International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Feature Extraction Techniques to Analyze Competitive Exam Result Data to Improvise Educational Policies

Richa[1], Jyoti Singh[2]

[1]richa12.nitrr@gmail.com, *Hewlett Packard Enterprise, STSD, Bangalore, Karnataka, 560016, India*
[2]jsraipur13@gmail.com, *Chhattisgarh Professional Examination Board, Raipur, Chhattisgarh, 492002, India*

---***---

Abstract - The increase in e-learning resources and online facilities for applications, examinations and evaluations has generated huge amount of data in educational field. As quality of education plays vital role in development of individuals, this work aims at analyzing the quantity as well as quality of education district wise, in a state, by utilizing a competitive exam result database. In some district, the results may be better quantitatively, while in others, number of students cleared may be less, but scores may be higher. The aim is to understand the strength and weakness of existing educational infrastructure and how to bring up its level in rural districts, thus helping policymakers to reform and improvise. The collected database undergoes pre-processing phase to ensure its quality and feature engineering is done for missing values, data validation, inconsistencies, data dependencies and outliers. Data inconsistency is handled to ensure data integrity and reliability. Outliers are detected and removed to avoid skewing the predictions on examination results. Machine learning and deep learning models will be applied to Exam-result database to analyse and extract valuable insights. This knowledge can aid in optimizing resource allocation, improving policies to enhance overall educational outcomes and implementing targeted interventions. The accurate predictions generated from this work can help to identify applicants' risk of underperforming and can provide personalized support

Key Words: Feature extraction, Information retrieval, Classification, Data pre-processing, Data analysis, Education assessment, Correlation analysis, Skewness, Kurtosis, KNN algorithm, Gaussian Naïve Bayes algorithm, Bernoulli Naïve Bayes algorithm

1. INTRODUCTION

Competitive exams are a good tool for evaluating knowledge, skills and ability of a candidate. They provide a transparent way to select candidates on the basis of their merit rather than favoritism. Merit based selections assures quality in place of quantity, providing equal opportunities to all candidates regardless of their gender, background or social status and enable individuals from diverse background to excel. Success in these exams requires hard work, dedication and discipline. Achieving good scores in exams lead to personal growth and development and bring recognition and prestige to individuals. A state has 30 administrative districts. Only 24% is urban and remaining 76% area is rural in state. Government agencies conduct three types of competitive exams: entrance exam, recruitment exam and eligibility test. Competitive exams offer efficient and transparent way to sort large number of candidates, helping organizations and institutions to streamline the selection process and identify the most suitable candidate.

Agencies conducting exams collect different data including, personal details, educational background, identification documents, payment information, choice of exam center, date of registration etc., using online application forms submitted by candidates. On the basis of application data, agencies issue digital Admit cards to candidates and allot exam centers to them. After examination, scores and result data is generated. Candidates get rank or percentile and result status.

Performance analysis of candidates on the basis of demographics and regions is possible. Data mining and machine learning techniques can help find trends, patterns, outliers and factors affecting performance. Predictive models can estimate the likelihood of candidate's success on various parameters as age, gender, category, caste and living area. Ranking analysis can provide insights into competitiveness of the individual.

Data mining can generate performance metrics as average marks and standard deviations which can be used by policymakers and institutes to upgrade education facilities in rural districts. It can also uncover candidate's behavior of not appearing in exam. Feature extraction techniques enable the transformation of raw data into a more manageable and insightful representation. Methods like Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor embedding (t-SNE) can reduce the dataset's dimensionality by preserving essential information. Pattern Recognition can help identifying candidates' strengths and weaknesses, aiding in personalized learning and targeted interventions within exam result data.

Here recruitment exam data is collected and analyzed. Objective is to upgrade rural candidates and increase facilities and awareness. The results may facilitate policymaker and institutions to decide which area or districts are too backward and weak in which education.

2. Data collection and data description

Data description and data collection are important steps in the process of conducting research in machine learning, data science and data analytics. Here data is collected from a competitive exam conducting organization's website. This organization conducts state level examinations and displays merit list on their website. Result table is of recruitment exam for filling post of lecturers in Mathematics department. Data is in structured form. In this project real data of candidates' results from exam conducting organization is used for analysis.

Location where the candidates live is termed as domicile district and staying has a direct impact for their examination results. There are 30 districts and applicants are divided among these 30 districts. Districts are either urban or rural. This state has 76% of rural area and only 24% of urban area. Thus most of the districts come under rural area. Education facilities and quality is much better in urban districts. Therefore number of applicants from rural districts are lesser in number and their scores are also less as compared to urban candidates.

2.1 Categorical and Numerical data

Shape of the considered result dataset is (12484, 40).

Of Categorial Data in the Maths dataset
['excd', 'exnm', 'appno', 'regno', 'cennm', 'cname', 'relative', 'rel_name', 'mname', 'dob', 'gender', 'gender_det', 'category', 'ph', 'phtype', 'ex', 'clas_type', 'domicile', 'dist_ep', 'dist_dom', 'pvtgtype', 'cadc_appl', 'cadc_sub', 'cadc_pgper', 'cadc_qual', 'cadc_exp', 'cadc_expyr', 'status', 'mrkfnl']

Of Numerical Data in the Maths dataset
['dst', 'rollno', 'cencd', 'cat', 'krishak', 'cadc_ugper', 'cadc_12per', 'mrk', 'experien', 'bonus', 'rank']

2.2 Class imbalances on the dataset:

2.2.1 Class imbalance based on marks:

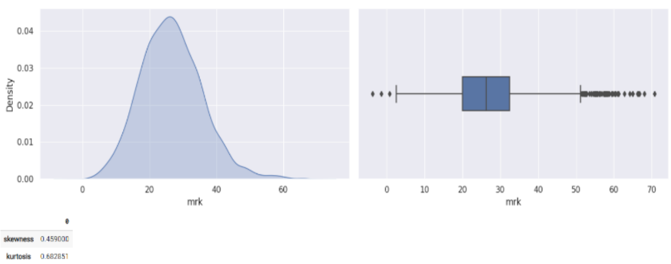

Fig-1: Class imbalance based on marks

As seen in Fig-1, skewness = 0.459, hence "marks" data is not skewed.

2.2.2 Class imbalance based on gender:

The distribution of males/females is 52% and 48% respectively, as seen in Fig-2.

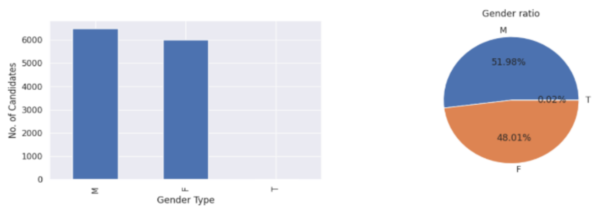

Fig-2. Class imbalance based on gender

2.2.3 Class imbalance based on category of candidates:

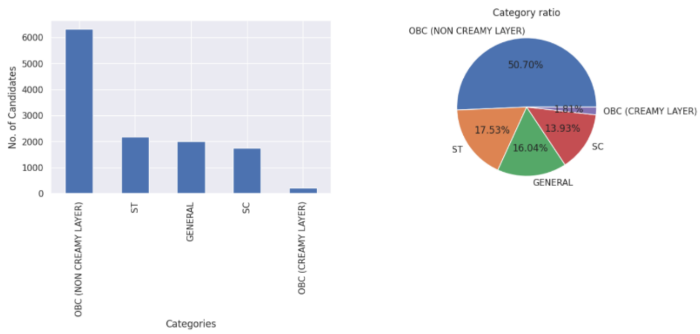

Fig-3. Class imbalance based in category of candidates

There is a lot of imbalance among the various categories. Hence preprocessing will be required here.

3. Data visualization

3.1 Relation between gender and category

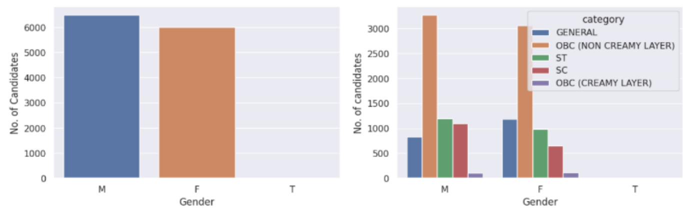

Fig-4. Relation between gender and category

3.2 Number of absent candidates from each category

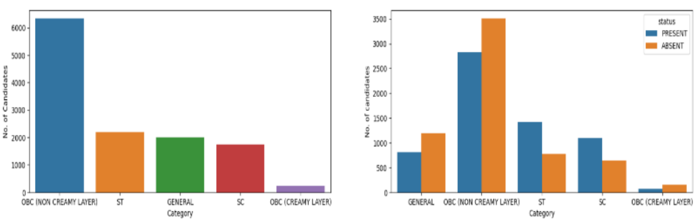

Fig-5. Absent candidates from each category

3.3 Numerical data analysis for district from which candidate appeared, UG and 12th percentages, marks, experience, bonus and category

	dst	cadc_ugper	cadc_12per	mrk	experien	bonus	cat
skewness	0.548867	61.973115	0.103143	0.459000	15.057423	15.057423	-0.524705
kurtosis	-0.424791	4065.090130	-0.561368	0.682851	235.437918	235.437918	-0.482073

Table-1. Skewness and Kurtosis of numerical data

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

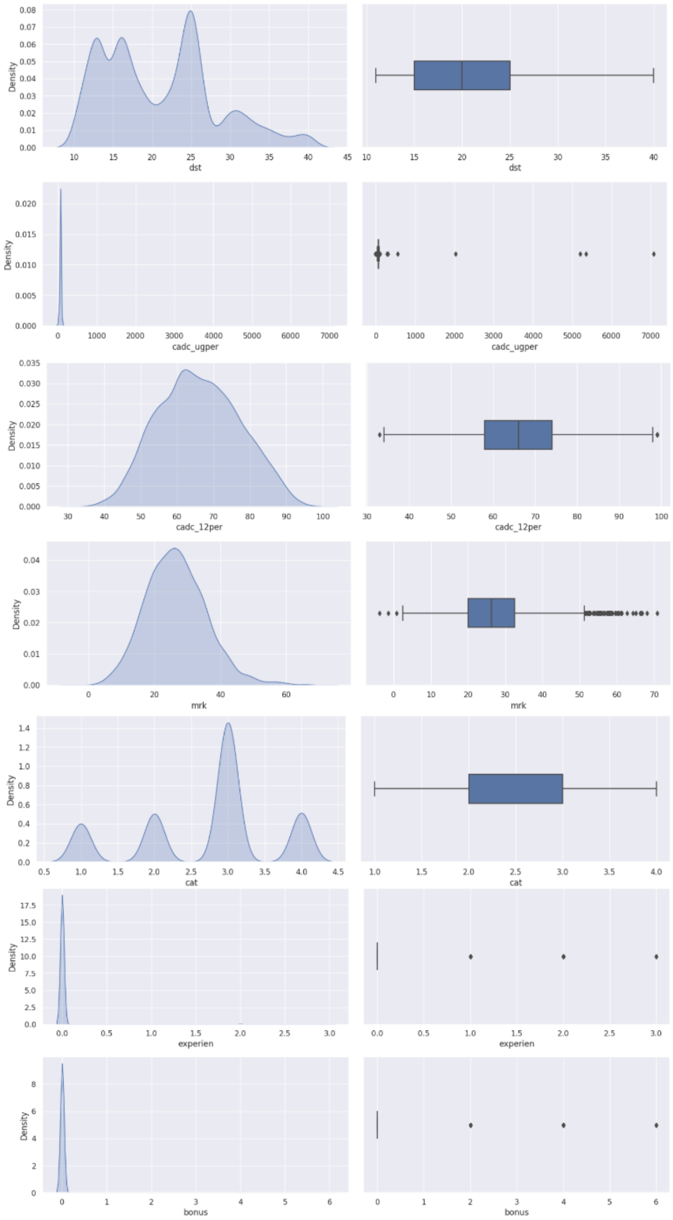

Fig-6. Numerical data analysis

3.4 Correlation matrix

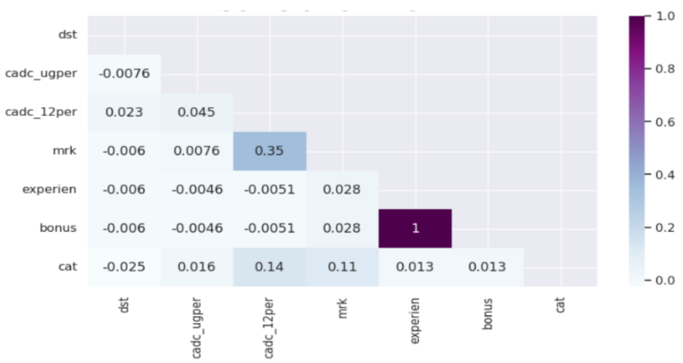

Fig-7. Correlation matrix for numerical data

From the correlation matrix, all the predictors more or less seems to be independent. Some have a correlation around 0.35 but those are acceptable margins. Features with moderate to high correlation are more linearly dependent and have almost the same effect on the dependent variable. So, when two features have high correlation we can drop one of the two features. Further, there isn't any strong relationship between the target (marks) and any of the predictors as the correlation coefficient is very close to 0.

4. Data pre-processing and cleaning

Data table contains many missing values, errors, outliers and inconsistencies. Cleaning and preprocessing is done before modeling.
Duplicate records doesn't provide any significant information for data mining algorithm, hence, will be removed from the dataset; retaining one instance of the duplicate record for modelling.

In data analytics, outliers are data objects that are significantly different from other objects. Outlier data needs to be edited, worked on, or manipulated before suitable analysis, as they can cause anomalies in the results obtained. They require special attention and need to be removed in order to analyze data effectively.

Outliers are categorized as Global outliers, Contextual Outliers and collective outliers.
Manually entering data, adding dummy data for test detection, extracting or mixing data for sampling, data processing errors, measurement errors, experimental errors are some causes of outliers.

4.1 Handling null, missing, duplicate and ambiguous records in dataset

The dataset has 31218 NA, null or missing values.

Since this is a result database, the roll number will be unique. But there are ambiguous cases where, a candidate applied for exam more than once from more than one district. This generated duplicate records. But even if a candidate had more than one record, he appeared from one location only. Thus, his status was marked present only at one place and rest were absent.

There were 113 such duplicate records found when compared on the basis of candidate name, date of birth and status (present/absent).

```
Record count before dropping duplicates:  (12484, 40)
Record count after dropping duplicates:  (12263, 40)
```

4.2 Removing absent candidates from the dataset

There are 6037 absentees in the exam.

```
Record count before dropping absentees:  (12263, 40)
Rows having absent candidates: [array([ 6224,  6225,  6226, ..., 12258, 12259, 12260])]
Record count after dropping absentees:  (6226, 40)
```

4.3 Removing outliers from the dataset

Here outlier records are the ones with negative marks and blank domicile districts.

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 12 | December - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

```
Record count before dropping candidates with -ve marks:  (6226, 40)
Rows having candidates with -ve marks: [array([6224, 6225])]
Record count after dropping candidates with -ve marks:  (6224, 40)
```

```
Record count before dropping candidates having blank dist_dom:  (6224, 40)
Rows having candidates having blank dist_dom: [array([], dtype=int64)]
Record count after dropping candidates having blank dist_dom:  (6224, 40)
```

4.4 Dropping columns having distinct values in the dataset

Columns with one distinct values doesn't provide any significant information while doing modelling. If any column contains one distinct value, then it can be dropped.

Following columns will be dropped:

```
'cname','rel_name','dob','cat','dst','status','excd'
,'exnm','appno','regno','rollno','cennm','cencd','re
lative','mname','dist_ep','gender_det','phtype','ex'
,'krishak','clas_type','domicile','pvtgtype','cadc_p
gper','cadc_appl','cadc_sub','cadc_qual','cadc_exp',
'cadc_expyr','mrkfnl','rank'
```

```
Data size after dropping columns from Maths dataset:  (6224, 9)
```

After the preprocessing and cleanup is done, the datasets were again checked for imbalances. There wasn't any significant differences observed. Feature engineering needs to be done in order to further apply modelling.

5. Feature Engineering

Feature Engineering is the act or process of selecting, manipulating and transforming data into features that can be used in supervised learning. In order to make models work well, it is necessary to design and train better features. Feature transformation aim is to plot and visualize data, speed up training and increase accuracy of the model.

5.1 Transforming categorical columns

The categorical features in the dataset cannot be used for directly for modelling and applying machine learning techniques. Hence, these features are to be converted to numerical columns.

Post preprocessing and cleanup, there are only four categorical features left – domicile district, category, gender and physically challenged.

There are 30 districts in the state being considered, which are classified as developed and under developed. Also, a lot of imbalance was observed for this feature. Since the main focus is on the relation between district and marks, the district is being converted to binary data based on the above classification.

For the remaining categorical features, ordinal encoding is applied, to transform them into numerical data before applying the ML techniques.

The new dataset looks like:

index	gender	category	ph	cadc_ugper	cadc_12per	mrk	experien	bonus	dist
0	0	0	1	72	75	70.75	0	0	1
1	1	2	1	79	88	68.0	0	0	0
2	0	2	1	80	90	66.75	0	0	0

Table-2. Converted categorical data to numerical

5.2 Handling outliers and skewness for the numerical features

As observed, "cadc_ugper", "experien", and "bonus" are heavily skewed. Performing following steps to handle it:

- **Outlier Capping:** We applied capping based on quartile value of 0.05 as lower limit and 0.95 as upper limit.
- Since most of the values in "experien" and "bonus" are 0, we can convert it as a categorical column containing 0 and 1.
- **Log Transformation:** Finally we applied log transformation to reduce the skewness and handle the outliers.

Skewness before processing:

	cadc_ugper	cadc_12per	mrk	experien	bonus
skewness	46.600599	0.104383	0.468128	12.270514	12.270514
kurtosis	2230.600598	-0.557250	0.671523	154.963215	154.963215

Table-3. Skewness and Kurtosis before processing

Skewness after processing:

	cadc_ugper	cadc_12per	mrk	experien	bonus
skewness	0.198007	0.104383	0.468128	0.0	0.0
kurtosis	-0.990403	-0.557250	0.671523	0.0	0.0

Table-4. Skewness and Kurtosis after processing

6. Implementing Machine Learning Models

Transforming data into train and test dataset:

	Sample Size(%)	Record Count (Training Set)	Feature Count (Training Set)	Record Count (Test Set)	Feature Count (Test Set)
0	0.5	6192	8	32	8
1	5.0	5912	8	312	8
2	10.0	5601	8	623	8
3	15.0	5290	8	934	8
4	20.0	4979	8	1245	8
5	30.0	4356	8	1868	8

Table 5. Record and Feature counts of Training and Test sets

Will be using 70:30 split ratio to test the machine learning models.

6.1 K-Neighbors (KNN)

K-nearest neighbor is a supervised machine language algorithm used to solve classification or regression problems. It is a non-parametric algorithm as it does not make any assumption on data. K-NN algorithm assumes the similarity between the new data and available data and put the new data into the category that is most similar to the available categories.
It is also called as lazy learner as it does not learn from the training set instead it stores dataset and perform actions at the

time of classification. The most preferred value for it is 5 and low values can be noisy and may lead to effects outliers.
KNN is simple to implement on large training data and robust to the noisy training data.

Following functions are implemented:

- **get_KNeighbors()** : Returns K Nearest Neighbors(based on Euclidean distance) for a given test record. In this function, we calculate the distance between test records with each record in the training dataset using the function calc_EuclideanDistance. Once the distance is calculated, we combine the distance with target (class) variable array of input training dataset. We sort the resultant data by distance and returns top K nearest neigbhour's class value.

- **knn_algorithm()** : Calls get_KNeighbors() to get the top K nearest neighbors and predicts the class label for a given test record based on the majority class represented by its identified K nearest neighbor.

- **calc_accuracy()** : Calculates accuracy as percent of correct predictions against actuals.

Slowness and computational intensiveness is one of the disadvantage of using KNN Algorithm. It was clearly evident from the execution time of the implement algorithm. It was difficult to capture data as the program tend to runs for hours even for 95:5 % training and test dataset split.

Hence, to demonstrate the functionality of the implement KNN solution, we have considered 0.3 times the dataset as test dataset.

Performance evaluation for KNN:

1. Comparing the results for implemented and sklearn library (Distance metric is Minkowski with p=2 value) for K=5. Also calculating the accuracy for both methods.
2. Finding the best accuracy for a various K values in a range.
3. Hyper-parameter tuning using K-fold cross validation.

6.1.1 For K=5

➢ Using functions implemented

```
Accuracy of Implemented KNN algorithm :  52.94  %
```

➢ Using sklearn library

```
Accuracy of KNN algorithm from sklearn library :  55.67  %

Accuracy Based On K-Neighbors on Training Dataset :  70.0  %
Accuracy Based On K-Neighbors on Test Dataset :  55.67  %
```

Classification Report:

	precision	recall	f1-score	support
0	0.62	0.68	0.65	1114
1	0.44	0.38	0.41	754
accuracy			0.56	1868
macro avg	0.53	0.53	0.53	1868
weighted avg	0.55	0.56	0.55	1868

Table 6. Classification report for KNN (K=5)

Confusion Matrix:

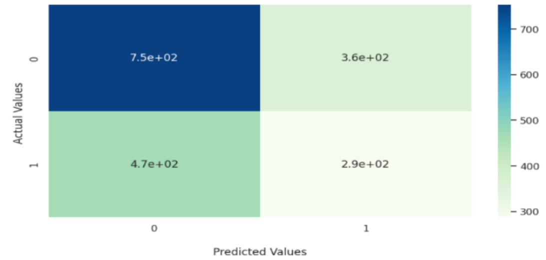

Fig. 8. Confusion matrix for KNN (K=5)

6.1.2 For K in range (20, 70)

➢ Using Grid Search

```
{'n_neighbors': 62} 0.6003225755485101
Accuracy of KNN algorithm from sklearn library :  60.03  %
```

The Grid Search Model was used to figure out the best metrics for use with K Neighbors Algorithm, in the range 20 to 70. As observed, the best accuracy is achieved using n_neighbors: 62, with accuracy 60.03%.

➢ Without using Grid Search

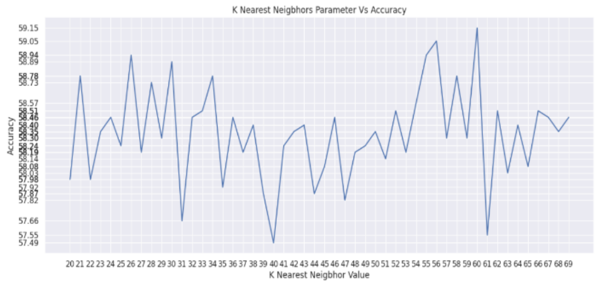

Fig. 9. KNN parameter vs accuracy

Above plot gives accuracy for various values of K, without using Grid Search. But the result is still comparable for same range (20, 70). Here max accuracy of 59.15% is observed for K value of 60.

Due to performance considerations, we ran the above steps in batches and presented result only for the batch size (K=20 to K=70) that represent the accuracy trend appropriately.

6.1.3 Hyper-parameter tuning using K-fold cross validation

Due to performance reasons, the tuning was run in batch size of 10 K-values and observed the accuracies was high in the range of 80 to 90. Hence, presented the required data for this batch size (K=80 to 90). The accuracy was observed to be high at 59.78 % for K Nearest Neighbor value of 68 using cross validation with K-Fold value = 5.

	K Nearest Neighbor	Mean Score (%)
4	84	59.78
5	85	59.62
2	82	59.55
6	86	59.50
3	83	59.44
1	81	59.37
9	89	59.32
7	87	59.25
0	80	59.18
8	88	58.98

Table 7. Mean score for KNN (80, 90)

Accuracy of KNN algorithm using optimal hyper parameter K Nearest Neighbor = 84 is 58.67%.

6.2 Naïve Bayes algorithm

Naïve Bayes classifiers are simple classifiers based on strong independence assumption between features. It is a supervised learning algorithm based on Bayes theorem for classification problem. It is mostly used in real world applications and easy to implement. Naive Bayes assumes conditional independence, meaning the relationship between all input features are independent. It is highly scalable and give accuracy in less amount of data within less training time.

Two classifiers are used in the Naïve Bayes:

1. Bernoulli Naïve Bayes Classifier (Grid Search CV is used to tune the hyper parameter "alpha")
2. Gaussian Naïve Bayes Classifier

The accuracy for both the classifiers is compared and then the better one is chosen for creating Classification report and Confusion matrix.

6.2.1 Bernoulli Classifier

After running the classifier for various values of "alpha", the best parameter value was found to be for alpha = 10, with accuracy of 59.52%.

```
Accuracy Based On Benoulli Naive Bayes Classifier on Training Dataset :  59.92  %
Accuracy Based On Bernoulli Naive Bayes Classifier on Test Dataset:  60.06  %
```

6.2.2 Gaussian Classifier

```
Accuracy Based On Gaussian Naive Bayes Classifier on Training Dataset :  56.82  %
Accuracy Based On Gaussian Naive Bayes Classifier on Test Dataset:  56.75  %
```

From the accuracy reports of above two, Bernoulli has better value compared to Gaussian classifier.

Classification Report: (based on Bernoulli classifier)

```
              precision    recall  f1-score   support

           0       0.61      0.93      0.74      1114
           1       0.53      0.11      0.18       754

    accuracy                           0.60      1868
   macro avg       0.57      0.52      0.46      1868
weighted avg       0.57      0.60      0.51      1868
```

Table 8. Classification report for Naïve Bayes Classifier

Confusion Matrix: (based on Bernoulli classifier)

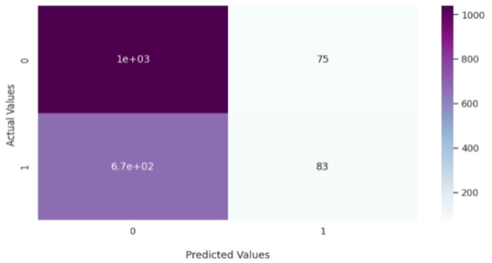

Fig. 10. Confusion matrix for Naïve Bayes Classifier

6.3 Comparing accuracy – KNN and Naïve Bayes

S.no.	Model Name	Accuracy (in %)
1.	K-neighbors (KNN)	55.67
2.	Gaussian Naïve Bayes	55.75
3.	Bernoulli Naïve Bayes	60.06

Table 9. Comparison of implemented ML models

In analysis of actual competitive exam result database, candidates successes result prediction is based on their living location, attendance and category. Prediction is done on machine learning algorithms KNN, Gaussian NB, and Bernoulli NB. Taking living location of candidate as key feature different prediction models are applied. The accuracies are calculated on applying different models and are in the range from 54% to 60% for the considered Maths database.

7. CONCLUSION

The objective of this research work was to predict the success rate of candidates appearing in state level competitive exams under defined parameters. Many candidates apply in online application but dropout on examination date. The reasons identified are lack of proper education, guidance, understanding of subject matter and also some social factors. This analysis uses different machine learning classification models for predicting the number of candidates appearing in the state level competitive exams from different districts to study, and compare rural and urban district candidates' performance.

Future implementation of this project work is to improve the accuracy percentage of the classification models. And also more parameters can be considered to predict better results. In this work only classification algorithms are used to do predictions but for future enhancement, use of regression algorithms in prediction using numerical values is possible.

Feature extraction techniques were used to analyze competitive exam result database and found that improvement in educational level and social facilities is required. This study is useful for policy makers and educational institutes in order to increase facilities in rural districts. Upgrading educational awareness in rural districts may increase their economic status by getting government jobs. This analysis helps in identifying the rural districts from where least candidates apply for competitive exams. It also identifies those few small populated rural district from where candidates never score well in competitive exams.

ACKNOWLEDGEMENT

Sincere thanks goes to Dr. Pradeep Choube, Advisor, Chhattisgarh Professional Examination Board, Raipur, for his support and motivation throughout this work.
Furthermore, we are grateful to our family for their unwavering support and understanding.

REFERENCES

1. Poudyal S; Mahnas J.; Mohammadi A; John E. Ball; Prediction of Student Academic Performance Using a Hybrid 2D CNN Model, Electronics 2022, 11, 1005. https://doi.org/10.3390/electronics11071005 https://www.mdpi.com/journal/electronics

2. Mustafa Y; Educational data mining: prediction of students' academic performance using machine learning algorithms, Smart Learning Environments (2022) 9:11 https://doi.org/10.1186/s40561-022-00192-z Smart Learning Environments-Springer open.

3. Aulck, L.; Velagapudi, N.; Blumenstock, J.; West, J. Predicting student dropout in higher education. In Proceedings of the ICML Workshop on #Data4Good: Machine Learning in Social Good Applications, New York, NY, USA, 24 June 2016; Available online: https://arxiv.org/pdf/1606.06364.pdf (accessed on 28 September 2021).

4. Kemper, L.; Vorhoff, G.; Wigger, B.U. Predicting student dropout: A machine learning approach. Eur. J. High. Educ. **2020**, 10, 28–47. [CrossRef].

5. Poudyal, S.; Mohammadi-Aragh, M.J.; Ball, J.E. Data Mining Approach for Determining Student Attention Pattern. In Proceedings of the IEEE Frontiers in Education Conference (FIE), Uppsala, Sweden, 21–24 October 2020; IEEE: Manhattan, NY, USA, 2020.

6. Poudyal, S.; Morteza, N.; Mohammad, N.; Ghodsieh, G. Machine Learning Techniques for Determining Students' Academic Performance: A Sustainable Development Case for Engineering Education. In Proceedings of the International Conference on Decision Aid Sciences and Applications (DASA), Online, 8–9 November 2020.

7. Goga, M.; Kuyoro, S.; Nicolae, G. A Recommender for Improving the Student Academic Performance. Soc. Behav. Sci. **2015**, 180, 1481–1488.

8. Pedro, S.; Luís, C.; Carlos, S.; João, M.; Rui, A. A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. In Proceedings of the International Conference on Educational Data Mining (EDM), Madrid, Spain, 26–29 June 2015.

9. SujanPoudyal , Mahnas J. Mohammadi-Aragh and John E. Ball, Hybrid Feature Extraction Model to Categorize Student - Attention Pattern and Its Relationship with Learning, Electronics 2022, 11, 1476. (https://doi.org/10.3390/electronics11091476, https://www.mdpi.com/journal/electronics).

10. Moustafa M. Nasralla , Basiem Al-Shattarat , Dhafer J. Almakhles Abdelhakim Abdelhadi and Eman S. Abowardah ,Futuristic Trends and Innovations for Examining the Performance of Course Learning Outcomes Using the Rasch Analytical Model. Electronics 2021, 10,727. (https://doi.org/10.3390/electronics10060727).

11. Mustafa Yağci, Educational data mining: prediction of students' academic performance using machine learning algorithms, Smart Learning Environments (2022) 9:11https://doi.org/10.1186/s40561-022-00192-zSmart Learning Environments-Springer open.

12. Afshan Latif, Aqsa Rasheed, Umer Sajid, Jameel Ahmed, Nouman Ali, Naeem Iqbal Ratyal, Bushra Zafar, Saadat Hanif Dar, Muhammad Sajid, Content-Based Image Retrieval and Feature Extraction: A Comprehensive Review, Hindawi Mathematical Problems in EngineeringVolume2019, Article ID 9658350, 21 pages https://doi.org/10.1155/2019/9658350.

13. Rohan Nikam, Ritesh Pardeshi, Yash Patel3, and Ekta Sarda, Deep Learning based Automatic Extraction of Student Performance from Gazette Assessment Data, ITM Web of Conferences 40, 03022 (2021)ICACC-2021 ttps://doi.org/10.1051/itmconf/20214003022.

14. Zhengyang Wu, Tao He, Chenjie Mao, Changqin Huang, Exam paper generation based on performance prediction of student group, ELSEVIER, https://doi.org/10.1016/j.ins.2020.04.043

15. Xiaoyong Li, Yong Zhang· Huimin Cheng · Mengran Li, Baocai Yin, Student achievement prediction using deep neural network from multi-source campus data, Complex & amp; Intelligent Systems (2022) 8:5143–5156 https://doi.org/10.1007/s40747-022-00731-8.

16. Vivek Anand, Saurav Kumar, A. Neela Madheswari; Students Results Prediction using Machine Learning Techniques , Int. J. Adv. Sci. Eng. Vol.3 No.2 325-329 (2016) 325, ISSN 2349 5359.

17. P.H Thilakaratne, Student Result Predictor Using Machine Learning Techniques, A Dissertation Submitted for the Degree of Master of Business Analytics, 2018/BA/035-18880358,University of Colombo School of Computing -2021

18. Sheena Angra, Sachine Ahuja, machine Learning and its Applications: A Review, International Conference on Big data Analytics and Comp Int.—ICBDACI-2017.