**SUPPORT VECTOR MACHINE ANALYSIS FOR INTEREST AND TALENT CLASSIFICATION WITH PYTHON LIBRARY**

**Devi Sartika1, Febie Elfaladonna2\*, Andre Mariza Putra3**

1,2,3 Manajemen Informatika, Politeknik Negeri Sriwijaya

*email*: febie\_elfaladonna\_mi@polsri.a.c.id

**Abstract:** Recognizing one's interests and talents early on is crucial in guiding an individual toward a prosperous future. While distinct, interests and talents share a close relationship. Interest denotes a genuine attraction to something without external pressure, and when consistently nurtured, it evolves into a skill or talent. Machine learning, specifically utilizing the SVM algorithm with the RBF kernel, can be applied to categorize interests and talents. Prior to SVM modeling, conducting Exploratory Data Analysis (EDA) is imperative for scrutinizing interests and talents. This analysis facilitates the identification of variables, enabling the elimination of missing values and ensuring the selection of appropriate interest and talent variables. The primary objective is to achieve optimal accuracy in modeling the classification of interests and talents. The insights gained from this research contribute to the creation of an application designed for categorizing interests and talents within SDN XYZ school. This application is designed for student use, assisting them in making informed decisions about their future education and career paths

**Keywords:** Interests and Talents, Machine Learning, SVM Algorithm, Exploratory Data Analysis

**Abstrak:** Mengenali minat dan bakat seseorang sejak dini sangat penting dalam membimbing individu menuju masa depan yang sukses. Meskipun berbeda, minat dan bakat memiliki hubungan yang erat. Minat mengindikasikan ketertarikan yang tulus terhadap sesuatu tanpa tekanan eksternal, dan ketika terus-menerus dibina, berkembang menjadi keterampilan atau bakat. Pembelajaran mesin, khususnya dengan menggunakan algoritma SVM dan kernel RBF, dapat digunakan untuk mengelompokkan minat dan bakat. Sebelum pemodelan SVM, melakukan Analisis Data Eksploratif (EDA) sangat penting untuk mengkaji minat dan bakat. Analisis ini memfasilitasi identifikasi variabel, memungkinkan penghilangan nilai yang hilang, dan memastikan pemilihan variabel minat dan bakat yang tepat. Tujuan utamanya adalah mencapai akurasi optimal dalam pemodelan klasifikasi minat dan bakat. Temuan dari penelitian ini berkontribusi pada pengembangan aplikasi yang ditujukan untuk mengkategorikan minat dan bakat di sekolah SDN XYZ. Aplikasi ini dirancang untuk digunakan oleh siswa, membantu mereka membuat keputusan yang terinformasi mengenai pendidikan dan karier masa depan mereka.

**Kata kunci:** Minat dan Bakat, Machine Learning, Algoritma SVM, Exploratory Data Analysis

**INTRODUCTION**

Interests and talents, two distinct yet often interrelated aspects, are typically managed by parents to enhance the future quality of their children's lives. Education is commonly employed by parents as a method to enrich their children's interests and talents. Education plays a crucial role in self-development and the improvement of human quality, shaping character and knowledge [1].

Interest, as an expression of liking or attraction without pressure, reflects self-acceptance of things in the surroundings [2]. The close and strong connection between interest and external experiences can strengthen an individual's interests [3]. On the other hand, talent involves abilities that require specific training, such as expertise, skills, and knowledge in a particular field. For example, someone interested in singing and undergoing vocal training demonstrates talent in the field of singing.

SDN XYZ, an elementary school in Palembang City, actively participates in various competitions at the district, regency, and even city levels. However, the management of students' interests and talents in this school still relies on a selection technique based on the observation of class teachers, without a clear classification system. The lack of mapping students' skills may be a cause of the school's inability to achieve excellence. Digitizing the interest and talent classification system through an application can have a positive impact on the school's progress, helping to achieve the goal of producing outstanding and highly competitive alumni.

The choice of method is crucial in developing an application for interest and talent classification. Currently, machine learning, especially the Support Vector Machine algorithm, is a common choice. Classification in machine learning serves to group data into specific classes or targets. Machine learning is a mathematical algorithm used to predict the future based on data and applied to computers. The learning process in machine learning consists of two main stages, namely testing and training (Ahmad Roihan, 2020). There are three categories of machine learning that have been investigated, namely Supervised Learning, Unsupervised Learning, and Reinforcement Learning [4]. Supervised Learning is a classification technique in which all data is labelled to facilitate the classification of unknown classes. In contrast, Unsupervised Learning, known as clustering method, does not require labelling the data and generates clusters without identifying known classes [5]. Reinforcement learning meanwhile uses methods that can operate in dynamic environments and achieve goals without waiting for instructions from a computer. This category often includes supervised and unsupervised aspects together. Support Vector Machine (SVM) is an algorithm designed to recognize a maximal hyperplane, which is a function that can separate two classes. SVM creates the maximum margin between the training pattern and the decision boundary [6]. The advantages of using SVM involve superior performance for both small and large amounts of data, as well as for data that has many attributes [7]. Although initially utilized for binary classification, SVMs have evolved to classify multiple classes and can be used in regression and outlier detection.

In the context of machine learning, we also face Exploratory Data Analysis (EDA), an analytical process that aims to improve understanding of data. EDA steps include providing maximum insight into the data, creating a data structure, extracting key variables, detecting anomalies and outliers, testing hypotheses, determining models, and achieving factor optimality [8].

Python has become a programming language that is in high demand by large corporations and developers to build a variety of applications, including desktop, web, and mobile-based ones. Created by Guido van Rossum in the Netherlands in 1990, Python takes inspiration from his favourite television show, Monty Python's Flying Circus. Although originally developed as a hobby, Python has since gained great popularity in industry and education due to its simplicity, conciseness, intuitive syntax, and abundant library support [9]. Python libraries collect related modules that contain a number of codes that can be used repeatedly in various programmes. The existence of this library makes it easier and more convenient for programmers, as there is no need to repeat writing the same code for each different programme [10]. This research uses several libraries provided by python such as Pandas, NumPy, SciPy, Sklearn, and others.

The use of the Support Vector Machine algorithm to address the classification issues of interests and talents in SDN XYZ students is the focus of this research. Through this study, it is hoped to produce data analysis that is beneficial as a foundation for the development of a more effective application for interest and talent classification.

**METHOD**

There are five stages in the research that are part of the machine learning process. The following is an explanation for each stage:

**Data Identification and Preparation Stage**

It takes a logical and mathematical approach to address the problem at hand. Data identification is necessary to form the basic question and ensure focus on the initial topic. Data availability plays a crucial role in machine learning modelling, where the more data available, the more valid the output. Data collection methods can use data mining and web scraping. In this study, data was collected through a questionnaire distributed to parents of SDN 204 Palembang students, consisting of 15 statements related to children's favourites. The answers to the statements are ordinal, reflecting the parents' level of belief in the statement. The number of respondents was 144 parents, and the data was processed manually by the school to classify interests and talents based on the knowledge of the counselling teacher.

**Data Acquisition**

Derived from a Google Form questionnaire consisting of 15 statements, resulting in the classification of two classes, namely interest and talent.

Table 1. Statements on the Questionnaire

|  |  |  |
| --- | --- | --- |
| **Id** | **Pernyataan** | |
| Q1 | | Student's grade |
| Q2 | | Age of student |
| Q3 | | Critical thinking |
| Q4 | | Likes trading and math |
| Q5 | | Likes certain types of sports |
| Q6 | | Likes to play musical instruments or sing |
| Q7 | | Likes to cook |
| Q8 | | Likes to write, read or tell stories |
| Q9 | | Likes to draw, paint or color |
| Q10 | | Often participate in coloring, drawing or painting competitions |
| Q11 | | Often participate in storytelling, speech competitions. Or poetry |
| Q12 | | Often participates in singing or acting competitions |
| Q13 | | Good at singing and playing music |
| Q14 | | Proficient in drawing, coloring, and painting |
| Q15 | | Proficient in acting activities |

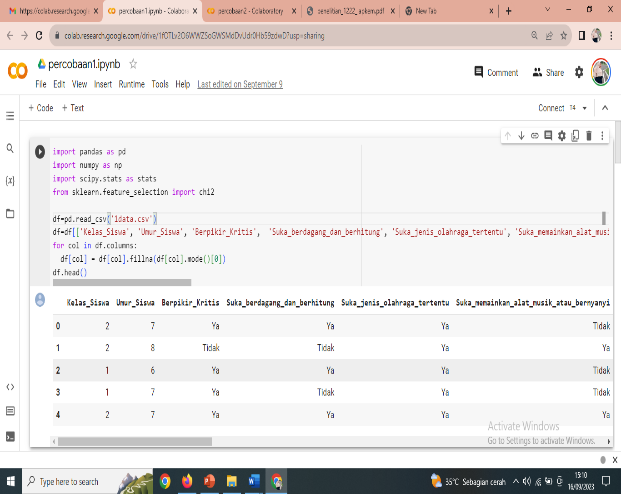
Questions are answered using a Likert scale, a common psychometric measurement method on questionnaires and often used in surveys. The counselling teacher scores five response options, with confidence levels from 0 to 100, to reflect the extent to which the statements in the questionnaire reflect the child's specific interests or talents.

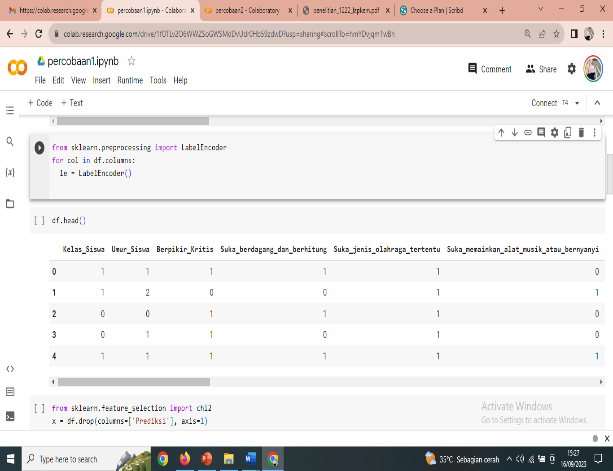
Tabel 2. Skala Likert Kuisioner

|  |  |
| --- | --- |
| **Scale** | **Description** |
| 0 | No idea |
| 20 | Probably |
| 40 | Most likely |
| 60 | Almost Certain |
| 80 | Definite |
| 100 | Very certain |

Top of Form

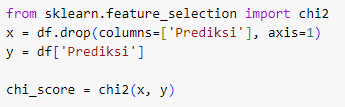
The next step involves transforming the survey data into a tabular format, where data tabulation refers to the process of converting information from questionnaires into tables. The aim is to make data analysis easier, especially since respondents' answers are recorded on a Likert scale. In this stage, there are two main variables, the dependent and independent variables. The data acquisition process uses several libraries in Jupyter Notebook, including Pandas, NumPy, and SciPy. Pandas, as an open-source library, works for data analysis and provides a data structure called DataFrame. DataFrame is required to clean the raw data and convert it into a format suitable for analysis.

Figure 1. Dataset view with Python library code

Figure 2. Running Python's LabelEncoder Function

**Exploratory Data Analysis**

In this research, the data exploration step involves analysing all the features in the dataset regarding the interests and talents of primary school students. This analysis uses the chi-square test and the T-test, where theoretically, chi-square is part of statistical testing that aims to assess how significant the relationship between two variables, namely the dependent variable and the independent variable. The symbol used for chi-square is x2. Meanwhile, the T-test is also a component of statistical testing, but it focuses on how significant the difference between the variables is. This allows the formation of hypotheses regarding whether the relationship between the variables is acceptable or not.

Figure 3. Run the chi square test sklearn function

The results of the chi-square code execution as shown in the figure above are:



Figure 4. Chi square results

In addition to producing a chi-square value which is an implementation of "chi2", the output also includes T-test results that show how significant the effect of the independent variable is on the dependent variable.

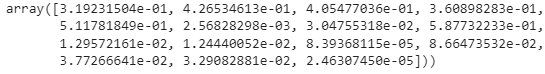


Figure 5. T-Test Results

The table below reflects the chi score test results for all variables.

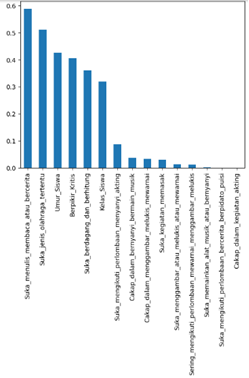
Table 3. Test Results of Comprehensive Variables

|  |  |
| --- | --- |
| **Variable** | **Value** |
| Student Class | 0.99209243 |
| Age of Student | 0.63224066 |
| Critical Thinking | 0.69202329 |
| Likes trading and math | 0.83476502 |
| Likes certain sports | 0.43042221 |
| Likes to play musical instruments and sing | 9.09129968 |
| Love Cooking | 4.68226551 |
| Likes to write, read or tell stories | 0.29389977 |
| Likes to draw, paint or color | 6.17494196 |
| Often participates in coloring, drawing, or painting competitions | 6.24648434 |
| Often participates in storytelling, speech, or poetry competitions | 15.46748504 |
| Often participates in singing or acting competitions | 2.93558343 |
| Good at singing and playing music | 4.31730334 |
| Good at drawing, coloring, and painting | 4.55054406 |
| Proficient in Acting Activities | 17.79284731 |

Table 4. T-Test Results

|  |  |
| --- | --- |
| **Variable** | **Value** |
| Proficient in Acting Activities | 17.79284731 |
| Often participates in storytelling, speech, or poetry competitions | 15.46748504 |
| Likes to play musical instruments and sing | 9.09129968 |
| Often participates in coloring, drawing, or painting competitions | 6.24648434 |
| Likes to draw, paint, or color | 6.17494196 |
| Likes to Cook | 4.68226551 |
| Good at drawing, coloring, and painting | 4.55054406 |
| Good at singing and playing music | 4.31730334 |
| Often participates in singing or acting competitions | 2.93558343 |
| Student Class | 0.99209243 |
| Likes to trade and do math | 0.83476502 |
| Critical Thinking | 0.69202329 |
| Age of student | 0.63224066 |
| Likes certain sports | 0.43042221 |
| Likes to write, read or tell stories | 0.29389977 |

Here's a diagram of the above results:

Figure 6. T-Test Graphics for Interest and Talent Variables

In this context, the higher the chi score value, the better the correlation between the variables. Chi Score and T-Test tests were conducted twice in this study, with the same dependent and independent variables. However, the independent variables used in subsequent tests are those selected after the first chi score and T-Test test.

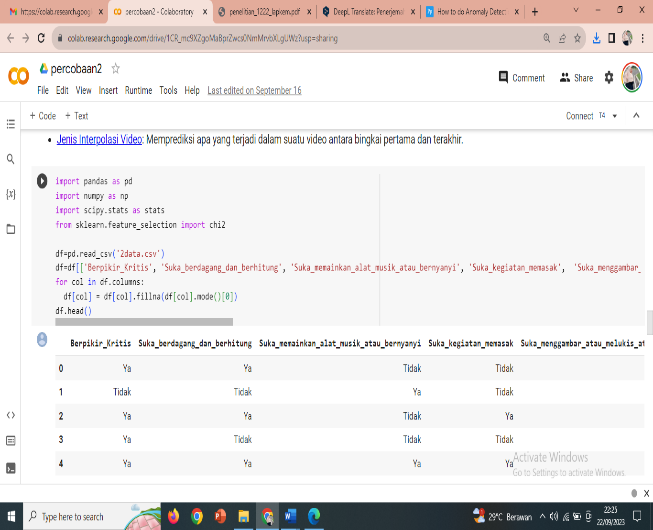


Figure 7. Import Library Code

Figure 8. Chi Square and T-Test Results

**Anomaly Identification**

Anomaly identification is an activity to recognise and capture inappropriate data from an irregular path, with the aim of weeding out irrelevant data. This data can be re-examined from a different perspective or deleted to keep the data clean before further processing. The chi square and T-Test results showed that of the many variables tested, there were only nine independent variables that were suitable for further processing.

**Recommended Variables**

Based on the results of the Chi square and T-Test tests, there are two dependent variables, namely "prediction" labelled as interest and talent, and nine independent variables that fall into the recommended variable category. The nine variables that passed the Chi square and T-Test tests are: "good at acting", "often participate in writing, reading, or storytelling activities", "playing musical instruments and singing", "often participate in colouring, drawing, or painting competitions", "drawing, painting, or colouring", "cooking activities", "good at drawing, painting, and colouring", "good at singing while playing musical instruments", "often participate in singing and acting competitions".

**RESULT AND DISCUSSION**

**Support Vector Machine with Python**

After completing the exploratory data analysis and identifying the recommended variables, the next step is to perform data analysis using the Support Vector Machine (SVM) algorithm. The SVM data analysis process was carried out using the Python programming language. There are four libraries used in Python, namely pandas, numpy, matplotlib, and seaborn. The function of matplotlib is to transform raw data into valuable information through various forms of graphs and diagrams, while seaborn is used to create graphs and statistics using Python.

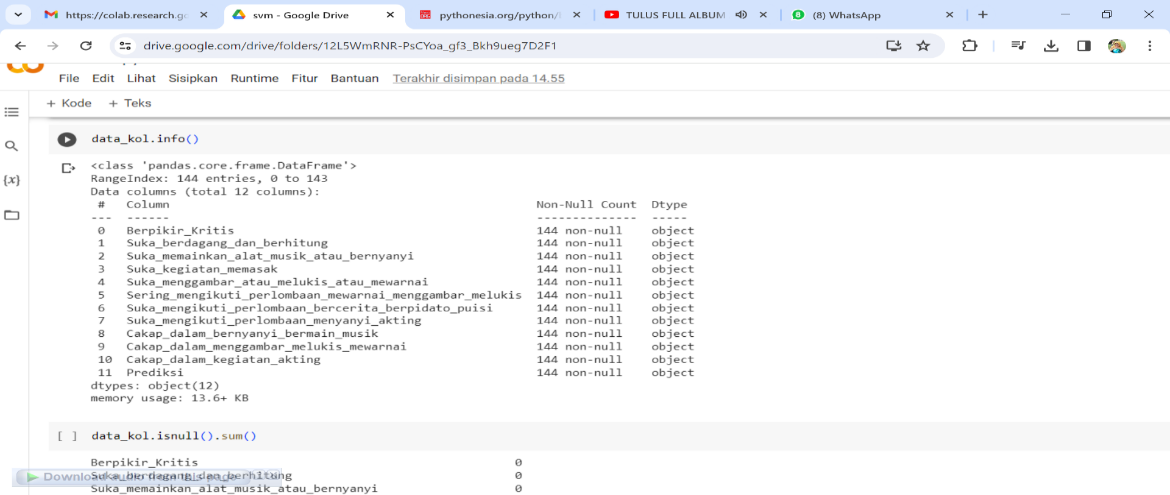


Figure 9. Data Frame

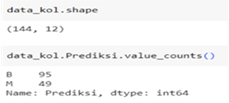


Figure 10. Collective Data Informatio

The figure above illustrates the number of rows in the dataset, reaching 144 data with 12 variables. In addition, this data is classified by aptitude prediction, with 95 people, and interest prediction, with 49 people.

**Data Preparation**

Some variables have an object data type. In this step, variable values are transformed using the label encoder function available in the scikit-learn library. This action aims to make it easier to observe the correlation between variables.



Figure 11. Correlation Between Variables in the Dataset

If the relationship between the variables in the above figure shows a strong linear pattern, then the correlation tends to be close to -1 or 1. Conversely, if the relationship is non-linear, the correlation will be close to 0 or show an irregular pattern. The dataset used shows an imbalance in class distribution. To address the imbalance, the SMOTE (Synthetic Minority Over-Sampling Technique) oversampling technique was applied. This process involves finding the nearest neighbour of the minority sample which becomes the reference in generating a new synthetic sample based on the pattern of the pre-existing sample. After the application of the SMOTE oversampling technique, the class distribution becomes balanced, as seen in the following figure:

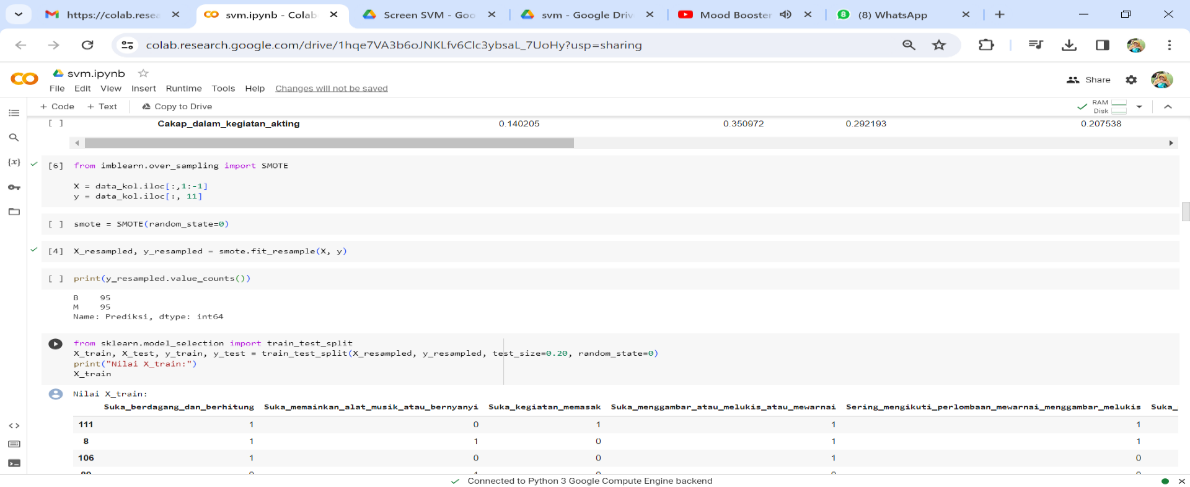


Figure 12. SMOTE process

In the early stages of modelling, the dataset was divided into 60% training data and 40% testing data. This division of data is done with more proportion for training data so that the model can be trained optimally. With more training data, the model has the opportunity to find more complex patterns and relationships in the dataset, thus increasing the model's ability to make predictions on data that has never been seen before.

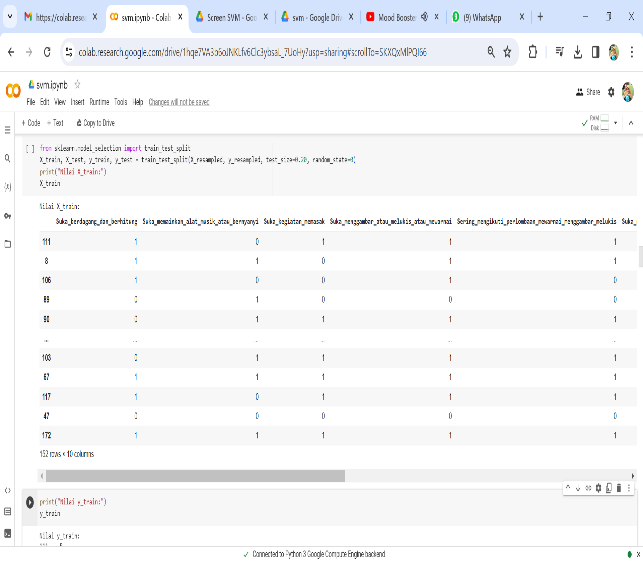


Figure 13. Training Data (X\_train and y\_train)

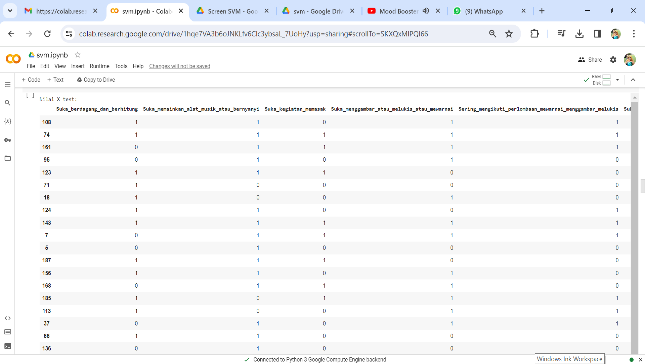


Figure 14. Data Testing (X\_test dan y\_test)

The dataset was split into a training set and a testing set using the 'train\_test\_split' function from the 'model\_selection' module in the scikit-learn library. The variable X\_train stores the subset of features from the training data, while X\_test stores the subset of features from the testing data. The variable y\_train is used to store the target subset of the training data, and y\_test stores the target subset of the testing data. The next step involves model building using a Support Vector Machine (SVM). SVM is not only effective on linear data but is also able to cope with non-linear data by applying kernel functions. Some of the commonly used kernel functions in SVM include Linear kernel, Gaussian kernel, RBF kernel, Polynomial kernel, and Sigmoid kernel. In this study, RBF kernel is used because it can handle classification problems on data that cannot be linearly separated. The RBF kernel performed well in the dataset used, with the training results showing a low error value. Furthermore, parameter adjustments were made to the SVM to obtain the optimal C and gamma values.

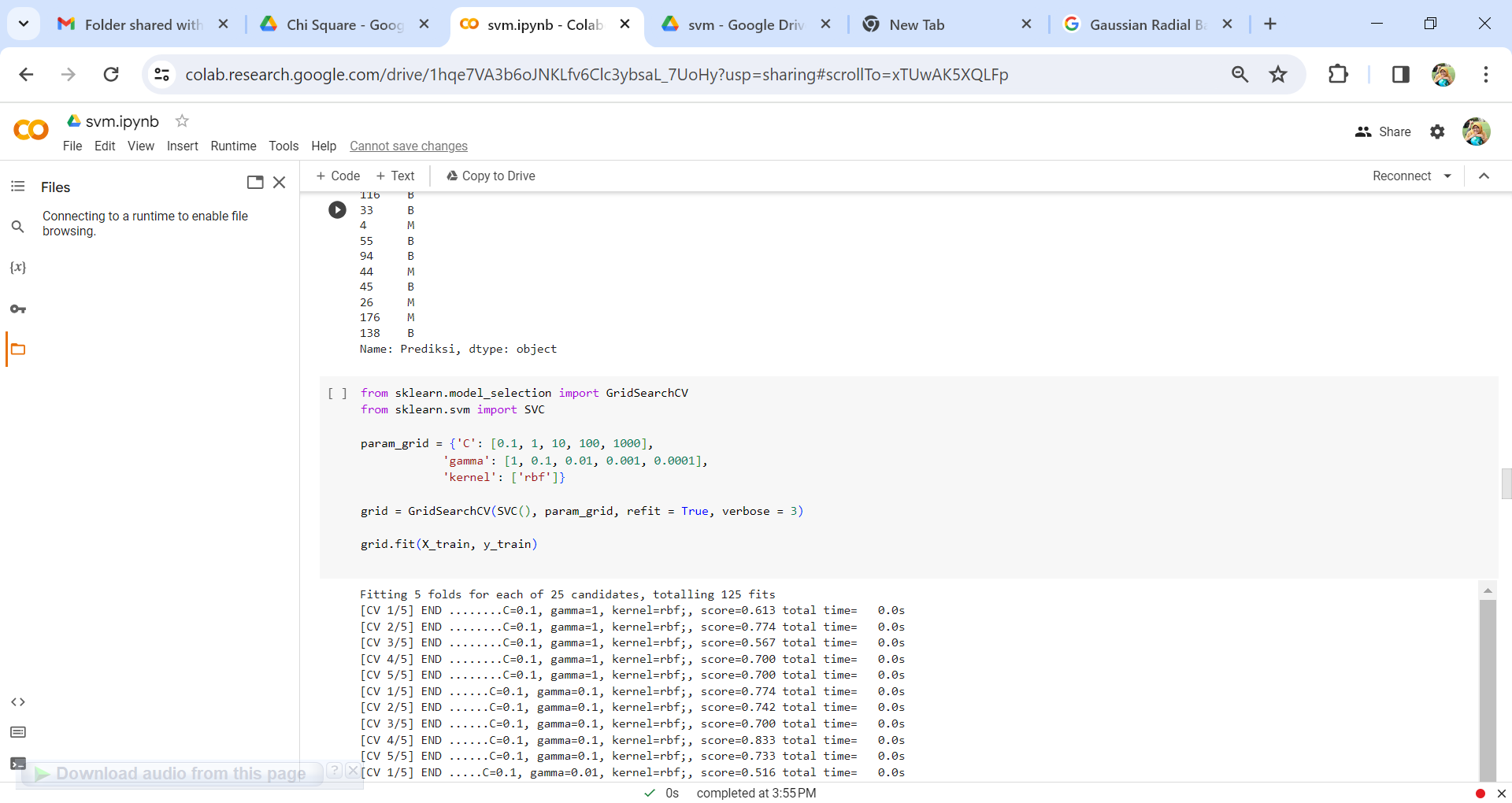
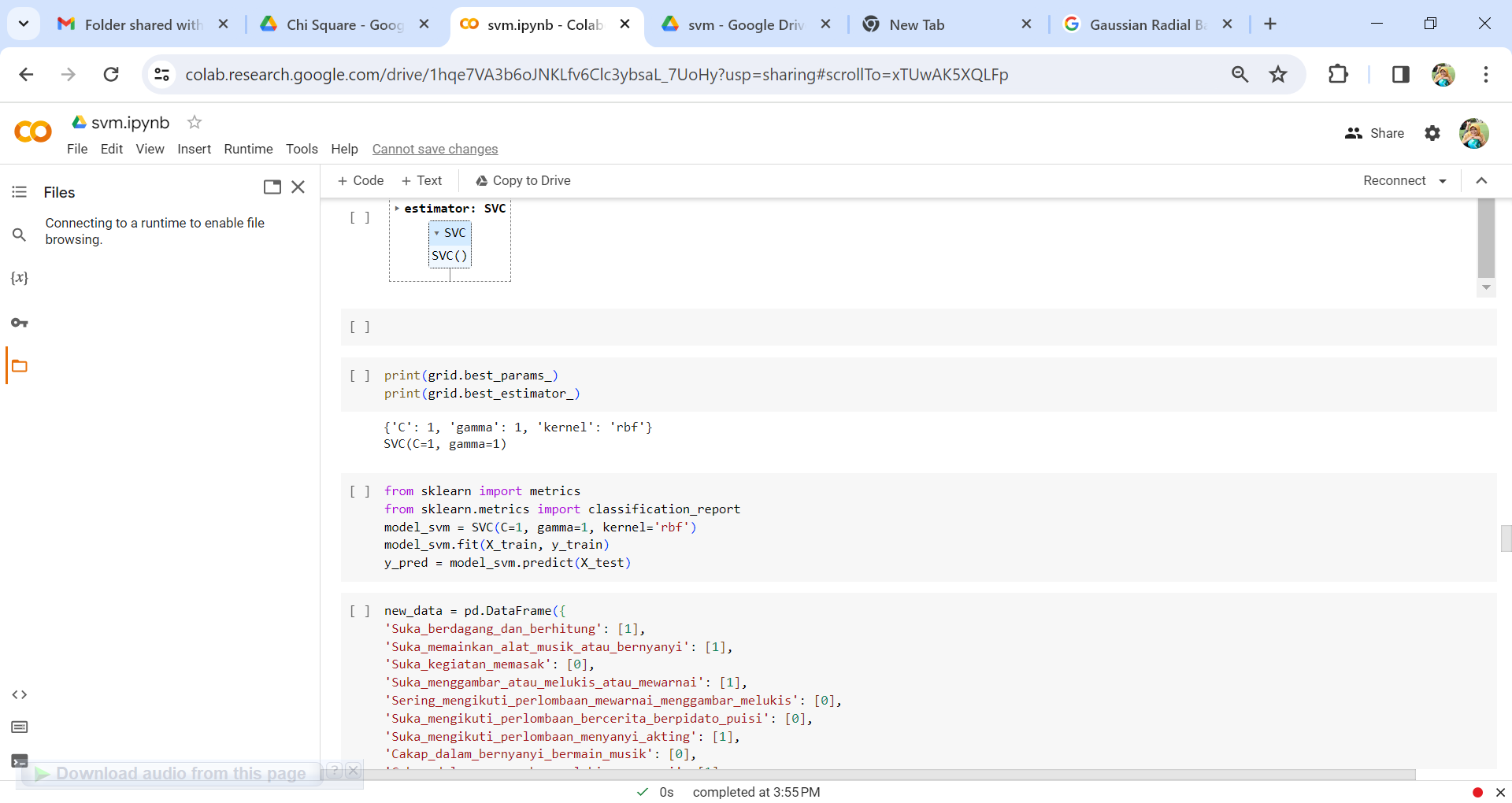
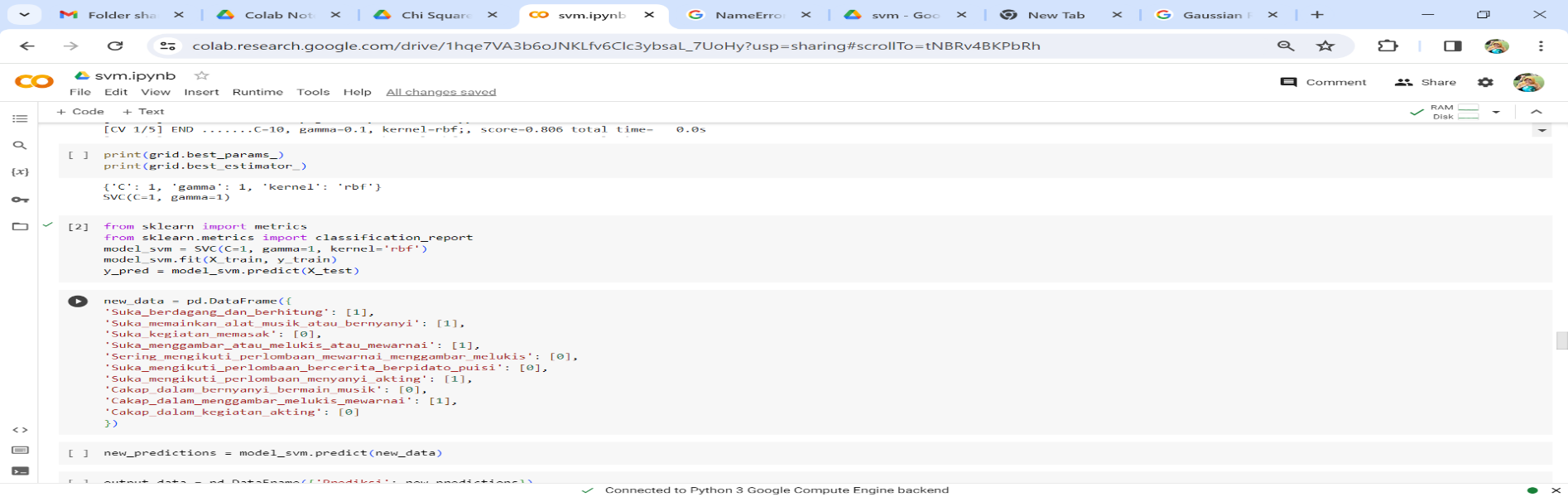


Figure 15. Parameters Tested on SVM

Parameter tuning can be observed through the application of the GridSearchCV function from the scikit-learn library. GridSearchCV performs a systematic search across the parameter space to find the best combination of parameters in the machine learning model. The tested C values include 0.1, 1, 10, 100, and 1000, while the tested Gamma values include 1, 0.1, 0.01, 0.001, and 0.0001.

Top of Form

Figure 16. Best Parameter Tuning

Figure 17. Building the SVM Model

The next step is to evaluate the performance of the SVM model using Confusion Matrix. Confusion Matrix is a table used to illustrate the predictive quality of a model, by comparing the predicted results with the actual values of the test data.

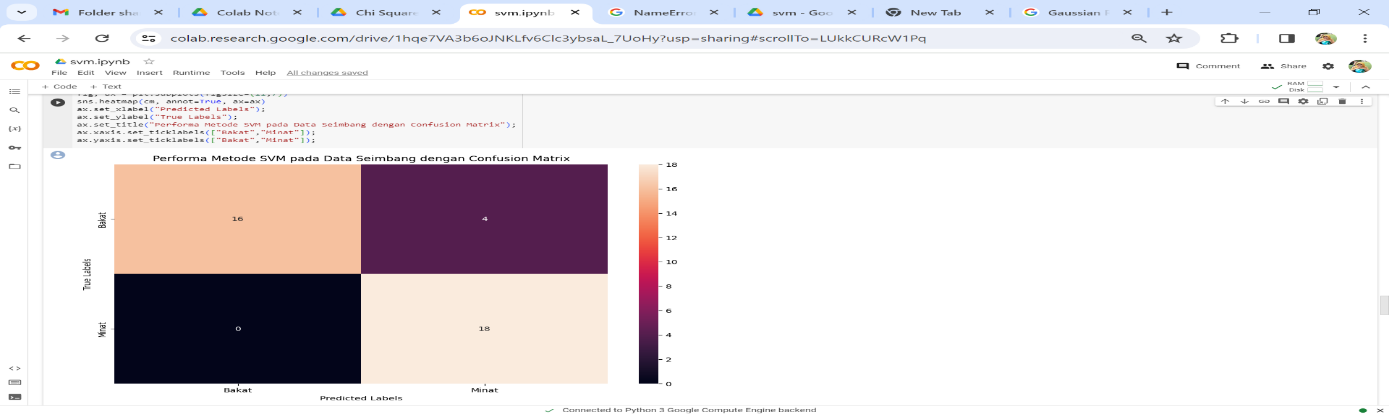
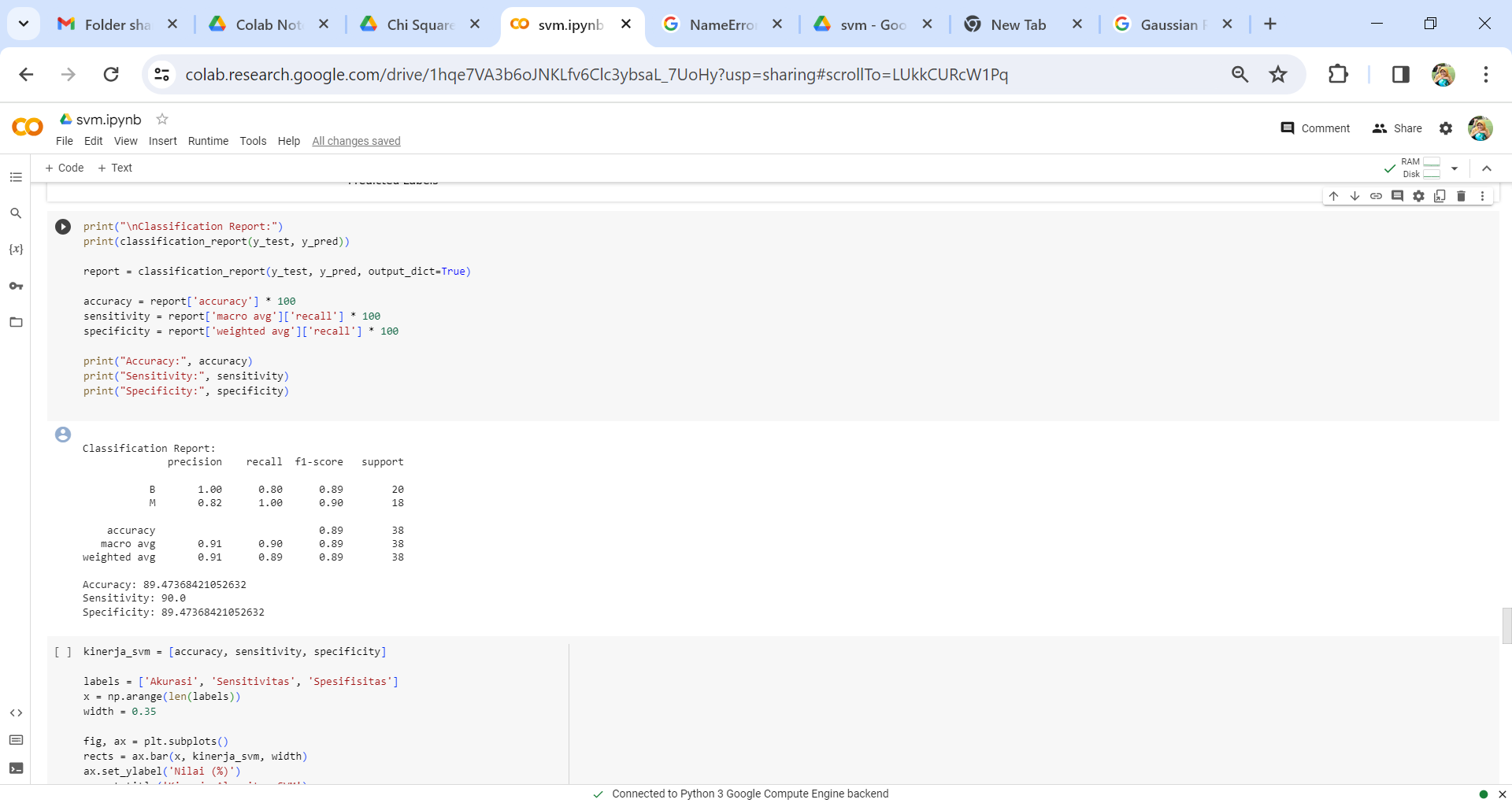
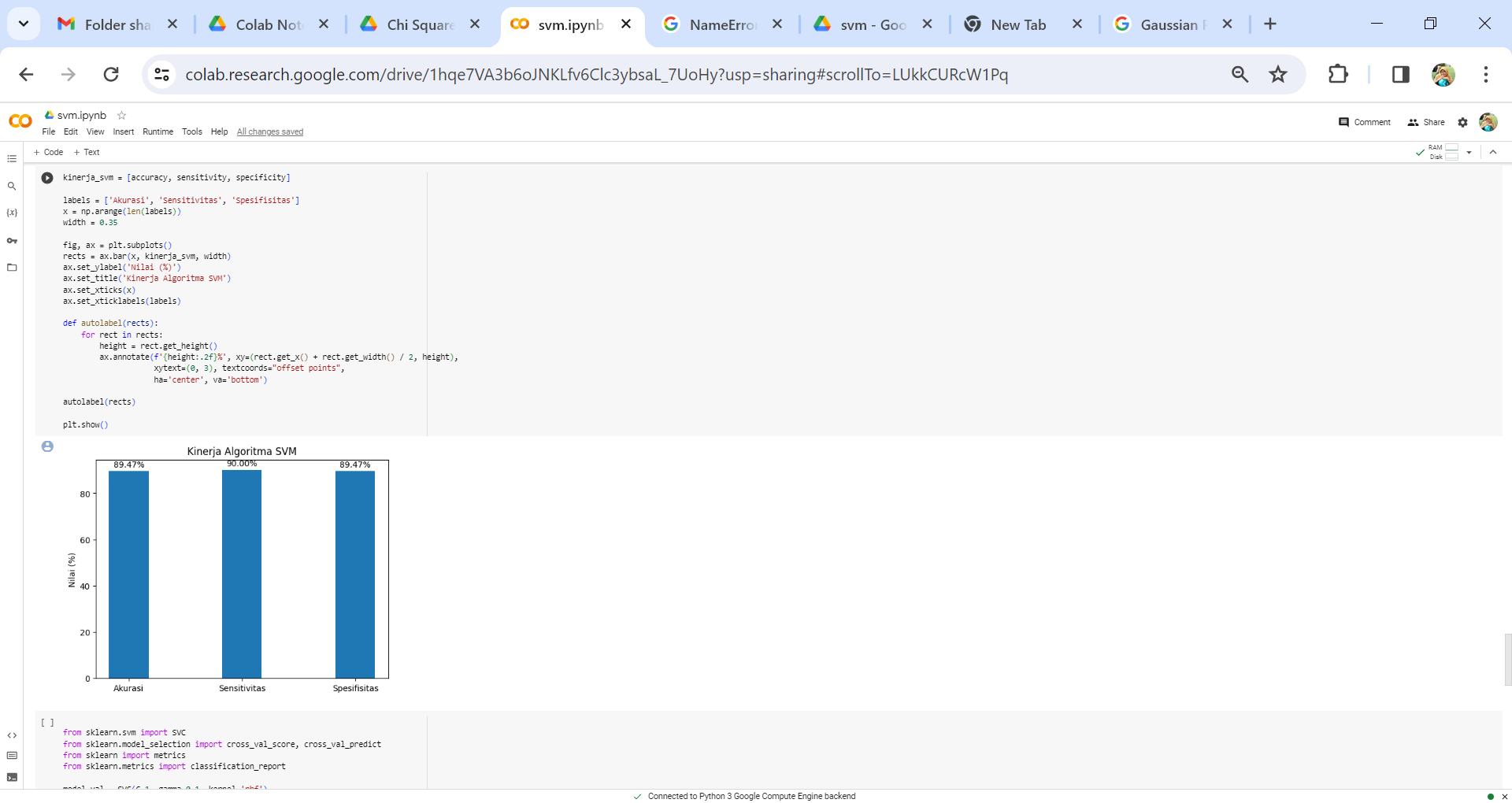


Figure 18. SVM Performance Test Using Confusion Matrix

The results of the confusion matrix in classifying interest and talent status with the Support Vector Machine method are illustrated in Figure 18. Furthermore, classification report is used to obtain a detailed report on the performance of the classification model in predicting interest and talent classes, as shown in the following figure.

Figure 19. Classification Report of SVM

Precision for the talent class reaches a value of 1.00, while precision for the interest class reaches a value of 0.82. This means that all predictions in the aptitude class are completely correct with an accuracy rate of 100%, while the interest class has an accuracy rate of 82%. Recall for the aptitude and interest classes showed that the model was able to identify all positive samples, with 80% and 100% accuracy respectively. Recall reflects the model's ability to recognise and recall actual positive samples. The high recall value indicates that the model was able to identify most of the positive samples, with little error in classifying the positive samples as negative. F1-score gives an idea of the extent to which the model can achieve a balance between precision and recall. In this case, F1-score has high values for all classes, indicating good performance in making positive predictions with high accuracy and recognising all positive samples.

Figure 20. Performance of the SVM Method

**Evaluation**

The assessment is done to verify that the generated model meets the business objectives and produces reliable results. In this evaluation stage, cross-validation techniques are used to measure the performance of the model in making predictions on data that has never been seen before.

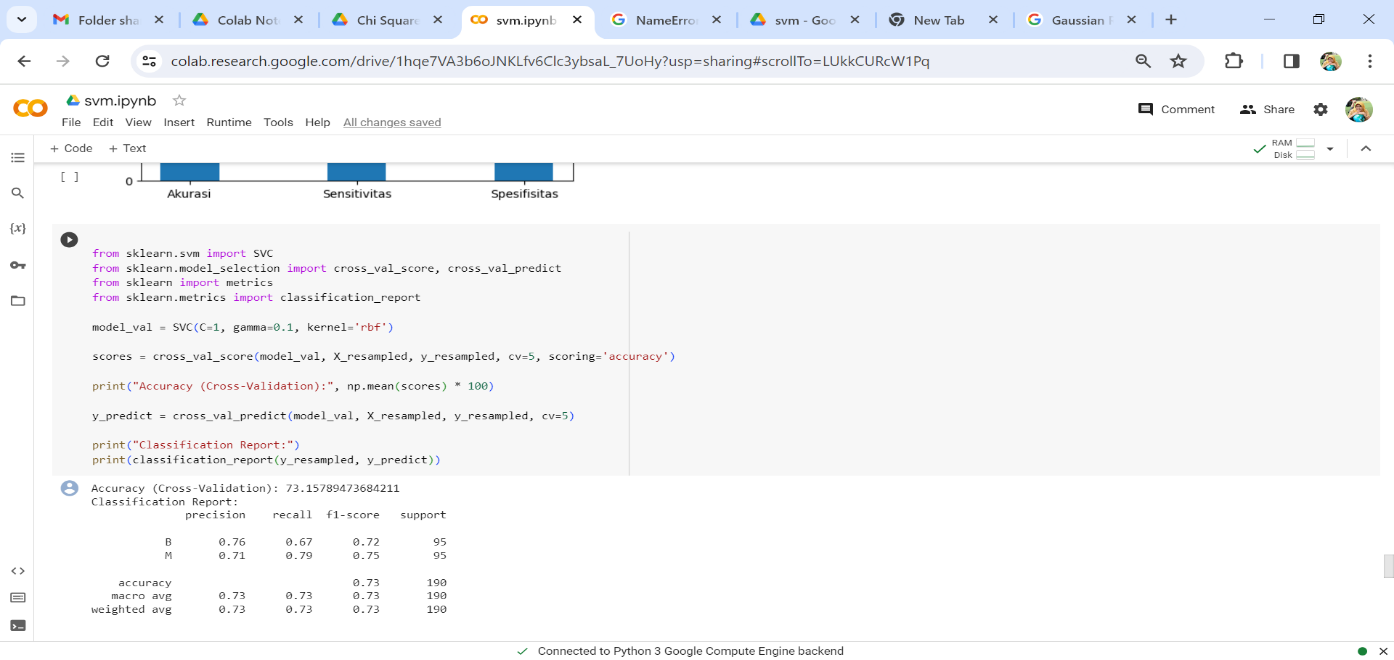
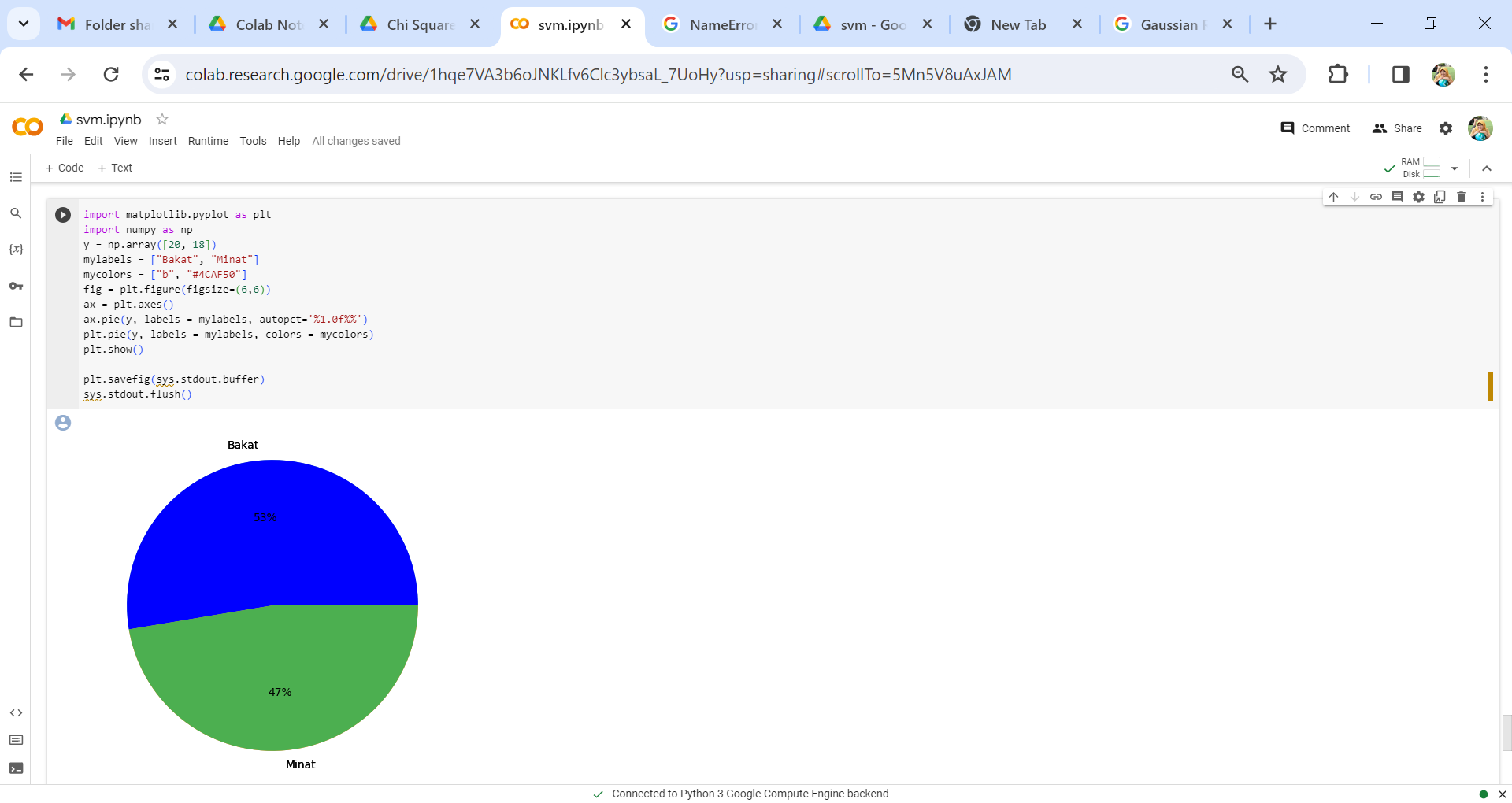


Figure 21. SVM performance

Evaluation of the model for interest and talent categories using cross-validation showed an accuracy rate of 96.66%. This result indicates that SVM is a suitable model and can be applied in the development of applications to support the classification decision of students' interests and talents at SDN XYZ Palembang.

Figure 22. Pie Chart of Interest and Talent Classification

Based on the implementation of Exploratory Data Analysis (EDA) and the application of the Support Vector Machine (SVM) algorithm on 144 questionnaire datasets regarding the interests and talents of students of SDN 204 Palembang, it can be concluded that the SVM algorithm is very suitable for interest and talent classification applications. This is reinforced by the accuracy rate reaching 73%. From the pie chart, it can be observed that out of a total of 144 datasets, aptitude reached 53%, while interest reached 47%. The implication is that schools can more effectively manage student competencies based on interests and aptitudes, supporting the school's goal of producing competent graduates.

**CONCLUSION**

Utilizing Exploratory Data Analysis (EDA) and the Support Vector Machine (SVM) algorithm on 144 questionnaire datasets for SDN XYZ Palembang students, it is concluded that SVM is highly suitable for interest and talent classification with a 73% accuracy rate. From the pie charts, it's evident that among the 144 datasets, aptitude accounts for 53%, while interest is at 47%. This facilitates efficient management of student competence in SDN XYZ Palembang based on interests and talents, aligning with the school's goal of producing competent graduates.

**BIBLIOGRAPHY**

|  |  |
| --- | --- |
| [1] | W. D. S. B. R. Indah Ayu Anggraini, "Analisis Minat dan Bakat Peserta didik terhadap Pembelajaran," *Terampil,* vol. 7, no. 1, pp. 23-28, 2020. |
| [2] | W. Warsito, "Peningkatan Minat Belajar Matematika Kelas Iv Melalui Alat Peraga Layang-layang," *Jurnal Sinetik,* vol. 2, no. 2, pp. 242-248, 2019. |
| [3] | W. d. Falah, "Peran Guru Bimbingan dan Konseling dalam Mengembangkan Minat dan Bakat Program Pilihan Studi Keterampilan Peserta Didik MAN 1 Magelang," *Jurnal Hisbah,* vol. 12, no. 2, pp. 21-28, 2015. |
| [4] | A. S. A. S. R. Ahmad Roihan, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review Paper," *IJCIT (Indonesian Journal on Computer and Information Technology,* vol. 5, no. 1, pp. 75-82, 2020. |
| [5] | S. d. Chavan, "A Review of Machine Learning Techniques Using Decision Tree And Support Vector Machine," 2016. |
| [6] | R. I. B. G. N. &. A. M. A. M. Thupae, "Machine Learning Techniques for Traffic Identification and Classifiacation in SDWSN: A Survey," 2018. |
| [7] | A. G. &. K. A. O. H. Rahman, "Klasifikasi Ujaran Kebencian pada Media Sosial Twitter Menggunakan Support Vector Machine," *Jurnal Rekayasa Sistem dan Teknologi Informasi (RESTI),* vol. 5, no. 1, p. 17 – 23, 2021. |
| [8] | Muhammad Radhi, dkk, "Notebook, Analisis Big Data Dengan Metode Exploratory Data Analysis (Eda) Dan Metode Visualisasi Menggunakan Jupyter," *Jusikom Prima (Jurnal Sistem Informasi dan Ilmu Komputer Prima),* vol. 4, no. 2, pp. 23-27, 2021. |
| [9] | M. Romzi and B. Kurniawan, "Pembelajaran Pemrograman Python Dengan Pendekatan Logika Algoritma," *JTIM: Jurnal Teknik Informatika Mahakarya,* pp. 37-44, 2020. |
| [10] | A. D. S. School, "Library Python Populer," p. 1, 2 February 2022. |