

# Prediction of Company Employee Resignation Using Naïve Bayes Algorithm

Widiyawati<sup>1</sup>, Linda Marlinda<sup>2</sup>

<sup>1</sup>Manajemen Informatika, Universitas Banisaleh, Bekasi, West Java, Indonesia

<sup>2</sup>Informatika, Universitas Nusa Mandiri, Jakarta, Indonesia

Email address: widiyawati@ubs.ac.id, linda.ldm@nusamandiri.ac.id

**Abstract**— The human resources management division supports company growth through recruiting quality employees. Despite stringent selection efforts, some employees still resign before their contracts expire, potentially harming the business. Therefore, this research uses the <https://www.kaggle.com/colara/hr-analytics> dataset to predict potential resignations, help human resource development and managers understand the causes of resignations, and take preventive action. Human Resource Development is a division that has responsibilities and duties in employee management in the company. The details of HRD's duties are to carry out recruitment, conduct training, determine salaries, and provide compensation. Dataset regarding Human Resources to predict Employee Loyalty, with the division of 10 attributes, namely: Satisfaction\_level, Last\_evaluation, Number\_project, Average\_monthly\_hours, Time\_spend\_company, Work\_accident, Promotion\_last\_5\_years, Job, Salary. This research uses fictitious company employee data, with research stages including data collection, pre-processing, and processing using research algorithms. Data analysis was carried out to look for correlations between attributes to individually understand the reasons why employees resign. The method used in this research is the naïve Bayes method, with accuracy results of 87% and an error rate of 13%.

**Keywords**— Employee Performance, Discipline Improvement, Naïve Bayes, Knime

## I. INTRODUCTION

Technological advances have an impact on various fields, especially human resource management. This transformation increases efficiency and results in HRD management. To remain relevant, the field needs to adapt to technology, optimize its use, and utilize analytical tools to solve problems. The decision to use data mining or big data technology can make a significant contribution to improving HRD management performance. HRD data analysis can include descriptive, predictive, and correlational approaches, enhanced by machine learning techniques for decision-making[1].

The large number of employee resignations in a company can cause business failure. Therefore, it is important for companies to know the reasons why employees leave the company. In this research, the company was able to predict employee resignation. It is important for HRD and managers to understand the factors that generally cause employee resignation and make improvement efforts based on these reasons[2]. By knowing the factors that cause employees to leave, companies can prevent employee turnover, so that companies can minimize spending money and also have no difficulty in finding employees. New employees need to be retrained from scratch.

Recruitment often invests heavily in training programs for prospective workers. However, when employees leave after six months, training costs become one of the company's most expensive expenses. High turnover rates are detrimental to employee retention, signifying a significant increase in workload and lowering motivation and self-confidence[3]. Personnel turnover impacts revenue by increasing costs, losing knowledge, and decreasing output, all of which impact profits. High turnover rates also result in higher recruitment, training, and employment costs, while finding competent replacements is not an easy task. Wise workforce reductions can result in

continued production, reduced recruiting costs, steady customer engagement, and increased morale for remaining employees. However, workforce reductions can also be a major challenge, especially when high-skilled and essential employees are lost. With advances in technology, the use of machine learning and predictive models can help in predicting the downsizing of human resources.

Many researchers have proven, the usefulness of human resource management (HRD) in work, production, and management scenarios, and in identifying relationships with productivity. In fact, the results confirm that the impact of HRM on productivity has a positive effect on capital growth and business intensity [4]. Most studies focus on analyzing and monitoring customers and their behavior and do not address the company's main assets, as represented by its employees. Many studies analyze employee attrition. Existing research [5] shows that employee demographics and work-related attributes are the factors that most influence employee attrition, as are salary and duration of the employment relationship. Another study evaluated the impact of demographic attributes and employee absenteeism on attrition[6]. The authors in only focused on job-specific factors. Authors in compared the Naïve Bayes classifier and the J48 decision tree algorithm in predicting the likelihood of an employee leaving the company. Specifically, two methodologies were evaluated for each algorithm: ten-fold cross-validation and percentage split researchers compared the Naïve Bayes classifier and the J48 decision tree algorithm in predicting the likelihood of employee departure [7]. Specifically, two evaluation methodologies were used for each algorithm: tenfold cross-validation and 70% percentage split. The results show an accuracy of 82.4% and incorrect classification of 17.6% with J48 using ten-fold cross-validation while achieving an accuracy of 82.7% and incorrect classification of 17.3% with a percentage split of 70%. In contrast, the Naïve Bayes classifier achieved an accuracy of

78.8% with and incorrect classification of 21.2% using tenfold cross-validation, and an accuracy of 81% with the incorrect classification of 19% using a percentage split of 70%. Furthermore, researchers in explored the application of Logistic Regression to predict employee turnover, resulting in an accuracy of 85% and a false negative rate of 14% [2][8][6].

To collect employee performance data, researchers used the Naïve Bayes algorithm. This algorithm identifies 10 attributes as parameters Level of Satisfaction, last ranking, number of projects, average monthly hours, length of service at the company, work accidents, departures, promotions in the last 5 years, sales, and salary. Using a large number of parameters will increase the accuracy of the results, whereas the Naïve Bayes algorithm is a form of structured decision-making that collects and creates decision rules.

## II. METHODOLOGY

In this study, an experiment was conducted using data from corporate employees. Research steps include data collection, preprocessing, processing through search algorithms, and data analysis to examine correlations between attributes to identify the most influential attributes in the study[9]. The following study diagram is shown in Figure 1.

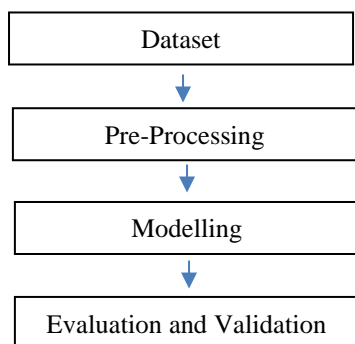


Fig. 1. Research flow

### A. Data Collection

In the initial stage, data was collected via the Kaggle site. The dataset used is a secondary data type Labor dataset obtained from the online repository <https://www.kaggle.com/colara/hr-analytics>, this data is similar to data in the human resources department of companies.

### B. Pre-Processing

Before the data is used, pre-processing is carried out to remove noise in the dataset and also add features. In the data used, there is noise such as duplicate data, inappropriate data types, and missing values. So, it is necessary to carry out pre-processing by deleting duplicate data, changing the data type, deleting data with empty values, and adding 2 features as well as normalizing the data to balance the values for each record, and dividing the data into training and testing data.

### C. Modeling

This step is a very crucial stage in conducting research related to machine learning or big data. At this stage, in-depth data exploration is carried out to identify potential outliers and

correlations between certain variables, which are then visualized. Next, at the modeling stage, using the Naïve Bayes method, various data modeling techniques are selected and implemented with parameters that have been calibrated according to the specified configuration. This step aims to achieve modeling results with an optimal and accurate level of accuracy[10]. The application of these modeling techniques requires the fulfillment of special requirements related to the condition of the data used, which must be complied with by each modeling technique applied.

### D. Evaluation and Validation

In this evaluation phase, the process is carried out using the KNIME tool, where data validation against training data and cross-validation are carried out. Next, data separation was implemented, and a comparative analysis of machine learning in the prediction classification of workers who resigned using the Naïve Bayes method. The dataset used has been cleaned of noise that could interfere with the construction of the classification model, so it does not require manipulation or changes. This dataset is used to build a prediction model, and the results are compared to produce the best model[11][1][12].

The results of supply training are obtained by classifying the dataset which is divided into two parts, namely training data and test dataset. The Supply Training scenario begins by determining the proportion of training data and test data, starting from 50% training data and 50% test data. This proportion is then increased by 10% for training data and reduced by 10% for test data until it reaches a division of 100% for training data. If the training data uses 100%, then the test data also uses 100%.

In the cross-validation scenario, the data is divided into several small sets which alternate as training data and test data. This scenario is run using k-fold, with k values of 3, 5, 7, and 9. Each subsection is used as training data and test data alternately. The process of changing the use of training data and test data continues until all parts have been used once as test data. Cross-validation scenario execution is carried out to eliminate algorithm results with k values that have the lowest average accuracy, thereby producing information about the best model using the Naïve Bayes method[11].

The performance value is obtained by comparing resources before and after classification. To obtain general information in cross-validation testing scenarios, the relevance and performance results will be averaged.

### E. Naïve bayes (NB)

Classification can be considered a form of machine learning because it has the ability to use pre-existing knowledge to make decisions regarding new objects. Basically, classification measures the system's ability to assign labels to objects based on previously identified patterns without the need to change the system when faced with a new object[6][7].

Naïve Bayes is a simple probabilistic classification method, where it calculates a series of probabilities by combining frequencies and combinations of values from a given dataset. This algorithm uses Bayes' theorem and assumes that all attributes are independent or not interdependent, which is given

by the value of the class variable[10][8][9]. In simple terms, the Naïve Bayes formula can be expressed as a conditional probability, as follows:

$$Posterior = \frac{Prior \times likelihood}{evidence}$$

To describe the Naive Bayes method, it is important to understand that in the classification process, a number of clues are needed to determine the most appropriate class for the sample being analyzed. Therefore, the formulation of the Naive Bayes method can be described as follows:

$$P\left(\frac{C}{E}\right) = \frac{P\left(\frac{E}{C}\right) \times P(C)}{P(E)}$$

- E : Evidence or evidence of existing data
- C : Assumption of objects with specific classes
- P(E|C) : Probability E based on C (Likelihood)
- P(C|E) : Probability C based on condition E (Posterior)
- P(E) : Probability E unknown C(Evidence)
- P(C) : The probability of the class object is true(Prior)

To explain the Naive Bayes theorem, it is necessary to know that the classification process requires several pieces of evidence to determine which class is suitable for the sample being analyzed. Therefore, the Bayes theorem above is adjusted for E as a set of evidence as follows:

$$P(C|F_1 \dots F_n) = \frac{P(C) \times P(F_1 \dots F_n|C)}{P(F_1 \dots F_n)}$$

The variable C represents the class, while the variables F1 ... Fn represent the characteristics of the evidence required to carry out the classification. The probability value of the attribute F from the equation above is obtained using the Laplacian smoothing probability function, where K=1 is used to avoid probabilities that are zero[2][11][7]. The Laplacian smoothing probability equation can be found in the following formulation:

$$P(F) = \frac{Count + K}{N + (K \times Z)}$$

- P : Probability of the variable F
- Count : Total emergence F
- K : Parameter smoothing
- N : Amount of data
- Z : Number of class types from the sample

In applying the Naïve Bayes method, before knowing the final result, it is necessary to calculate the likelihood value using the following formula:

$$P(E|C) = P(F_1|C) \times P(F_2|C) \times \dots \times P(F_n|C)$$

- P : Probability
- E : Evidence Value
- C : Object class
- Fi : Attribute of E

The results of the equation above produce simple values that can be used to determine new object classes[1][5]. The normalization function of this equation can be seen in the following formula:

$$P(X) = \frac{Likelihood \ prior}{Likelihood \ prior + Likelihood \ Posterior}$$

- P : Probability
- X : New Object

Likelihood prior : Previous possibilities

Likelihood posterior : Next possibility

By comparing the probability values of P(C), objects X can be classified into specific classes

### III. RESULTS AND DISCUSSION

The machine learning comparison system found will be divided into two aspects, namely the results of interface implementation and the results of algorithm implementation. The interface implementation section discusses the appearance of the system to maximize the information presented to the user. Meanwhile, the algorithm implementation section will detail the classification process. The system performance flow generally starts by receiving large amounts of workforce data input which is called a dataset. This dataset is divided into training data and test data, with training data used for training and model formation, while test data is used during the classification process by the model.

The Naïve Bayes classification process on test data will produce relevance and resource usage output. The classification results are then compared and compiled to form comparison output information.

#### A. Data Collection

In this research, the author succeeded in building a system for comparing classification methods. The dataset used is 14,999, with 10 attributes, 2 classes, 2 binary attributes, 2 continuous attributes, 2 category attributes, and 2 numeric attributes. The dataset is clean from noisy data anomalies, nulls, or mud. The dataset used is a secondary data type Labor dataset obtained from the online repository <https://www.kaggle.com/colara/hr-analytics>. The workforce dataset consists of two target classes, namely 1 for quitting and 0 for remaining, with an active work period of less than 10 years. The dataset has 10 features, namely satisfaction level, last evaluation, project number, average monthly hours, time spent company, work accident, promotion last 5 years, division, and salary.

#### B. Pra Proses Data

Pre-processing data in learning to process the possibility of employees resigning using the Naïve Bayes method in Knode machine learning tools.

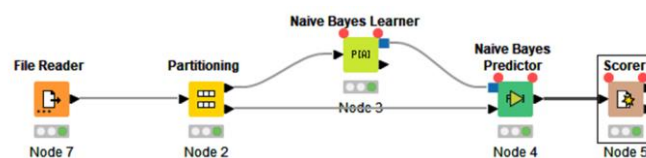


Fig. 2. Naïve bayes workflow to test employee likelihood to resign

In the picture above, after the dataset is used in Excel reading to pre-process any missing or invalid data or values carried over during processing.

#### C. Data Analysis

Process dataset reader

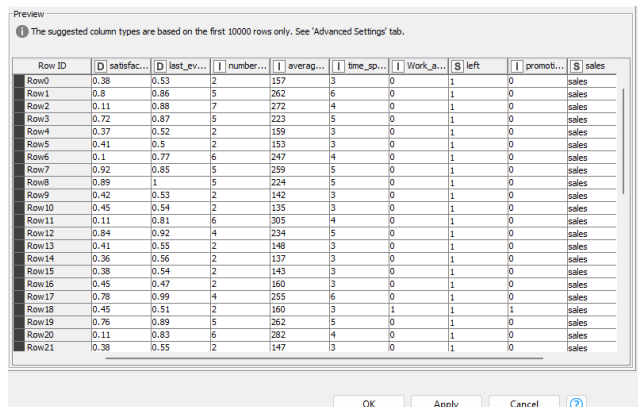


Fig. 3. File Reader Configuration

The Supply Training scenario begins by determining the proportion between training data and test data, with tiered divisions ranging from 50%-50%, 60%-40%, 70%-30%, 80%-20%, and 90%-10%, to achieve full distribution of 100%-100%. Data collection at each division stage is not carried out randomly but with the aim of giving each method a fair opportunity to build a prediction model. The distribution of data is calculated through the average equation with the aim of obtaining general information.

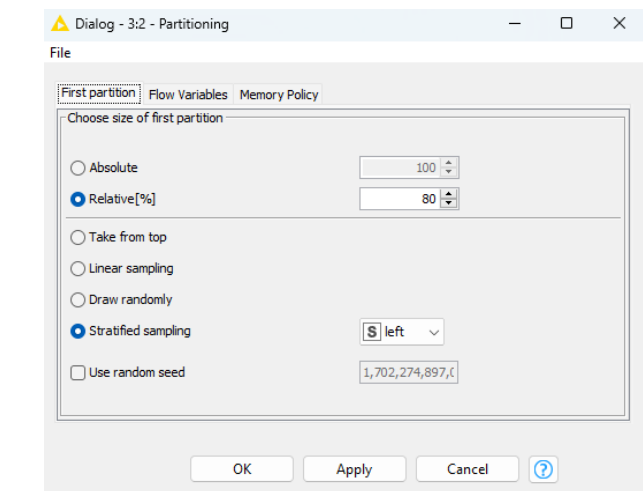


Fig. 4. Data partitioning and selection process

At this stage the data partition process is carried out, where in the picture above 80% of the data is testing data the remaining 20% is training data and the sampling data taken is the Left column (Leave/Resign in this scenario)

D. Modelling

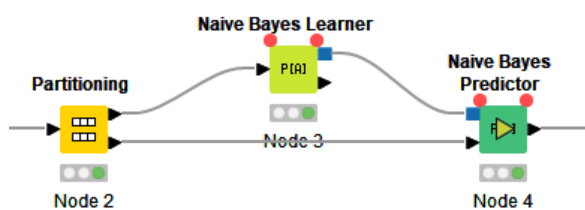


Fig. 5. Linkage between nodes

The next stage is connecting nodes between partitioning to Naïve Bayes Learner and Naïve Bayes Predictor for the processing stage using the Naïve Bayes algorithm.

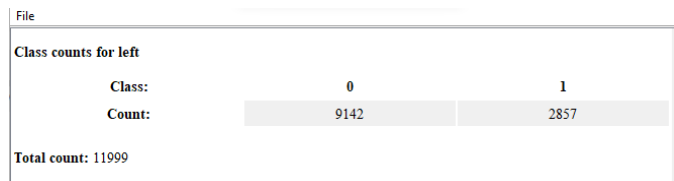


Fig. 6. Output Naive bayes on Knime

In the picture above, from a total of 11999 data, there are 9142 not Left (marked with the number "0") and there are 2857 Left (marked with the number "1")

In the CrossValidation scenario, the dataset will be divided into k parts, each of which will be used as test data in turn, while the rest will be used as training data. The cross-validation scenario provides information when the model receives the entire dataset as training data and test data. The results of each model with the best "k" parameter values will be used in the analysis. Next, the average value is calculated with the aim of finding information on the best prediction model in general, without any bias. The results of relevance, memory usage, and time usage are combined and used as a reference in the comparison process.

P(salary   class=?)			
Class/salary	high	low	medium
0	937	4075	4130
1	67	1723	1067
Rate:	8%	48%	43%

Fig 7. Output Naives bayes on Salary

In the picture above, Knime categorizes the amount of output into several categories, including 937 employees with high income not left, 4075 employees with low income not left, and 4130 employees with medium income not left. Then there were 67 employees with high salaries who chose left, 1723 employees with low incomes chose left, and 1067 employees with medium incomes chose left. From the conclusion of the data above, 48% or the majority of employees who resigned were low-income, followed by medium-income employees and the last was high income.

P(sales   class=?)										
Class/sales	IT	RandD	accounting	hr	management	marketing	product_mng	sales	support	technic
0	789	525	455	438	431	517	557	2503	1330	1597
1	214	93	160	172	70	164	162	820	430	572
Rate:	8%	5%	5%	5%	4%	6%	6%	28%	15%	18%

Fig. 8. Output Naives Bayes on Sales

The picture above is explained and explains in detail about which employees in sales or divisions resigned. In the picture above, the division with the most remaining or not left is the



sales division with the number 2503 not left, then the division with the most left is technical as many as 572 employees chose left.

#### D. Evaluation and validation

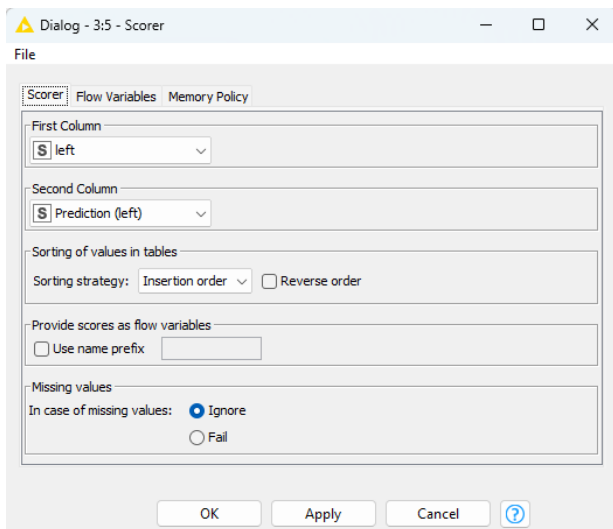


Fig. 9. Node Naïve Bayes Predictode to Scorer

The picture above is a display of the scorer configuration where the node is used to predict results using the true positive false positive and true negative false negative methods to predict the possibility that there will be employees who want to choose left

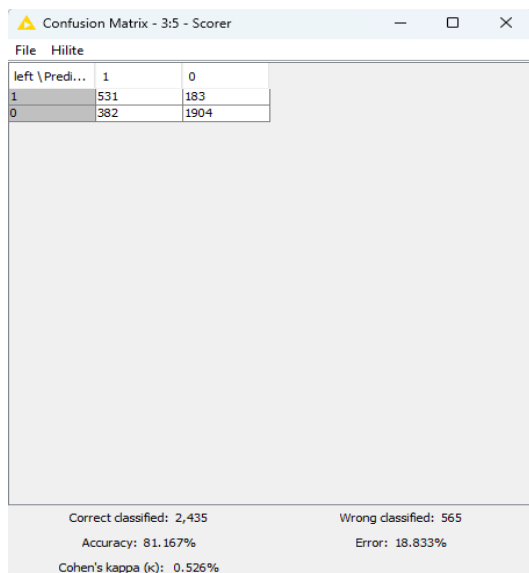


Fig. 10. Output Scorer estimated probability left

The image above is the output result of the naïve Bayes predictor, where 20% of the training data was previously selected for the training data, namely 2,435 data, and the accuracy level of this estimate is 81,167% with a possible error of 18,833. where these results provide information that a total of 531 employees are left and right left, 183 employees are left

but not left, then there are 382 employees who are not left but left and 1904 employees are not left.

#### IV. CONCLUSION

The naïve Bayes method can be used for classification models from a set of input data, especially predicting whether an employee will leave a company. Naïve Bayes uses a learning algorithm to obtain a classification model. So the input data becomes an important influence in forming the decision tree and the resulting rules. The resulting level of accuracy. In the picture above, it is explained and explains in detail regarding which employees in sales or divisions resigned. In the picture above, the division with the most remaining or not left is in the sales division with the number 2503 not left, then for the division with the most left is technical as many as 572 employees chose left. Using KNIME can provide a high level of sharpness from training data and test data, the accuracy of the Yes label is 81.1% while the No is 18.9%. Meanwhile, for the Get Score process on test data, the error result is 0.5%.

#### REFERENCES

- [1] P. P. Gulabbhai, M. Gangil, and M. Tech, "Employees Skills Inventory using Deep Learning for Human Resource Management," / *Res. J. Eng. Technol. Manag.*, vol. 02, no. 04, pp. 2582–2610, 2019, [Online]. Available: [www.rjetm.in/](http://www.rjetm.in/).
- [2] L. Sommer, "How Artificial Intelligence can be used in International Human Resources Management: A Case Study," *GATR Glob. J. Bus. Soc. Sci. Rev.*, vol. 11, no. 1, pp. 09–17, 2023, doi: 10.35609/gjbsr.2023.11.1(2).
- [3] W. Paper and W. Papers, "management practices for a firm ' s innovation," 2013.
- [4] R. Eduvie, J. C. Nwaukwa, F. Uloko, and E. Taniform, "Predicting Employee Attrition Using Decision Tree Algorithm," vol. 9, no. 9, pp. 1305–1318, 2021, [Online]. Available: [www.globalscientificjournal.com](http://www.globalscientificjournal.com).
- [5] B. Sri Harsha, A. Jithendra Varaprasad, and L. V. N. Pavan Sai Sujith, "Early prediction of employee attrition," *Int. J. Sci. Technol. Res.*, vol. 9, no. 3, pp. 3374–3379, 2020.
- [6] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. De Luca, "Predicting employee attrition using machine learning techniques," *Computers*, vol. 9, no. 4, pp. 1–17, 2020, doi: 10.3390/computers9040086.
- [7] R. Chauhan, "Prediction of Employee Turnover based on Machine Learning Models," *Math. Stat. Eng. Appl.*, vol. 70, no. 2, pp. 1767–1775, 2021, doi: 10.17762/msea.v70i2.2469.
- [8] A. Mansurali, M. Rajagopal, and R. Subbaiah, "Employee Attrition and Absenteeism Analysis Using Machine Learning Methods," vol. 11, no. 8, pp. 155–167, 2023, doi: 10.4018/978-1-6684-8942-0.ch011.
- [9] R. Jayadi, H. M. Firmantyo, M. T. J. Dzaka, M. F. Suaidy, and A. M. Putra, "Employee performance prediction using naïve bayes," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 6, pp. 3031–3035, 2019, doi: 10.30534/ijatcse/2019/59862019.
- [10] R. V. Raja, A. D. Kumar, I. Thamarai, S. N. Mohammed, and R. R. Kanna, "Analytic Approach of Predicting Employee Attrition Using Data Science Techniques," *J. Theor. Appl. Inf. Technol.*, vol. 101, no. 9, pp. 3380–3391, 2023.
- [11] V. Kakulapati and S. Subhani, "Predictive Analytics of Employee Attrition using K-Fold Methodologies," *Int. J. Math. Sci. Comput.*, vol. 9, no. 1, pp. 23–36, 2023, doi: 10.5815/ijmsc.2023.01.03.
- [12] E. P. F. Lee *et al.*, "An ab initio study of RbO, CsO and FrO (X2Σ<sup>+</sup>; A2[ $\Pi$ ]) and their cations (X3Σ<sup>-</sup>; A3[ $\Pi$ ])," *Phys. Chem. Chem. Phys.*, vol. 3, no. 22, pp. 4863–4869, 2001, doi: 10.1039/b104835j.