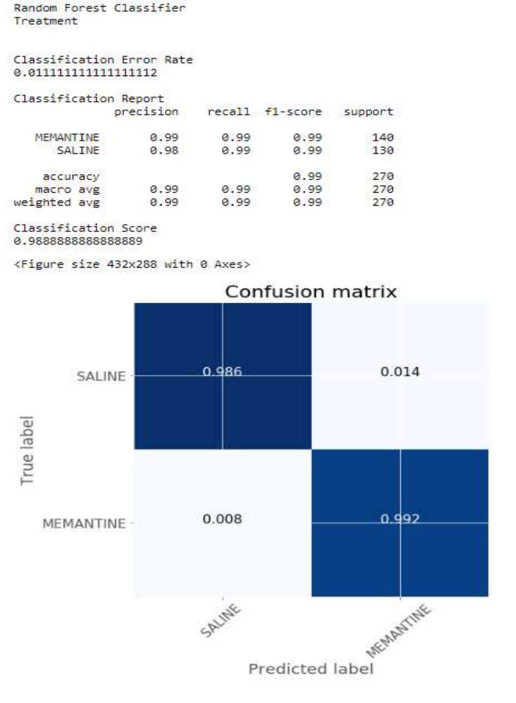
SVM



```
Support Vector Machine
Genotype


Classification Error Rate
0.018518518518518517

Classification Report
              precision    recall  f1-score   support

     Control       0.97      0.99      0.98       134
      Ts65Dn       0.99      0.97      0.98       136

    accuracy                           0.98       270
   macro avg       0.98      0.98      0.98       270
weighted avg       0.98      0.98      0.98       270

Classification Score
0.9814814814814815

<Figure size 432x288 with 0 Axes>
```

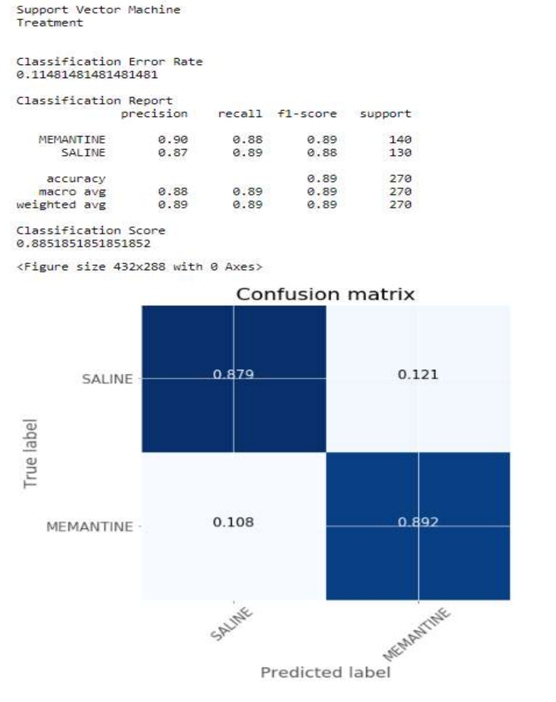
➢ TREATMENT

RANDOM FOREST



```
Random Forest Classifier
Treatment


Classification Error Rate
0.011111111111111112

Classification Report
              precision    recall  f1-score   support

   MEMANTINE       0.99      0.99      0.99       140
      SALINE       0.98      0.99      0.99       130

    accuracy                           0.99       270
   macro avg       0.99      0.99      0.99       270
weighted avg       0.99      0.99      0.99       270

Classification Score
0.9888888888888889

<Figure size 432x288 with 0 Axes>
```

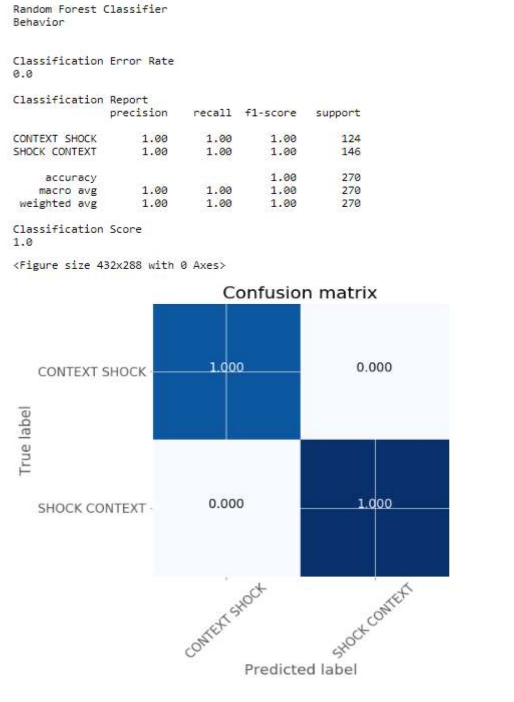
SVM



```
Support Vector Machine
Treatment


Classification Error Rate
0.11481481481481481

Classification Report
              precision    recall  f1-score   support

   MEMANTINE       0.90      0.88      0.89       140
      SALINE       0.87      0.89      0.88       130

    accuracy                           0.89       270
   macro avg       0.88      0.89      0.89       270
weighted avg       0.89      0.89      0.89       270

Classification Score
0.8851851851851852

<Figure size 432x288 with 0 Axes>
```

➢ BEHAVIOR

RANDOM FOREST

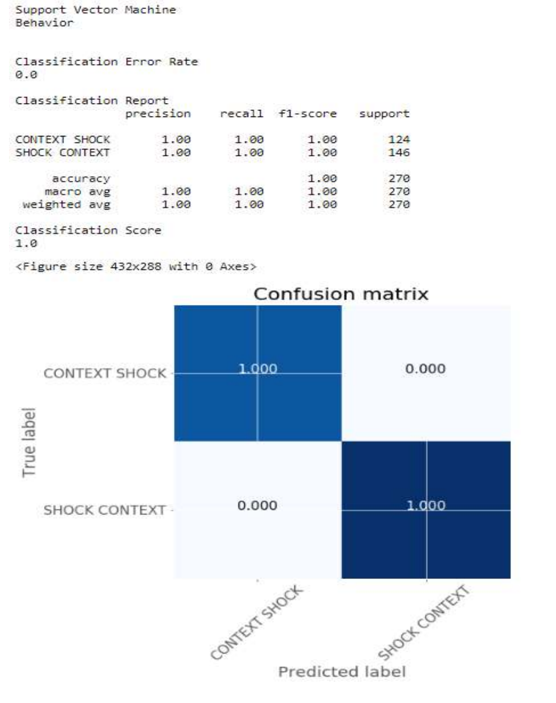


```
Random Forest Classifier
Behavior


Classification Error Rate
0.0

Classification Report
               precision    recall  f1-score   support

CONTEXT SHOCK       1.00      1.00      1.00       124
SHOCK CONTEXT       1.00      1.00      1.00       146

     accuracy                           1.00       270
    macro avg       1.00      1.00      1.00       270
 weighted avg       1.00      1.00      1.00       270

Classification Score
1.0

<Figure size 432x288 with 0 Axes>
```

Confusion matrix

	CONTEXT SHOCK	SHOCK CONTEXT
CONTEXT SHOCK	1.000	0.000
SHOCK CONTEXT	0.000	1.000

SVM



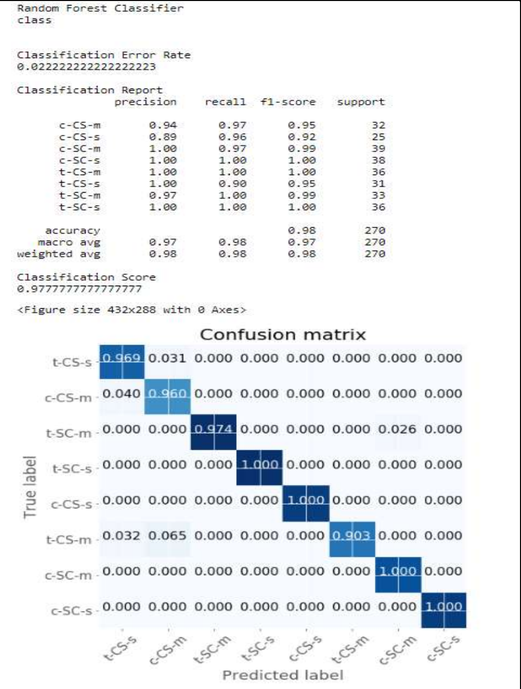
```
Support Vector Machine
Behavior


Classification Error Rate
0.0

Classification Report
               precision    recall  f1-score   support

CONTEXT SHOCK       1.00      1.00      1.00       124
SHOCK CONTEXT       1.00      1.00      1.00       146

     accuracy                           1.00       270
    macro avg       1.00      1.00      1.00       270
 weighted avg       1.00      1.00      1.00       270

Classification Score
1.0

<Figure size 432x288 with 0 Axes>
```

Confusion matrix

	CONTEXT SHOCK	SHOCK CONTEXT
CONTEXT SHOCK	1.000	0.000
SHOCK CONTEXT	0.000	1.000

➢ CLASS

RANDOM FOREST

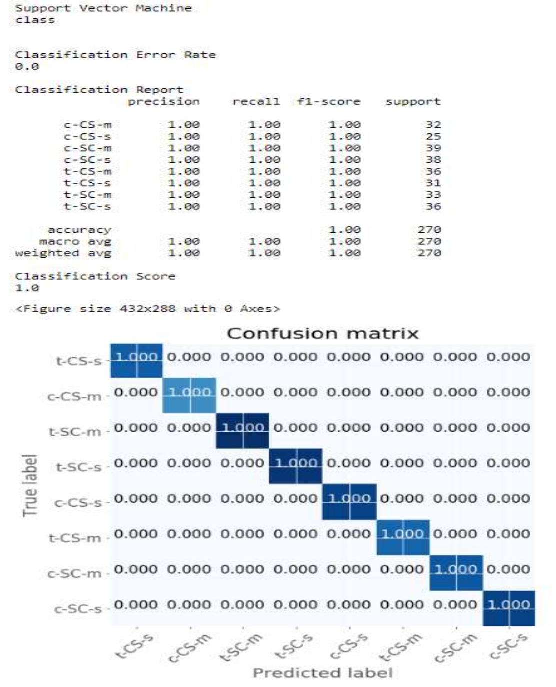


```
Random Forest Classifier
class


Classification Error Rate
0.022222222222222223

Classification Report
              precision    recall  f1-score   support

      c-CS-m       0.94      0.97      0.95        32
      c-CS-s       0.89      0.96      0.92        25
      c-SC-m       1.00      0.97      0.99        39
      c-SC-s       1.00      1.00      1.00        38
      t-CS-m       1.00      1.00      1.00        36
      t-CS-s       1.00      0.90      0.95        31
      t-SC-m       0.97      1.00      0.99        33
      t-SC-s       1.00      1.00      1.00        36

    accuracy                           0.98       270
   macro avg       0.97      0.98      0.97       270
weighted avg       0.98      0.98      0.98       270

Classification Score
0.9777777777777777

<Figure size 432x288 with 0 Axes>
```

Confusion matrix

	t-CS-s	c-CS-m	t-SC-m	t-SC-s	c-CS-s	t-CS-m	c-SC-m	c-SC-s
t-CS-s	0.969	0.031	0.000	0.000	0.000	0.000	0.000	0.000
c-CS-m	0.040	0.960	0.000	0.000	0.000	0.000	0.000	0.000
t-SC-m	0.000	0.000	0.974	0.000	0.000	0.000	0.026	0.000
t-SC-s	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
c-CS-s	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
t-CS-m	0.032	0.065	0.000	0.000	0.000	0.903	0.000	0.000
c-SC-m	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
c-SC-s	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000

SVM



```
Support Vector Machine
class


Classification Error Rate
0.0

Classification Report
              precision    recall  f1-score   support

      c-CS-m       1.00      1.00      1.00        32
      c-CS-s       1.00      1.00      1.00        25
      c-SC-m       1.00      1.00      1.00        39
      c-SC-s       1.00      1.00      1.00        38
      t-CS-m       1.00      1.00      1.00        36
      t-CS-s       1.00      1.00      1.00        31
      t-SC-m       1.00      1.00      1.00        33
      t-SC-s       1.00      1.00      1.00        36

    accuracy                           1.00       270
   macro avg       1.00      1.00      1.00       270
weighted avg       1.00      1.00      1.00       270

Classification Score
1.0

<Figure size 432x288 with 0 Axes>
```

Confusion matrix

	t-CS-s	c-CS-m	t-SC-m	t-SC-s	c-CS-s	t-CS-m	c-SC-m	c-SC-s
t-CS-s	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
c-CS-m	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
t-SC-m	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
t-SC-s	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
c-CS-s	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
t-CS-m	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
c-SC-m	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
c-SC-s	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000

The metrics and methods that help in deciding which classifier is effective on this data are as follows,

- ***Confusion Matrix***
- ***Classification error rate***
- ***Classification Score***
- ***Precision***
- ***Recall***
- ***F1-score***

1. Random Forest –

- *The behavorial classification is more accurate as the value is equal to 1. 1 indicates 100% classification and 0 implies 0% classification.*
- *"Memantine" has 0.992 – 99.2% classification whereas for "Saline" 0.986 which implies 98.6% classification.*
- *Coming to Genotype "Control" got more classification result of nearly 1.000 accounting to 100% accuracy and "Ts65DN" -> 0.971 – 97.1%.*
- *The "Precision" from the "Classification report" for class label "Genotype" – Control 0.97 and Ts65Dn 1.00, for "Treatment" - MEMANTINE 0.99 and SALINE 0.98, for "Behavior" - CONTEXT SHOCK 1.00 and SHOCK CONTEXT 1.00 and for "Class" c-CS-m 0.94, c-CS-s 0.89, c-SC-m 1.00, c-SC-s 1.00, t-CS-m 1.00, t-CS-s 1.00, t-SC-m 0.97, t-SC-s 1.00.*
- *The "Recall" for class label "Genotype" – Control 1.00 and Ts65Dn 0.97, for "Treatment" - MEMANTINE 0.99 and SALINE 0.99, for "Behavior" - CONTEXT SHOCK 1.00 and SHOCK CONTEXT 1.00 and for "Class" is as c-CS-m 0.97, c-CS-s 0.96, c-SC-m 0.97, c-SC-s 1.00, t-CS-m 1.00, t-CS-s 0.90, t-SC-m 1.00, t-SC-s 1.00.*
- *The "F1 - score" for class label "Genotype" – Control 0.99 and Ts65Dn 0.99, for "Treatment" - MEMANTINE 0.99 and SALINE 0.99, for "Behavior" - CONTEXT SHOCK 1.00 and SHOCK CONTEXT 1.00.*
- *While coming to "Class" Random Forest has different F1 – scores for each class "c-CS-m 0.95, c-CS-s 0.92, c-SC-m 0.99, c-SC-s 1.00, t-CS-m 1.00, t-CS-s 0.95, t-SC-m 0.99, t-SC-s 1.00".*

2. Support Vector machine (SVM) –

- *SVM predicted the classification of all 8 classes accurately.*
- *The behavorial classification is more accurate as the value is equal to 1. 1 indicates 100% classification and 0 implies 0% classification.*
- *"Memantine" has 0.892 – 89.2% classification whereas for "Saline" 0.879 which implies 87.9% classification.*
- *Coming to Genotype "Control" got more classification result of nearly 0.993 accounting to 99.3% accuracy and "Ts65DN" -> 0.971 – 97.1%.*
- *The "Precision" from the "Classification report" for class label "Genotype" – Control 0.97 and Ts65Dn 0.99, for "Treatment" - MEMANTINE 0.90 and SALINE 0.87, for "Behavior" - CONTEXT SHOCK 1.00 and SHOCK CONTEXT 1.00 and for "Class" is 1 for all classes.*
- *The "Recall" for class label "Genotype" – Control 0.99 and Ts65Dn 0.97, for "Treatment" - MEMANTINE 0.88 and SALINE 0.89, for "Behavior" - CONTEXT SHOCK 1.00 and SHOCK CONTEXT 1.00 and for "Class" is 1 for all classes.*
- *The "F1 - score" for class label "Genotype" – Control 0.98 and Ts65Dn 0.97, for "Treatment" - MEMANTINE 0.89 and SALINE 0.88, for "Behavior" - CONTEXT SHOCK 1.00 and SHOCK CONTEXT 1.00 and for "Class" is 1 for all classes.*

d)Comparative conclusion

- *From the above results we can say that the 2 classifiers performed well for different class labels.*
- *The "Classification Error Rate" for class labels "Genotype" and "Treatment" is more for SVM compared to Random Forest. Whereas for class label "Behavior" it is 0 for both classifiers. Suggesting their good prediction ability.*
- *Again, SVM has 0 classification error rate for class label "Class", making it more accurate than Random forest in this case.*
- *The "Classification score" for class labels "Genotype" and "Treatment" is more for Random Forest compared to SVM. Whereas for class label "Behavior" it is 1 for both classifiers. Suggesting their good prediction ability.*

- *Again, SVM has classification score of 1 for class label "Class", making it more accurate than Random forest in this case.*
- *The "Precision" is more in Random Forest compared to SVM for both "Genotype" and "Treatment". Whereas SVM tops when classified by "Class".*
- *The "Recall " is more in Random Forest compared to SVM for both "Genotype" and "Treatment". Whereas SVM tops when classified by "Class".*
- *The "F1 – score" from the "Classification report" is more in Random Forest compared to SVM for both "Genotype" and "Treatment". Whereas SVM tops when classified by "Class".*
- *All the metrics for "Behavior" class label are 1.00 suggesting the models 100% classification.*
- *The "Trace" of the confusion matrix of random Forest is greater than that of SVM for both "Genotype" and "Treatment". Whereas for "Class" SVM has the greater trace.*
- *I got confusion matrix values in the scope of 0 to 1 since I set the scale to 0 to 1 utilizing StandardScaler – Normalizer – MinMaxScaler. For better understanding of data values.*
- *As we know that "if the trace of the matrix is very high compared to the sum of all matrix entries, then the predictions for majority of the parts are correct". Here this is seen in Random Forest than in SVM in most cases.*
- *From these insights we can clearly say that Random Forest is more accurate in predicting results and classifying the proteins than SVM when classified by "Genotype" and "Treatment".*
- *When coming to "Class" it is SVM which classified accurately.*
- *Both Classifiers performed well when classified based on Behavior.*

References

[1] Merck Sharp & Dohme Corp., "Merck INVENTING FOR LIFE," Merck Sharp & Dohme Corp., 2019. [Online]. Available: https://jobs.merck.com/us/en/job/MERCUSR64869ENUS/Senior-Scientist-Computational-Biology?utm_source=indeed&utm_medium=phenom-feeds&utm_source=Indeed&utm_medium=organic&utm_campaign=Indeed. [Accessed 2 August 2020].

[2] G. K. C. K. Higuera C, "UCI Machine Learning Repository," 2015. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression. [Accessed 1 August 2020].

[3] Wikipedia , "Wikipedia," 22 June 2020. [Online]. Available: https://en.wikipedia.org/wiki/Computational_biology. [Accessed 2 August 2020].

[4] Tutorials Point Simply Learning, "Tutorials Point," 2020. [Online]. Available: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm. [Accessed 2 August 2020].

[5] Tutorials Point Simply Learning, "Tutorials Point," 2020. [Online]. Available: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_support_vector_machine.htm. [Accessed 2 August 2020].