Personality prediction using machine learning classifiers

Xin Yee Chin
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
cindychin0724@gmail.com

Han Yang Lau
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
hans-lau2@hotmail.com

Zhi Xin Chong
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
zhixin20000803@gmail.com

Man Pan Chow
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
chrischow1995@gmail.com

Zailan Arabee Abdul Salam
School of Computing
Asia Pacific University of Technology and Innovation (APU)
Kuala Lumpur, Malaysia
zailan@apu.edu.my

Abstract— **Personality is a fundamental basis of human behaviour. At most basic, personality including patterns of thought, feeling, behaviours that make an individual unique. Personality will directly or indirectly influence the interaction or preferences of a person. This research using different learning algorithms and concepts of data mining to mine on the data features and learn from the pattern. The aim of this experiment is to explore different options of the algorithm on modifying the personality prediction source code by using logistic regression algorithm, and to find whether the accuracy of the classification can be improved. There are five characteristics of different people that are known as the Big Five characteristic, which is openness, neuroticism, conscientiousness, agreeableness and extraversion that have been stored in the dataset used for training. Then, an overview and comparison will be provided on the different measures taken to reduce the issues faced by researchers in this field. Classification methods implemented are Support Vector Machine, Ridge Algorithm, Naive Bayes, Logistic Regression and Voting Classifier. Testing results showed that the Logistic Regression still outperformed the other methods.**

Keywords—*machine learning; personality prediction; Big Five Personality; regression*

I.   Introduction

Personality is all about the different characteristic of an individual's pattern of feeling, behaving and thinking. Personality embraces the mood, opinion and attitude of someone, it is the best expression way to express clearly and in an understandable form when interacting with someone. Personality provides the ability to distinguish one person from another that can be observed in the workplace environment and so on. Although there are many more ways to explain what personality exactly is, from the psychological point of perspective, there are two main explanations. First pertains to the consistency of differences between humans.

In this way, the study of personality can focus on classifying and identifying human's psychological patterns. Second belongs to the emphasis of quality which is mostly likely to make people alike and that will help to distinguish psychological man from the other species. Personality theorists are then directed to research about those regulations among people that can usefully define the nature of man and other factors that influence the course of live. Understanding personality is important and useful. Personality provides people the idea of how leading, influencing communication can take place in certain conditions. For example, personality traits such as agreeableness and extraversion are mostly going to improve the chance of communication. Whereas personality traits such as high self-esteem are most likely going to remain silent at the workplace.

Therefore, personality shows that it can be useful in many situations. In this experiment, machine learning is being apply to judge and classify personality. Based on the previous source code, logistic regression was used to classify the big five personalities. Big five personality traits include openness, neuroticism, conscientiousness, agreeableness and extraversion. Classifying these personality traits is useful in many ways, one of the reasons to classify personality is to check the suitability of an employee. Employee's personality is often tested in real time to determine which position of the job he or she is particularly fitting in well.

In this research, different algorithms are added to further explore the dataset to test if higher accuracy can be found and created. Classification methods will be added to the original code is Support Vector Machine, Ridge Algorithm, Naïve Bayes, Logistic Regression and Voting Classifier. Logistic Regression is being the default algorithms to the source code.

Critical analysis was performed on similar projects/papers that used different methods. [2] used Multiclass Support Vector Machine (SVM) to perform personality classification based on handwriting. The personalities are Optimistic, Extrovert, Introvert, Sloppy, Energetic. Multiclass classification. Histogram of oriented gradient performs feature extraction on handwriting data, and noise removal performed using adaptive thresholding followed by resizing to reorient image, then multiclass SVM classification was applied using polynomial kernel to map the feature space to a higher dimension, then a hyper-plane will be created which classifies the handwriting features to different classes. Even

with a limited dataset and feature extraction, it achieves 80% accuracy. [2] study used the position of the user's iris based on Eye Accessing Cues in Neural Language Processing to predict their personality. The iris position is an indicator of the mind's internal representational system of what brain sections' currently active. Support Vector Machine is used to take a rectangle crop of the eye with 9000 pixels as an input. Visual, auditory and kinaesthetic (VAK) learning style is used because it best conveys the personality of the person. 215 images of eyes (features) are pre-processed with an eye detection procedure called Cascade Object detector and resized into smaller size and classified with SVM. The results show Radial Basis Function kernel (Standard Gaussian Kernel) has best accuracy at 84.9% followed by linear kernel at 83.7%, and a train-test split of 75:25 gives the best accuracy at 84.9%, followed by 70:30 at 82.8%.

Allouch, Azaria and Azoulay's [4] used a voting classifier algorithm to assist children with special needs with communication in social encounters by recognising possibly insulting or harmful sentences. Dataset consists of interviews with parents of ASD children, categorised to five categories. Audio is translated to text for text classification by the voting classifier (ensemble method), it performs a voting protocol and chooses the result that the majority of algorithms suggest. Algorithms used in voting classifier included random forest, SVM, ridge classifier, extra trees, Bayesian inference method, MLP and K-nearest neighbours. Voting classifier achieved the best accuracy at 71.2%, followed by random forest at 71% accuracy and Embedded Convolutional Neural Network at 69.6% [4] . Using a combined set of neural networks achieved even better results at 71.4% accuracy, downside being significantly prolonged training period and requiring more training data.

It can be concluded that SVM with polynomial and rbf kernel are good algorithms with high accuracy that have been used successfully in other projects. Voting classifier is a promising method that can achieve the highest accuracy with the downside of taking longer training and computational resources, but still faster than typical convolutional neural networks. Thus it is recommended that these algorithms and methods be incorporated into our project to further improve and iterate on our results.

II. MATERIALS AND METHODS

A. Hardware

For this experiment, the system was implemented on an Acer Swift 3 laptop. The technical specifications are listed in Table I.

TABLE I.   SPECIFICATION OF ACER SWIFT 3

Specification	Description
Processor	2.30GHz Hexa-core Intel Core i5(8th Gen) processor
Storage	512GB Solid State Drive Capacity (SSD)
Memory	8GB of LPDDR4 onboard memory
Operating System	Windows 10

B. Programming Language

Python language is used in this research program as python is easy to understand as it has English-like syntax. Besides that, it has free and open source. For example, source code can be downloaded and modified because python has the open-source license which is under OSI approval. It also can improve productivity as it is the productive language which developers can execute the code line by line. Furthermore, it can import almost all the library and it also consists of 200,000 packages in the library.

C. Software

In this research, we are using Anaconda navigator to run the packages as the anaconda navigator allows us to manage the conda packages, environments and channels. Data scientists always use different versions of packages and environments so that anaconda navigator can help them to separate different versions. For instance, we can use the navigator to import the library or packages without typing the conda commands. We also can modify the parameter of the source code and run in the navigator as shown in Table II.

TABLE II.   DATASET DESCRIPTION

S.NO	ATTRIBUTE	TYPE	RANGE
1	Gender	nominal	Male/Female
2	Age	numeric	17-28
3	Openness	numeric	1-8
4	Neuroticism	numeric	1-8
5	Conscientiousness	numeric	1-8
6	Agreeableness	numeric	1-8
7	Extraversion	numeric	1-8

Class label description:
No. of class labels: 5
Type: Nominal
Values: ● Extraverted ● Serious ● Responsible ● Lively ● dependable

D. Original system

The steps involved in the system work are:

- Import all the libraries needed.

- Load the train dataset from the train dataset file.

- Pre-processing of the train datasets

  (a) Data transformation is performed by encode all nominal data type (Male and Female) into binary numbers (0 and 1).

  (b) Among the datasets, first 7 columns are used for training purposes and remaining one column for testing purpose.

- Train the Logistic Regression Classification using the train dataset.

- Load the test datasets from the test dataset file.

- Pre-processing of the test datasets.

  (a) Data transformation is performed by encode all nominal data type (Male and Female) into binary numbers (0 and 1).

- Evaluate the performance of the model (Logistic Regression) by predicting the testing Datasets.

- Stored the prediction output to a file.

III.   RESULT AND DISCUSSION

In order to classify the personalities, a number of techniques are implemented. During the training phase, the models are trained by using the features of Big 5 personality observation. When it comes to validation phase, the model is used to predict the unlabeled features.

A.   Logistic Regression

Logistic regression is one of the classification techniques and it is a common and useful regression method to solve multinomial classification problems which is to handle the issues of multiple classes. For example, we can use logistic regression to predict personality or predict cancer as logistic regression is predictive analysis most of the time. Multinomial regression is used to present the relationship between nominal dependent variable and independent variable.

(multi_class = 'multinomial', solver = 'newton-cg', max_iter =1500)
Multi_class:{'auto', 'ovr', 'multinomial'}

Multinomial allows the minimum loss of the whole probability distribution even if the data is binary. When using the solver 'liblinear', it cannot apply to the multinomial as solver 'liblinear' is accepting binary data only. If the data is binary, it will change to 'auto' or 'ovr' and solver 'liblinear' otherwise it will select 'multinomial'. It is because multinomial logistic regression can generate the variable which is more than two levels. Our assignment is using multinomial as it needs to predict the personality which is more than 2 nominal independent variables.

Solver:{'newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'}

There is 5 solver is multinomial logistic regression which is newton-cg, lbfgs, liblinear, sag and saga. Only 'newton-cg', 'lbfgs', 'sag', 'saga' solver can be applied in multiclass problems in order to handle the multinomial loss and there supports the penalty of 'l2' or none penalty. We have choose 'newton-cg' solver in our assignment and it is the method to exact hessian matrix. Among all of the solvers, 'newton-cg' can generate a higher rate of accuracy when compared with another 4 solvers.

Max_iter : int, default= 100

Maximum number of iterations which are taken by the solver to coverage. In our assignment, we are using 1500 iterations. If we change the parameter of iteration to 1000, the accuracy will become lower. In the multiclass case, the cross-entropy loss is used in the training algorithms if the option is set to the 'multinomial'.
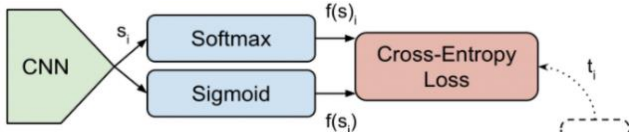

Fig. 1.   cross-entropy loss

$$CE = -\sum_i^C t_i \log(f(s)_i) \qquad (1)$$

$$CE = -\sum_{i=1}^{C'=2} t_i \log(f(s_1)) - (1 - t_1)\log(1 - f(s_1)) \qquad (2)$$

From Fig 1, CNN will have C output neurons and Scores will be gathered in a vector s. After that, it will generate the output of sigmoid and softmax. Sigmoid is also called as logistic function and it will squash the vector in the range of (0,1) and it will apply to the elements of s and s(i). However, softmax is a function but not a loss. It will also squash at the vector which is in the range of (0,1) and the result will add 1 in the vector s which is shown in Figure 3.0.1.2. Lastly, the target will be stored in a vector t and along with a positive class and C- 1 negative class. As it is using multi-class classification, each sample can only belong to one C classes and it is also known as a single classification problem.
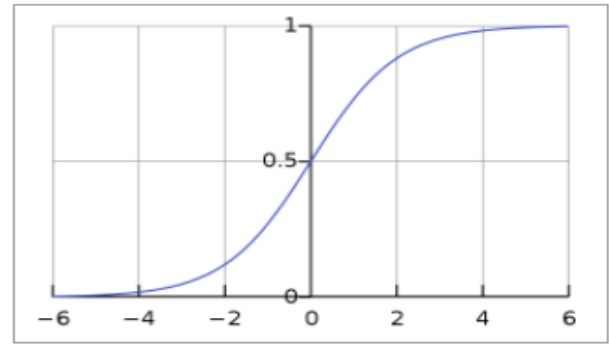

Fig. 2.   Sigmoid and Softmax

$$f(s_i) = \frac{1}{1+e^{-s_i}} \qquad (3)$$

$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \qquad (4)$$

Cross-entropy loss

$$CE = -\sum_i^C t_i \log(f(s)_i) \qquad (5)$$

The value t_i and s_i is the groundtruth of the cross-entropy loss and the value of C is the CNN score of each class. Besides that, $f(s_i)$ refers to the activations and it is applied to the scores when sigmoid and softmax is calculated and it is also applied before calculating the CE loss calculation.

B.   Support Vector Machine

Support vector machines are one of the supervised learning algorithms are mainly used in classifying and solving regression problems which also can be named as support vector classification (SVC) and support vector regression (SVR)[5] . SVM will come out with a hyperplane that split the features into different domain in higher dimensions(). Type of kernel that have been implemented will be kernel polynomial. Fig 1. shows the formula of polynomial kernel.

$$K(X_1, X_2) = (a + X_1^T X_2)^b \qquad (6)$$

x and y will be the input space, a determines the constant while b sets the degree of the polynomial. After applying this formula, x and y will be mapped into a higher dimension Z that may seem like in Fig 2.

$$Z_a = \emptyset(X_a) = (1, a_1, a_2, a_1^2, a_2^2, a_1 * a_2) \qquad (7)$$

$$Z_b = \emptyset(X_b) = (1, b_1, b_2, b_1^2, b_2^2, b_1 * b_2) \qquad (8)$$

To solve the SVM , we would have to perform the dot product on each of the data points and do multiplication with the next dot product, by using kernel trick, dot product can be simply calculated by increasing the value of the power [5] .

$$Z_a^T Z_b = k(X_a, X_b) = (1 + X_a^T X_b)^2 \qquad (9)$$

$$Z_a^T Z_b = 1 + a_1 b_1 + a_2 b_2 + a_1^2 a_1^2 + a_2^2 a_2^2 + a_1 a_1 b_1 b_1 \qquad (10)$$

The 2 dimensional relationships can be used to find a support vector classifier while in a higher dimensional relationship, hyperplane is needed to split the features into different domain. If there are "m" dimensions, the equation of the hyperplane is shown in Equation 11.

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \ldots$$
$$= w_0 + \sum_{i=1}^{m} w_i x_i$$
$$= w_0 + w^T X$$
$$= b + w^T X \qquad (11)$$

Wi will be determine as a vector (W1,W2,W3….Wm) , b will be the biased term which is W0, while X is variables. Thus we can conclude that for point wTx + b >= 0 for di = +1 and wTx + b >= 0 for di = -1. W
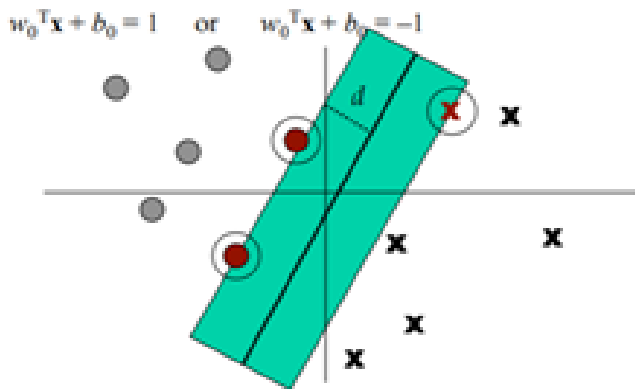

Fig. 3. Hyperplane

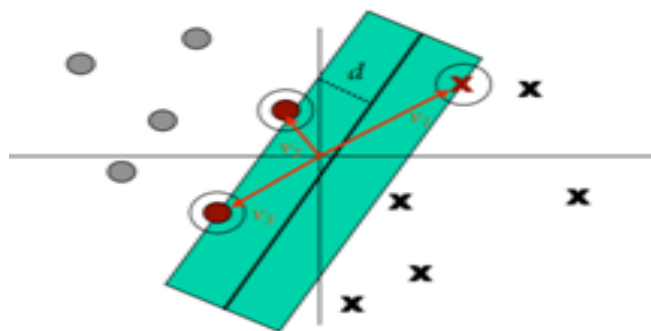The input vector that touch the margin in Fig 3. will be pick as the "tips" of the vectors.

Fig. 4. hyperplane with actual support vector

Finally, by using the algorithm, hyperplane or a line will be created which can separates the data into classes. By using hyperplane can classify the test dataset. SVM is really effective in the higher dimension.

C. Ridge Classification

Ridge Classification is simply ridge regression, just with the response values converted into -1 and +1, then treated as a normal regression task. It's a traditional machine learning algorithm, improving on the least squares regression, often used to handle multicollinearity.

With presence of multicollinear data, regression coefficients will have significantly huge standard errors, which reduces prediction accuracy of the coefficients. A small constant value is modified into original Least Squares estimator, $\beta = (X'X)^{-1} X'Y$, forming:

$$\beta_{ridge} = (X'X + \lambda I_p)^{-1} X'Y \qquad (12)$$

Ridge regression performs L2 regularization by penalizing the square of features' coefficients' magnitude to reduce the error between actual and predicted observations. β^ridge is selected to minimize the penalized square sums:

this is equivalent to minimization

$$\sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2) \qquad (13)$$

Of $\sum_{i=1}^{n} (y_i - \sum_{j=i}^{p} X_{ij} \beta_j)^2 \qquad (14)$

subject to, for some, $C > 0, \sum_{j=1}^{p} \beta_j^2 < c. \qquad (15)$

Therefore, by placing a constraining/penalty term on certain parameters, ridge regression further minimizes the residual sum of squares. A constant is chosen as the penalty term, which is multiplied with the squared vector. The larger the value of the vector, the more the optimization function is penalized.

D. Naive Bayes

Naïve bayes is one of the statistical algorithms used for classification based on Bayes Theorem. For example, a fruit may be considered as orange if it is in orange colour and 3 inches in diameter. Even though these features will rely on each other, these properties will be contributing independently to the probability that the fruit is orange, and this is the meaning of why it is known as "Naïve' [3].

Naive bayes algorithm commonly applies to large amounts of data. Naïve Bayes algorithm is suitable in classifying various applications. For example, text classification, recommender system, spam filtering and so forth. It is fast and accurate, providing high accuracy of prediction. A few models of naïve bayes had been created to perform different tasks based on suitability. Equation below shows the formula for calculating probability that a document occurs in a class by using multinomial Naïve Bayes classification.

$$P(t|d) = P(C) \times P(t_1| \times c) \times P(t_2| \times c) \times P(t_3| \times c) \ldots \times P(t_n| \times c) \qquad (16)$$

In order to calculate the probability that a document occurs in a class, prior probability and probability of nth word are needed to calculate [6] . Equation 16 shows the formula of prior probability P(C). Equation 17 shows the formula of probability of the nth word.

$$P(C) = \frac{N_c}{N} \qquad (17)$$

$$P(t_n|C) = \frac{count(t_n,c)+1}{count(c)+|V|} \qquad (18)$$

These are the probability was being specific in a particular category from a set of documents. Equation 18 will show the formula for the calculation of TF-IDF (term frequency-inverse document frequency). This is a statistical method to observe how relevant a word is to its document [6] .

$$tfidf_t = f_{t,d} \times log\frac{N}{df_t} \qquad (19)$$

Equation above will show the formula for calculating conditional probability which is known as likelihood. It is calculating the conditional probability of a word appearing in a document that the document belongs to a class.

$$(t_n|C) = \frac{W_{ct}+1}{\sum_{W' \in V} W'_{ct}+B'} \qquad (20)$$

E. Voting Classifier

Also known as an ensemble method, voting classifier is a wrapper for a set of different models that are trained in parallel. It then predicts the output class based on highest probability. Ensemble methods tend to produce lesser error and result in less over-fitting. There are 2 basic types of voting classifiers:

$$y = mode\{C_1(x), C_2(x), \ldots, C_{m(x)}\} \qquad (21)$$

Majority voting: Simplest case, the class that receives the highest majority of votes by each model predicted is the output class.

$$y = arg\ _i^{max} \sum_{j=1}^{m} w_j p_{ij} \qquad (22)$$

where wj is the weight that can be assigned to the jth classifier

Soft voting: The probability vector for each predicted class from all the classifiers are summed up and averaged. Soft voting is recommended only if each classifier is well-calibrated. Voting classifier makes the most of the different algorithms and if done right should yield better performance than any single model. It is important that the set of classifiers are diverse so errors do not aggregate.

IV.   DISCUSS ON IMPLEMENTATION

After the pre-processing of data is done, five types of algorithms are implemented. Nominal types of data were encoded into binary where 1 represents male, 0 represents female. Hence, there are 7 types of prediction models were append to an model array, which were Logistic Regression, Multinomial Naive Bayes, Support Vector Machine,

Polynomial Kernel (SVM), Radial Basis Function Kernel(SVM), Ridge Regression and Voting Classifier.

All the prediction models were trained by using the train datasets. After the process of train models, the predictions model were used to predict the test datasets. Performance result of each prediction model will then print as a classification report.

V.   RESULT

To evaluate the performance of different classifiers, in order to make the result more precise, there are some different metrics used in the result, recall, F1-score, and accuracy.

Classifier: Logistic Regression

Classification Report:

TABLE III.   PERFORMANCE METRIC FOR LOGISTIC REGRESSION

	Precision	Recall	F1-Score	Support
Accuracy			0.86	315
Macro Avg.	0.87	0.72	0.77	315
Weighted Avg.	0.86	0.86	0.85	315

Classifier: Multinomial Naive Bayes

Classification Report:

TABLE IV.   PERFORMANCE METRIC FOR MULTINOMIAL NAIVE BAYES

	Precision	Recall	F1-Score	Support
Accuracy			0.67	315
Macro Avg.	0.47	0.39	0.38	315
Weighted Avg.	0.56	0.67	0.58	315

Classifier: Support Vector Machine (SVM)

Classification Report:

TABLE V.   PERFORMANCE METRIC FOR *FOR SVM*

	Precision	Recall	F1-Score	Support
Accuracy			0.72	315
Macro Avg.	0.71	0.47	0.51	315
Weighted Avg.	0.73	0.72	0.68	315

Classifier: SVM with Polynomial kernel

Classification Report:

TABLE VI.   PERFORMANCE METRIC FOR SVM WITH POLYNOMIAL KERNEL

	Precision	Recall	F1-Score	Support
Accuracy			0.57	315
Macro Avg.	0.35	0.28	0.27	315
Weighted Avg.	0.51	0.57	0.48	315

Classifier: SVM with RBF kernel

Classification Report:

TABLE VII.   PERFORMANCE METRIC FOR SVM WITH RBF KERNEL

	Precision	Recall	F1-Score	Support
Accuracy			0.58	315
Macro Avg.	0.24	0.28	0.25	315
Weighted Avg.	0.43	0.58	0.48	315

Classifier: Ridge Classifier

Classification Report:

TABLE VIII.   PERFORMANCE METRIC FOR RIDGE CLASSIFIER

	Precision	Recall	F1-Score	Support
Accuracy			0.83	315
Macro Avg.	0.90	0.67	0.71	315
Weighted Avg.	0.85	0.83	0.81	315

Classifier: Voting Classifier

Classification Report:

TABLE IX.   PERFORMANCE METRIC FOR VOTING CLASSIFIER

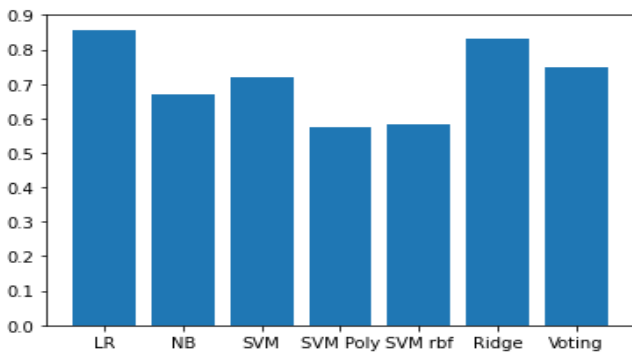|	Precision	Recall	F1-Score	Support
Accuracy			0.75	315
Macro Avg.	0.78	0.51	0.55	315
Weighted Avg.	0.77	0.75	0.71	315

Fig. 5.   Bar chart of accuracy for each algorithm

Based on Fig 5, Logistic Regression remains the best performing algorithm at 85%. This is followed by Ridge Classifier at 83%, and then Voting classifier at 75%. Logistic Regression continues to outperform the other models because as a linear regressor it does not overfit to the training model. Ridge also performed well because it avoids overfitting a model by using L2 regularisation and can deal with multicollinearity. Voting classifier came in third, proving that ensemble methods aren't always necessarily better than the individual classifiers, especially when some of the classifiers are subpar.

SVM RBF KERNEL AT 58% AND SVM POLYNOMIAL KERNEL 57% PERFORMED. It is also possible that the dataset problem is very close to being linearly separable. From this study, we know that the ensemble could be worse than the individual models. For example, taking the average of the true model and a bad model would give a fairly bad model. There is no absolute guarantee a ensemble model performs better than an individual model, but if you build many of those, and your individual classifier is weak. Your overall performance should be better than an individual model. real gains come from so-called unstable models such as decisions trees, where each observation usually has an impact on the decision boundary. More stable ones like SVMs do not gain as much because resampling usually does not affect support vectors much. Though It tends to improve performance on average.

VI.   CONCLUSION

In this research, we have studied 5 machine learning classifiers to classify the human personality. The main purpose of this research is to study the new algorithms and improve the accuracy of the original algorithm which is Logistic Regression. New algorithms have been added are Support Vector Machine, Naive Bayes Classifier, Ridge Classifier, Voting Classifier. Precision, recall, accuracy and f1 score are used to measure the performance of all the classifiers. The overall results have shown that Logistic Regression still outperformed all new algorithms added. Moreover, predicting personality is an abstract model which means that it will describe the phenomena while concrete models will have a direct analogue result in machine learning. Therefore, abstract models cannot obtain a very high accuracy result when compared to concrete model.

REFERENCES

[1] A. Chitlangia, G. Malathi, "Handwriting Analysis based on Histogram of Oriented Gradient for Predicting Personality traits using SVM," Procedia Computer Science 165, pp. 384-390, 2019.

[2] S. Ramli, S Nordin, "Personality Prediction Based on Iris Position Classification Using Support Vector Machines," Indonesian Journal of Electrical Engineering and Computer Science, 9(3), pp. 667-672, 2018.

[3] A. P. Wibawa, A. C. Kurniawan, D. M. P. Murti, R. P. Adiperkasa, S. M. Putra, S. A. Kurniawan, Y. R. Nugraha, " Naïve Bayes Classifier for Journal Quartile Classification" International Journal of Recent Contributions from Engineering, Science & IT (iJES), 7(2), pp. 91-99, 2019.

[4] M. Allouch, A. Azaria, R. Azoulay., "Detecting Sentences that May be Harmful to Children with Special Needs," 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1209-1213, 2019.

[5] N. Guenther, M. Schonlau.," Support vector machines," The Stata Journal, 16(4), pp.917-937, 2016.

[6] Y. Artissa, I. Asror, S.A. Faraby.," Personality Classification based on Facebook status text using Multinomial Naïve Bayes method," In Journal of Physics: Conference Series, vol. 1192, no. 1, p. 012003. IOP Publishing, 2019.