Gender Classification

Abstract

This project aims to develop a gender classification model based on facial features, encompassing long hair presence, forehead dimensions, nose characteristics, lip thinness, and the distance between the nose and lips.

On a societal level, the project seeks to contribute to gender equality. In the realm of security, integrating automated gender classification can enhance identification processes in public spaces and assist law enforcement agencies in investigations. Furthermore, the project's application extends to marketing and consumer analytics, offering potential benefits such as targeted advertising and insights into consumer behavior.

The approach involves data preprocessing including data visualization, data cleaning (handling missing values, handling the duplicates), predicting the outliers by box plot and dimensionality reduction using PCA, LDA, and SVD. The implementation of different machine learning algorithms and neural network are also achieved, performance will be gauged in terms of the model's accuracy, precision, recall, and confusion matrix.

The models were trained and evaluated on a labeled dataset and achieved different accuracies. Before training the models, the data was split into 80% for training and 20% for testing. The Naive Bayes algorithm demonstrated the highest accuracy of 97.21%, outperforming other algorithms such as Bayesian Model (55.79%), Decision Tree Classifier with entropy (96.29%), Decision Tree Classifier with error estimation (95.05%), LDA (96.59%), NN (94.89%), KNN with Euclidean distance (97.06%), while KNN with Minkowski distance (97.06%), and KNN with Manhattan distance (96.90%).

The findings suggest that Naive Bayes and k-Nearest Neighbors with various distance metrics are effective in gender prediction, showcasing their potential utility in real-world applications. The comparative analysis provides insights into the strengths and weaknesses of each algorithm, aiding in the selection of the most suitable approach for gender prediction tasks.

Introduction

main problem:

The main problem is to classify gender into female and males based on the facial measurements. Which can help in targeted advertising, achieving gender equality, and in identification process.

The techniques used in the project are:

The objective of this undertaking is to construct a model for classifying gender using facial characteristics, utilizing machine learning algorithms and statistical methods. In addition to traditional techniques like Chi-square test, ANOVA, Z-test and calculation of minimums, means and standard deviations; dimensionality reduction approaches such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Singular Value Decomposition (SVD) are essential in optimizing the effectiveness of the model.

When it comes to machine learning algorithms, the Naive Bayes algorithm leads the pack with a high accuracy rate of 97.21%. Alongside this top performer are other models including Decision Tree Classifier variations, Linear Discriminant Analysis (LDA), Neural Network (NN) and k-Nearest Neighbors (KNN) which measure distance metrics accordingly. These algorithms undergo a meticulous assessment using an extensive range of assessments such as confusion matrix readings and evaluating precision along with recall rates for effective analysis purposes.

Contribution:

The primary contribution of this project is the meticulous comparison of diverse models for gender classification based on facial features. By systematically applying various machine learning algorithms and neural network approaches, the analysis aimed to reveal the strengths and weaknesses of each model, offering insights into performance metrics like accuracy, precision, recall, and confusion matrices. Notably, the study highlighted the exceptional performance of the Naive Bayes algorithm and various k-Nearest Neighbors (KNN) configurations, with practical implications for applications in security, marketing, and consumer analytics. Additionally, the project demonstrated a commitment to optimizing model efficiency and computational performance through the exploration of dimensionality reduction techniques, including Principal Component Analysis (PCA).

Organization of the rest of the Project:

The rest of the paper is organized into the following sections. Section two concentrates on the literature work reviewing the studies on gender classification using different datasets. The related methodology is reported in Section three. Section four delineates the proposed models, and results achieved, and discussion are in section five. The final section concludes the research work and talks about the scope for future work.

Related Work

Paper	year	methods	Dataset type	Results (accuracy)
[1]	2023	decision tree classification	measurements	0.9842
[2]	2023	StyleGAN and VGG19 networks	images	0.9478
[3]	2022	Random forest	Signature	0.962
[4]	2020	KNN	voice	0.968
[5]	2021	CNN	Fingerprint	0.99
[6]	2021	SVM	Measurements	0.96
[7]	2021	CNN	images	0.9512
[8]	2021	BLSTM	Voice	0.9176
[9]	2023	MS Kinect	Measurements	0.975
[10]	2023	back propagation algorithm	Measurements	0.7894

This paper [1] makes gender classification based on measurements of physical body parts like bodyweight, height, and circumferences.

methods:

- **Data Preparation:**
  Use the Body Fat dataset for benchmarking and split the dataset into training and testing sets.
- **Boosting Algorithm Initialization:**
  Initialize a boosting algorithm to enhance the performance of classification algorithms.

- **Sample Ranking:**
  For the training dataset, rank samples based on error rates, sorting them from best to worst.
- **Generate a root node:**
  Generate a root node N, which becomes a leaf node with the required class if the sample goes to the same class.
- **Attribute Selection:**
  Test attributes with the highest information gain are selected.
- **Sub-Tree Development:**
  For each possible outcome, develop a sub-tree based on the partition P.
- **Repeat**

paper [2] make a gender classification based on images dataset.

Methods:

The proposed method comprises two main components: a pre-trained StyleGAN network generating realistic 1024 × 1024 facial images, labeled based on observed gender criteria, and a gender classifier trained from scratch, further fine-tuned using pre-trained VGG16 and VGG19 networks. The StyleGAN network is acquired from the official GitHub page, trained on the FlickrFaces-HQ dataset. The training process involves a min-max game between the generator and discriminator, with loss functions determining their performance.

Paper [3] makes gender classification based on signature.

Methods:

feature extraction reduces resource requirements when describing extensive datasets, with specific features detailed for online handwritten signature. Random Forest (RF), leverages the bagging of tree-structured classifiers, demonstrating robust performance in prediction models, especially in handling numerous training samples and high-dimensional spaces.

Methodology

After preparing the data and preprocessing, different models will be applied to train on the data. The models used in the project are :

Naive Bayes Algorithm:

Utilizing probabilistic reasoning, it facilitates gender classification by assuming independence among the various facial features.

Bayesian Model:

It is the same as naïve bayes, but it assumes some dependency between labels and the target.

Decision Tree Classifier:

Employing Decision Trees with entropy and error estimation criteria, this technique discerns gender based on hierarchical decision-making processes.

Linear Discriminant Analysis (LDA):

Using LDA as a dimensionality reduction technique, this method enhances discriminative power by maximizing class separation. And it is also used in classification, as it finds the best fit line that separates the data.

Principal Component Analysis (PCA):

It aimed to capture the most significant variations in facial features while reducing the overall dimensionality of the data.

k-Nearest Neighbors (KNN):

Implementing KNN with various distance metrics (Euclidean, Minkowski, Manhattan) to classify gender based on proximity to neighboring data points, as KNN classifies based on the vote of the similarity items.

Neural Network (NN):

Introducing a Neural Network into the framework, a deep learning paradigm capable of learning intricate patterns within facial features

Proposed Model

Preprocessing phase:

Prepare data for processing, the duplicates in the data have been dropped, the data contains no missing value. then visualizing the data to get insights from it about the features to be used and the relation about then. The box plot is drawn to capture the outliers, in this dataset the outliers is so small, so it preferred to not to remove it to provide variety to the data.

The statistical data are also calculated like min, max mean, standard deviation, covariance matrix and correlation matrix.

Feature selection phase:

In this phase the features are chosen to predict. But the column in this dataset is few and all of them will be included in the training so there is no feature selection.

Feature Reduction phase:

Before feature reduction the feature gets Standardize, then LDA is done which find the best fit line that separate the two classes female and male, PCA that capture to which features the target got variant. Finaly SVD is a factorization method to get the most important features in the data.

Classification phase:

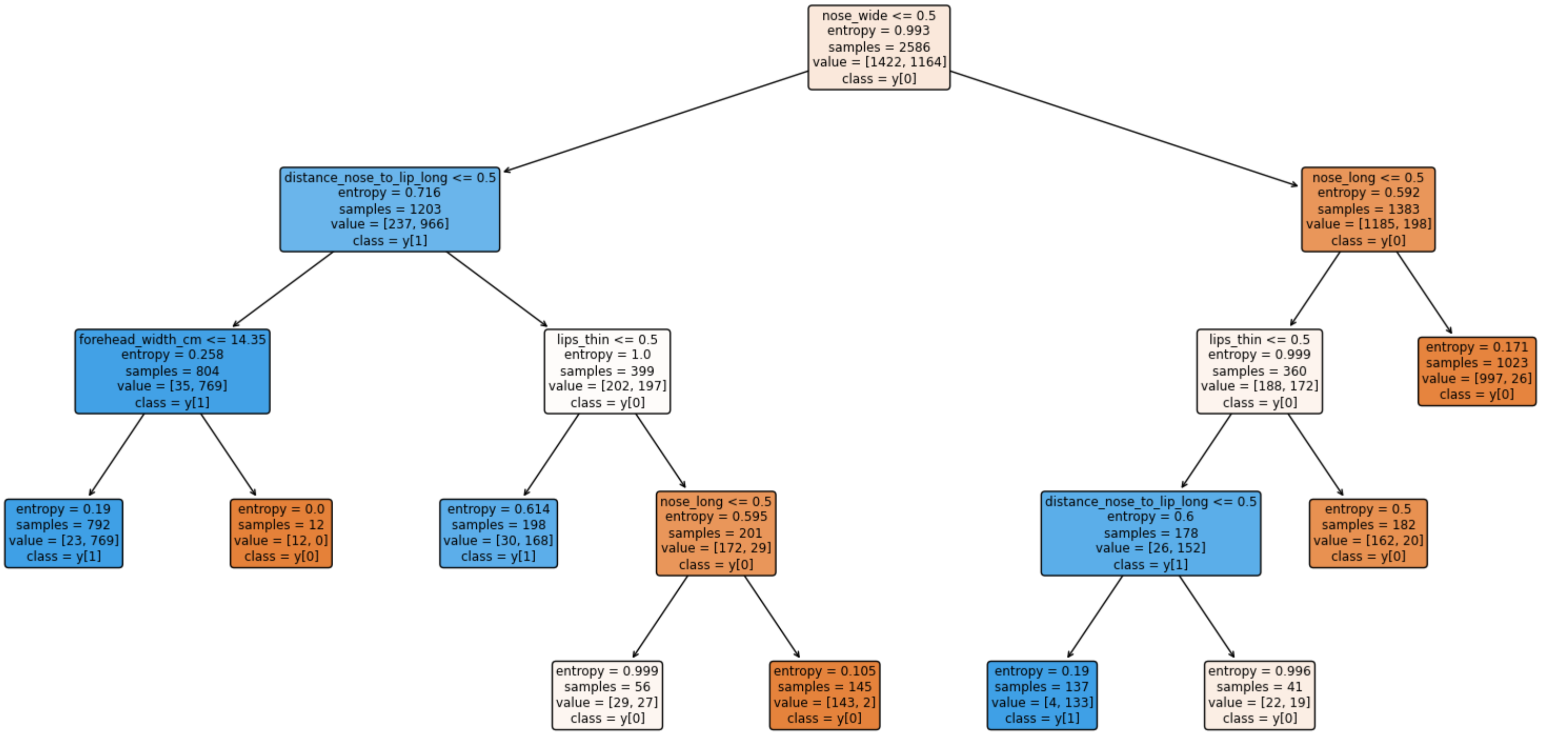
Before using model to train on the datasets, the dataset is split into 80% for training and 20% for testing.

Then the models will be trained on the data once and will be trained with it using kfold cross validation, the k here is 5.

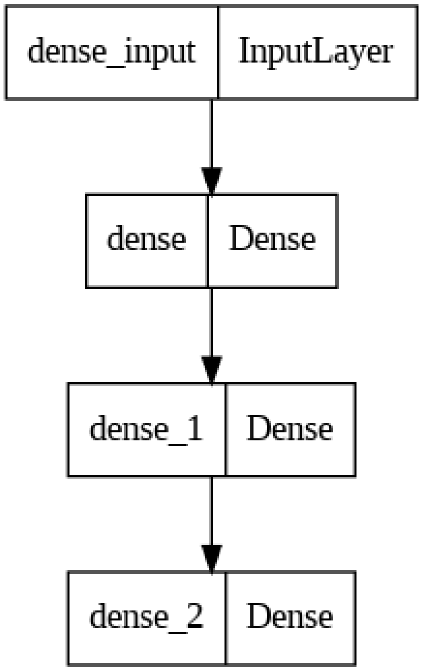
The first model is the naïve bayes model which assumes independently between features of the data.

The second model is Bayesian belief network which oppose to naïve bayes, it assumes some dependency between the features, as target will be dependent on all the features.

The third model is a decision tree which build tree of feature to make it like condition to reach the target.



The fourth model is LDA which try to separate the two classes by a straight line.



The fifth model is a neural network which consists of an input layer, hidden layer, and output layer.

The sixth model is KNN, which makes prediction make on the vote of k similarity.

The know the nearest k items, the distance is calculated, because the data is numerical, there is three possible distances, Euclidean distance, Minkowski distance and Manhattan distance.

After this each model will be evaluated using confusion matrix and ROC curve.

Evaluation metrics:

Considering factors such as correctness, false positives, false negatives, and the balance between precision and recall, metrics collectively provide a fine examination of a machine learning model performance.

Confusion Matrix:

Breaks down predictions into four categories, True Positives, True Negatives, False Positives, and False Negatives.

Accuracy:

measures the overall correctness of predictions.

Calculation: (True Positives + True Negatives) / Total Instances.

Precision:

assesses the model's ability to correctly identify positive instances among all instances predicted as positive.

Calculation: True Positives / (True Positives + False Positives).

Recall (Sensitivity):

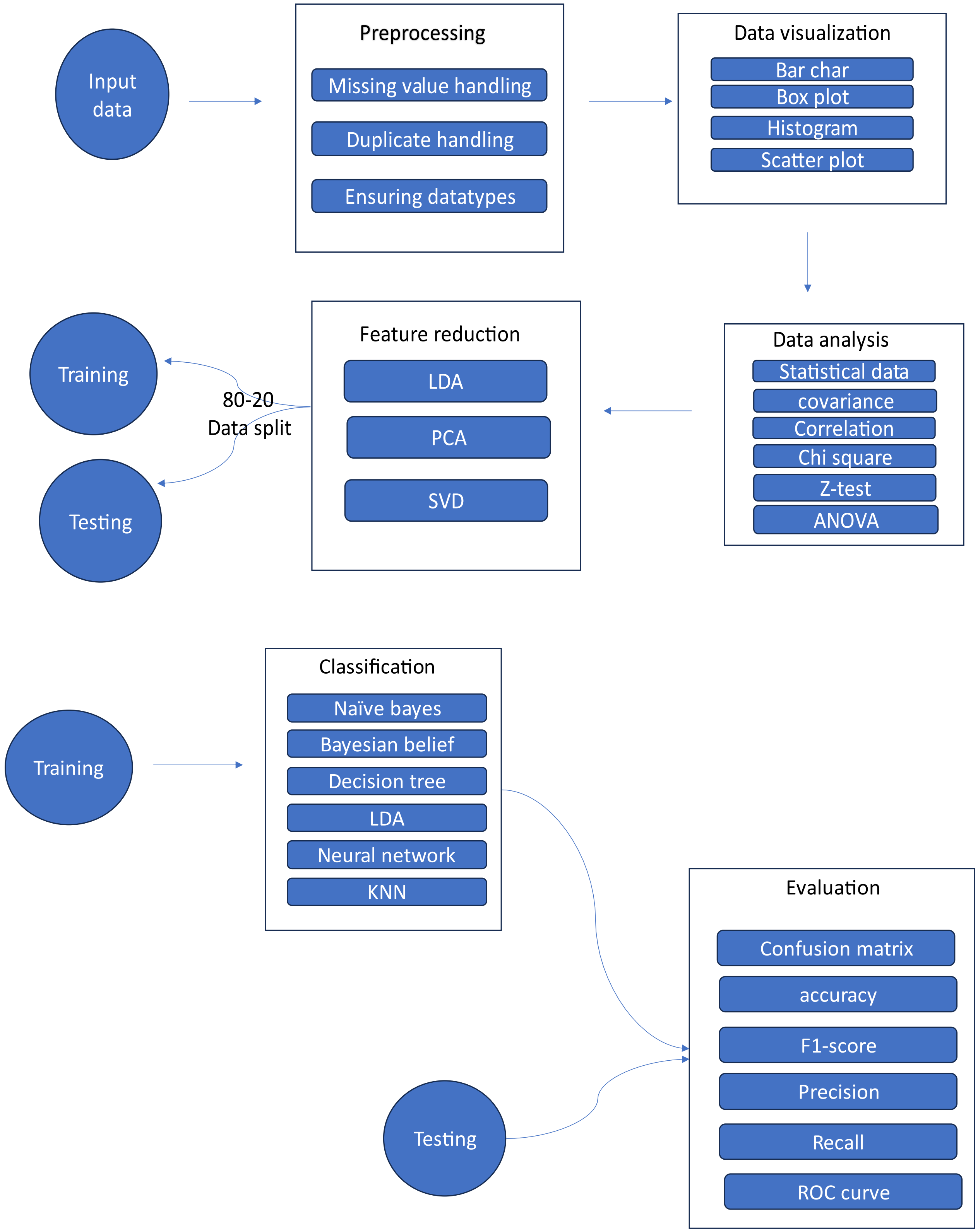
measures the model's capability to capture all positive instances.

Calculation: True Positives / (True Positives + False Negatives).

F1-Score:

balances precision and recall, providing a mean between the two metrics.

Calculation: 2 * (Precision * Recall) / (Precision + Recall).


Input data
Preprocessing
Missing value handling
Duplicate handling
Ensuring datatypes
Data visualization
Bar char
Box plot
Histogram
Scatter plot
Data analysis
Statistical data
covariance
Correlation
Chi square
Z-test
ANOVA
Feature reduction
LDA
PCA
SVD
80-20
Data split
Training
Testing
Training
Classification
Naïve bayes
Bayesian belief
Decision tree
LDA
Neural network
KNN
Evaluation
Confusion matrix
accuracy
F1-score
Precision
Recall
ROC curve
Testing

Results and discussion

dataset description:

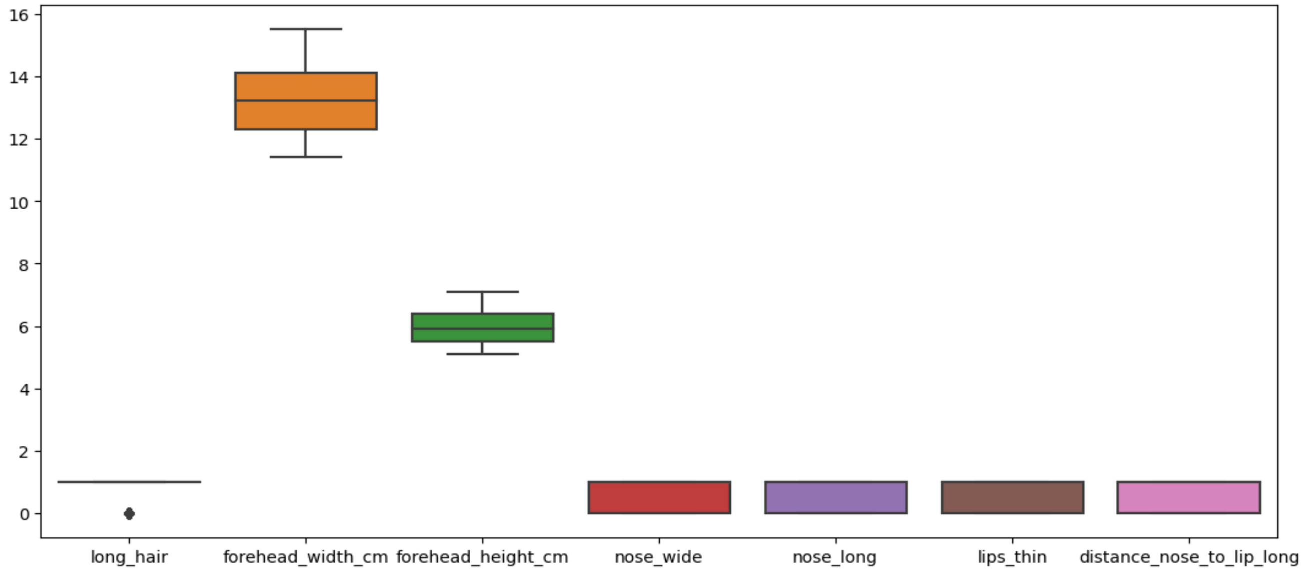
This dataset contains 5001 rows and 7 features and a label column(gender):

- long_hair: This column contains 0's and 1's where 1 is "long hair" and 0 is "not long hair".
- forehead_width_cm: This column is in CM's. This is the width of the forehead.
- forehead_height_cm: This is the height of the forehead and it's in Cm's.
- nose_wide: This column contains 0's and 1's where 1 is "wide nose" and 0 is "not wide nose.
- nose_long: This column contains 0's and 1's where 1 is "Long nose" and 0 is "not long nose".
- lips_thin: This column contains 0's and 1's where 1 represents the "thin lips" while 0 is "Not thin lips".
- distance_nose_to_lip_long: This column contains 0's and 1's where 1 represents the "long distance between nose and lips" while 0 is "short distance between nose and lips".
- gender: This is either "Male" or "Female".

The data contains no null values, and it has about 1768 duplicate. there is no need for binding because the algorithms will work with numerical data.

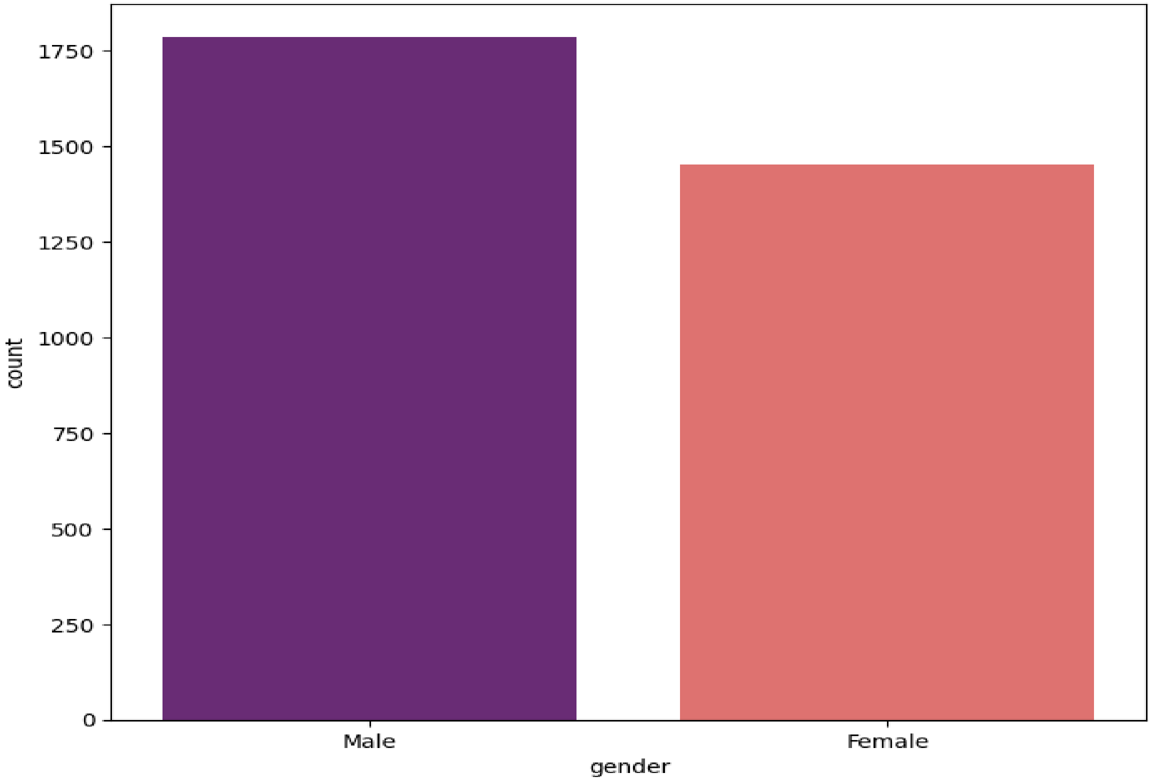
data visualization:

box plot



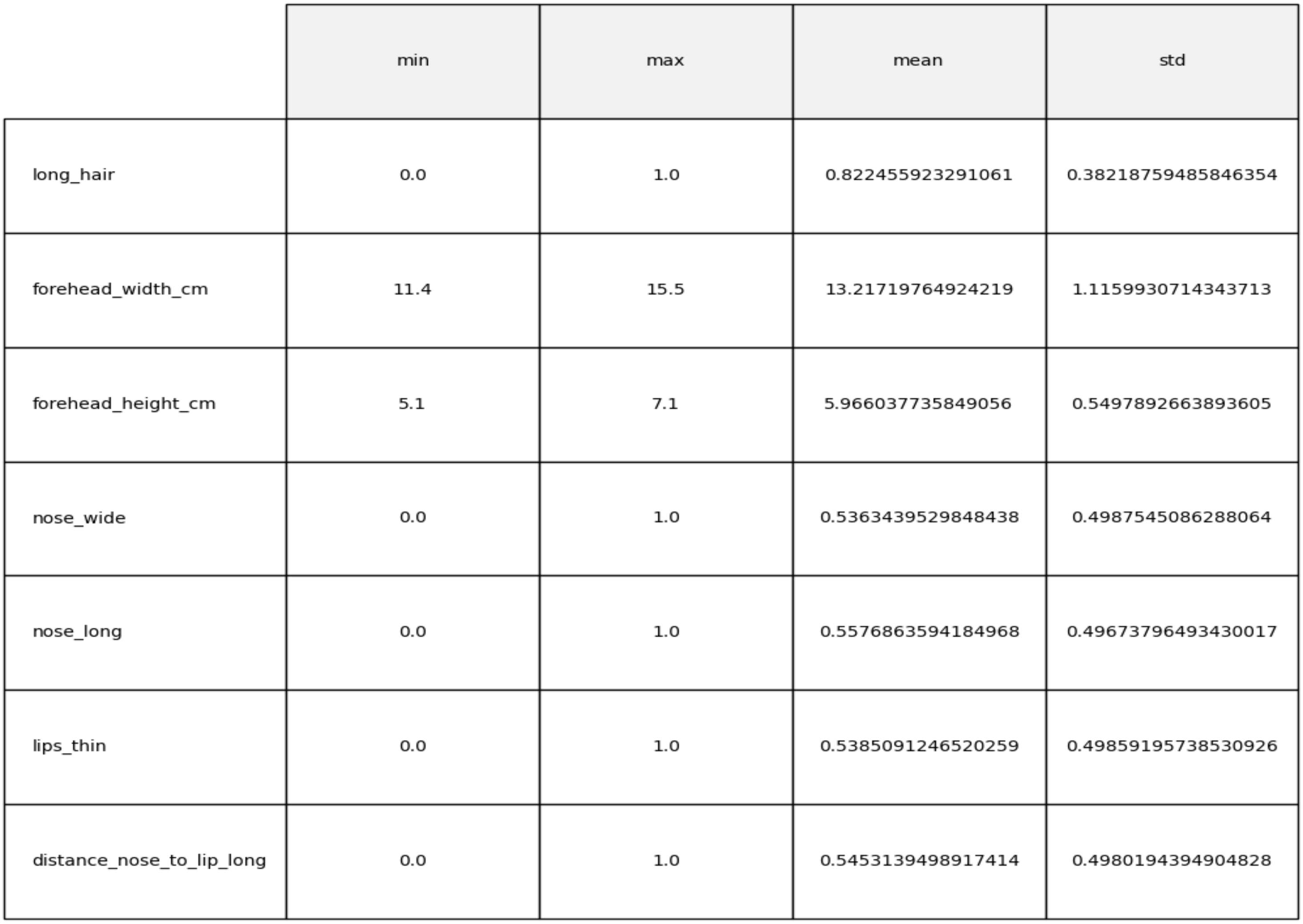
bar char

showing the number of females and males in the dataset.



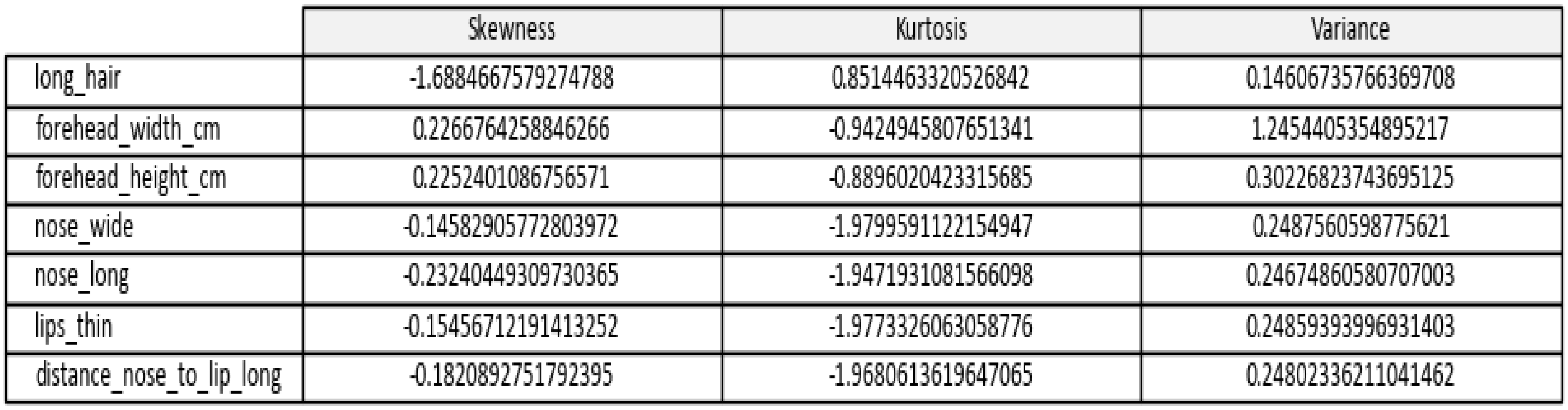
data analysis:

min, max , mean , std



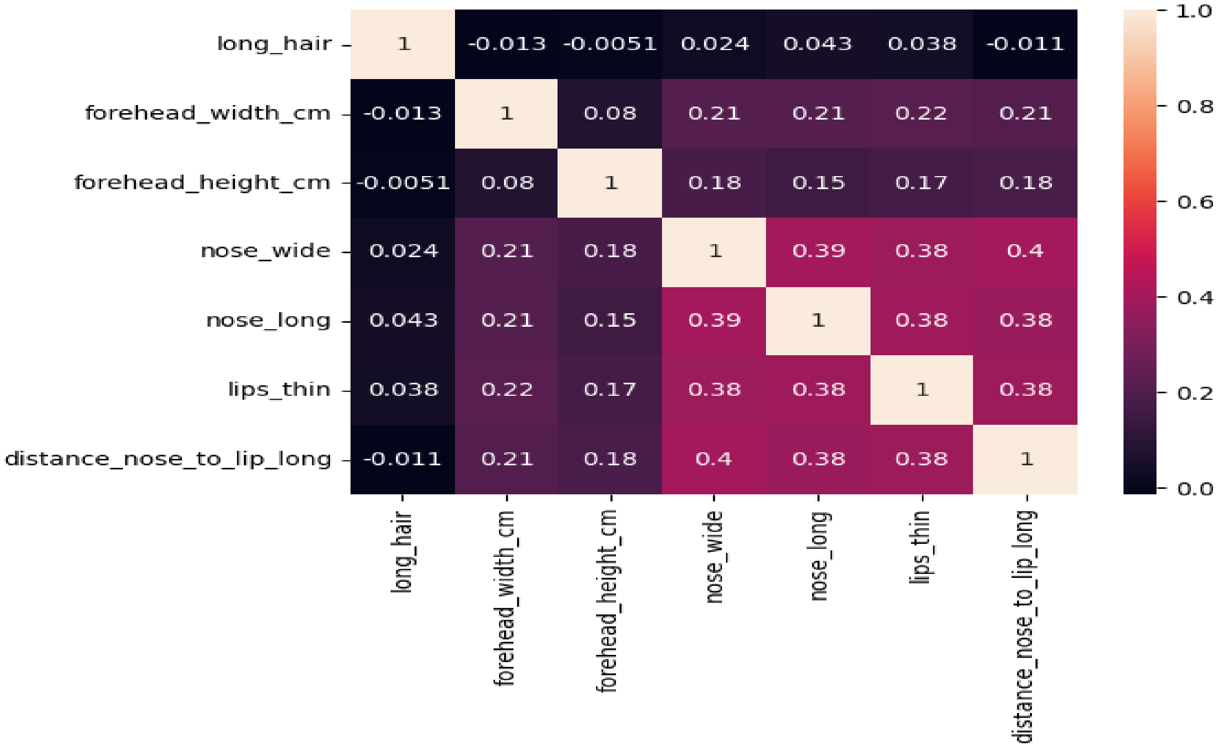
	min	max	mean	std
long_hair	0.0	1.0	0.822455923291061	0.38218759485846354
forehead_width_cm	11.4	15.5	13.21719764924219	1.1159930714343713
forehead_height_cm	5.1	7.1	5.966037735849056	0.5497892663893605
nose_wide	0.0	1.0	0.5363439529848438	0.4987545086288064
nose_long	0.0	1.0	0.5576863594184968	0.49673796493430017
lips_thin	0.0	1.0	0.5385091246520259	0.49859195738530926
distance_nose_to_lip_long	0.0	1.0	0.5453139498917414	0.4980194394904828

skewness, kurtosis, variance

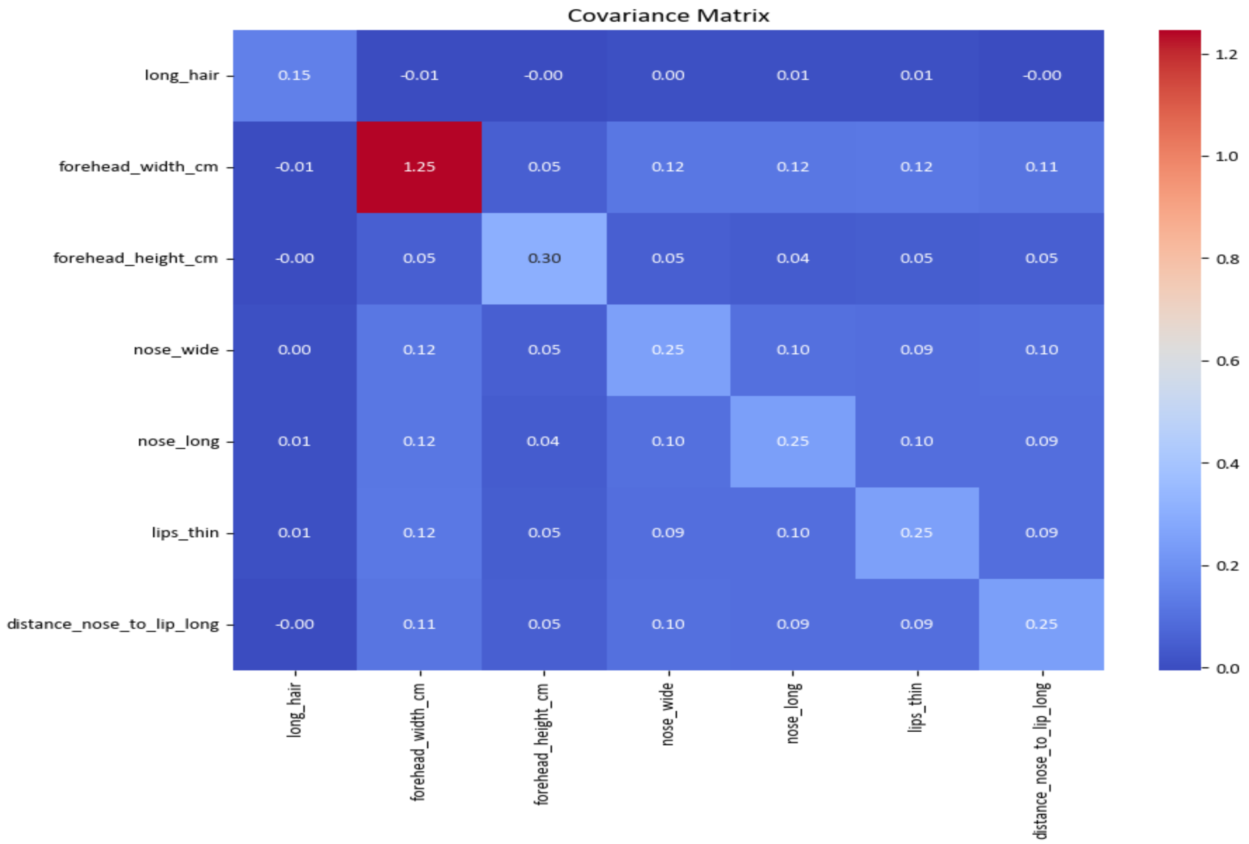


	Skewness	Kurtosis	Variance
long_hair	-1.6884667579274788	0.8514463320526842	0.14606735766369708
forehead_width_cm	0.2266764258846266	-0.9424945807651341	1.2454405354895217
forehead_height_cm	0.2252401086756571	-0.8896020423315685	0.30226823743695125
nose_wide	-0.14582905772803972	-1.9799591122154947	0.2487560598775621
nose_long	-0.23240449309730365	-1.9471931081566098	0.24674860580707003
lips_thin	-0.15456712191413252	-1.9773326063058776	0.24859393996931403
distance_nose_to_lip_long	-0.1820892751792395	-1.9680613619647065	0.24802336211041462

heat map



Covariance Matrix



Chi- square Test, Z-test, ANOVA

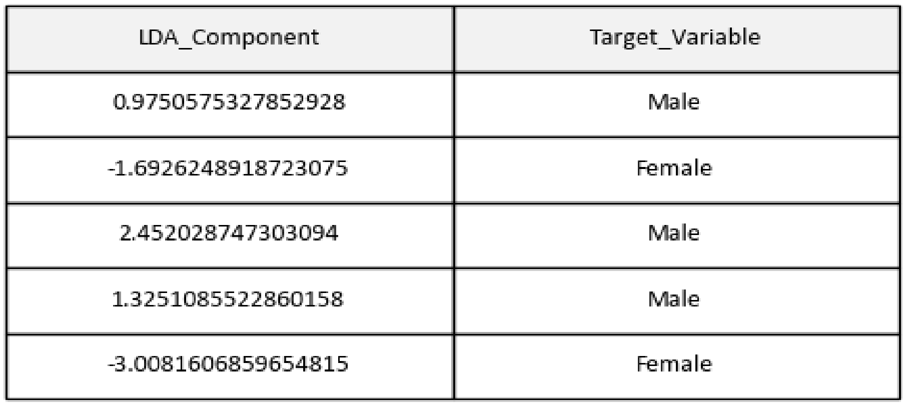
Chi- square Test	**the p value is 1.0** independence
Z-test (sample size = 300)	Z-Score: -0.41343019804851133 Critical Z-Score: 1.6448536269514722 Fail to Reject Null Hypothesis
ANOVA	F Statistic: 199199.9751074745 P-value: 0.0 Reject the null hypothesis. There is a significant difference between the means

Feature Reduction

Before applying any feature reduction algorithm, data scaling must be done by removing each value from the mean and scaling to unit variance.

Linear Discriminate Analysis (LDA):

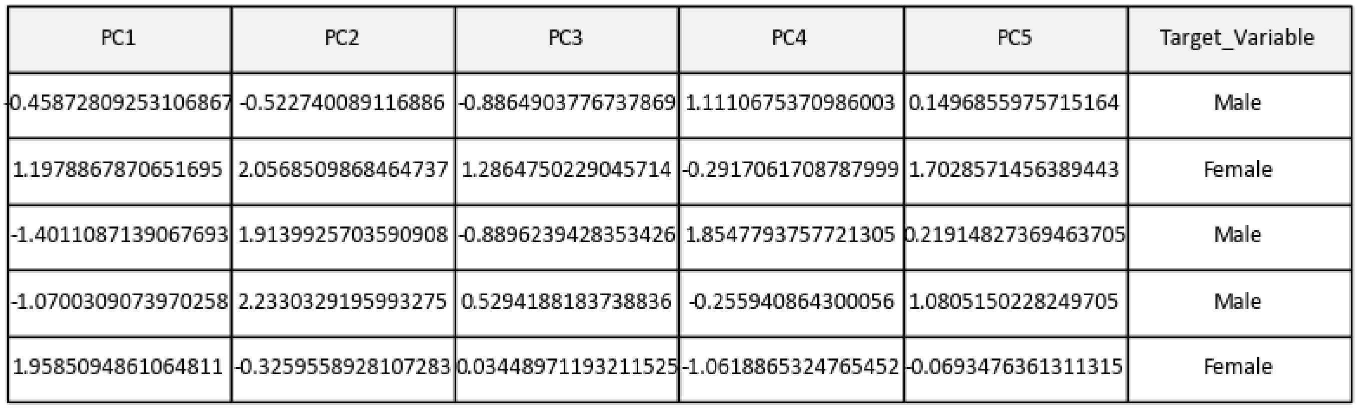
It reduced the data to (3233, 1)



LDA_Component	Target_Variable
0.9750575327852928	Male
-1.6926248918723075	Female
2.452028747303094	Male
1.3251085522860158	Male
-3.0081606859654815	Female

Principle Component Analysis (PCA):

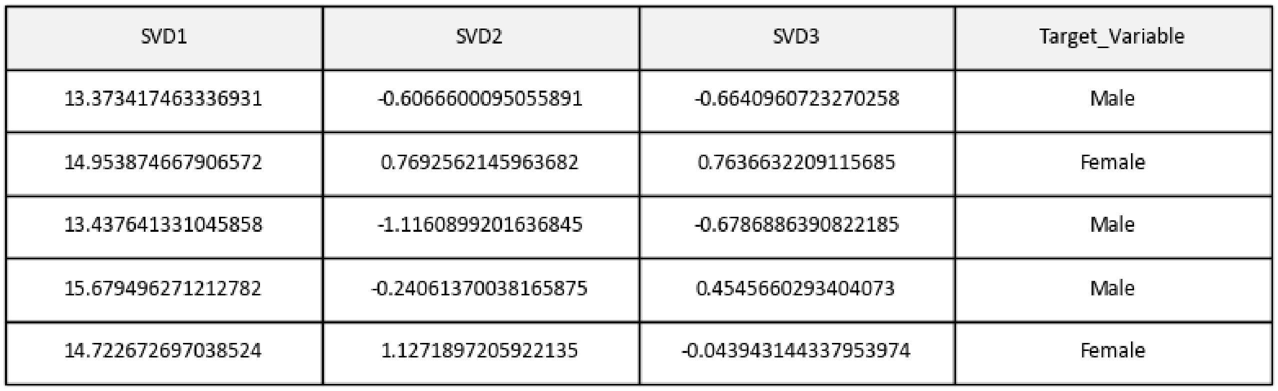
It reduces the data to (3233, 5)



PC1	PC2	PC3	PC4	PC5	Target_Variable
-0.45872809253106867	-0.522740089116886	-0.8864903776737869	1.1110675370986003	0.1496855975715164	Male
1.1978867870651695	2.0568509868464737	1.2864750229045714	-0.2917061708787999	1.7028571456389443	Female
-1.4011087139067693	1.9139925703590908	-0.8896239428353426	1.8547793757721305	0.21914827369463705	Male
-1.0700309073970258	2.2330329195993275	0.5294188183738836	-0.255940864300056	1.0805150228249705	Male
1.9585094861064811	-0.3259558928107283	0.03448971193211525	-1.0618865324765452	-0.0693476361311315	Female

Singular Value Decomposition (SVD)

It reduces the data to (3233, 3)



SVD1	SVD2	SVD3	Target_Variable
13.373417463336931	-0.6066600095055891	-0.6640960723270258	Male
14.953874667906572	0.7692562145963682	0.7636632209115685	Female
13.437641331045858	-1.1160899201636845	-0.6786886390822185	Male
15.679496271212782	-0.24061370038165875	0.4545660293404073	Male
14.722672697038524	1.1271897205922135	-0.043943144337953974	Female

The data is split into training and testing, before training the models and evaluating them. In the training part, the data contains 2586 rows which is 80% of the data, it provides large examples for the model to capture patterns and relationships within the dataset. Moreover, the testing part consists of the remaining 647 rows which is 20% of the data and it acts as an independent sample to evaluate how well is the model. By utilizing this division, we ensure that the model experiences robust training and extensive evaluation.

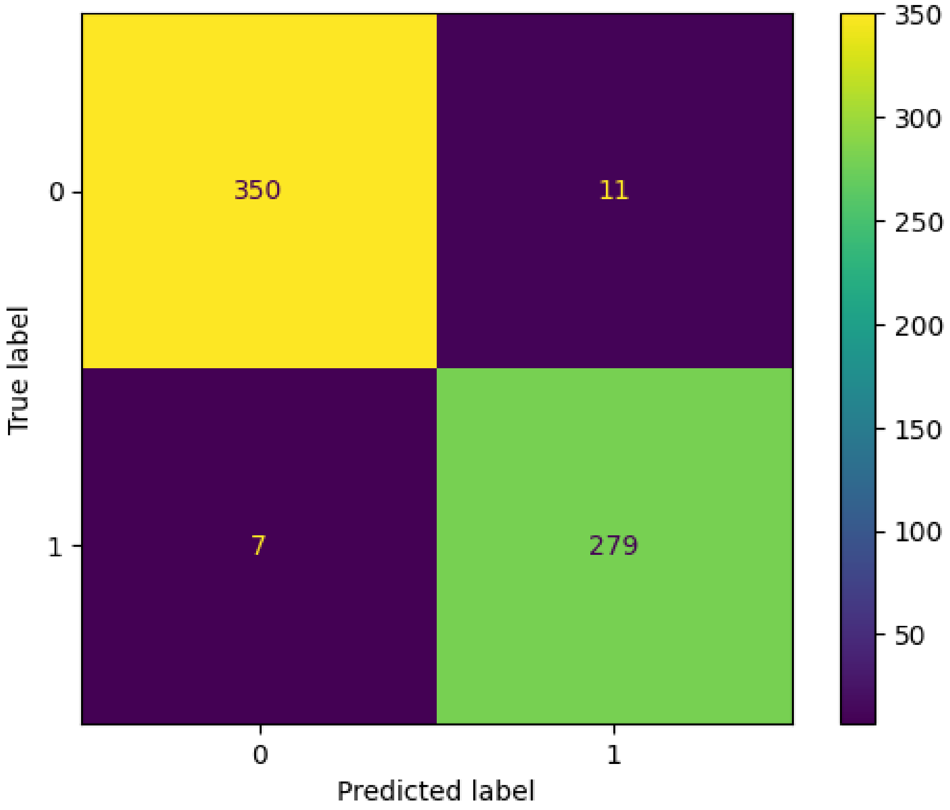
K-fold cross-validation is like giving your model a varied workout, training it on different parts of your data to ensure it's well-prepared for any situation. The average accuracy then sums up how it performed across all these training scenarios.

Model	Accuracy	Average accuracy
Naïve bayes	95.282406%	97.217929%
Decision Tree (Entropy)	93.503432%	96.290572%
Decision Tree (Error)	93.194029%	95.054096%
LDA	94.663301%	96.599691%
Neural network	94.663301%	96.445131%
KNN	95.166426%	97.063369%

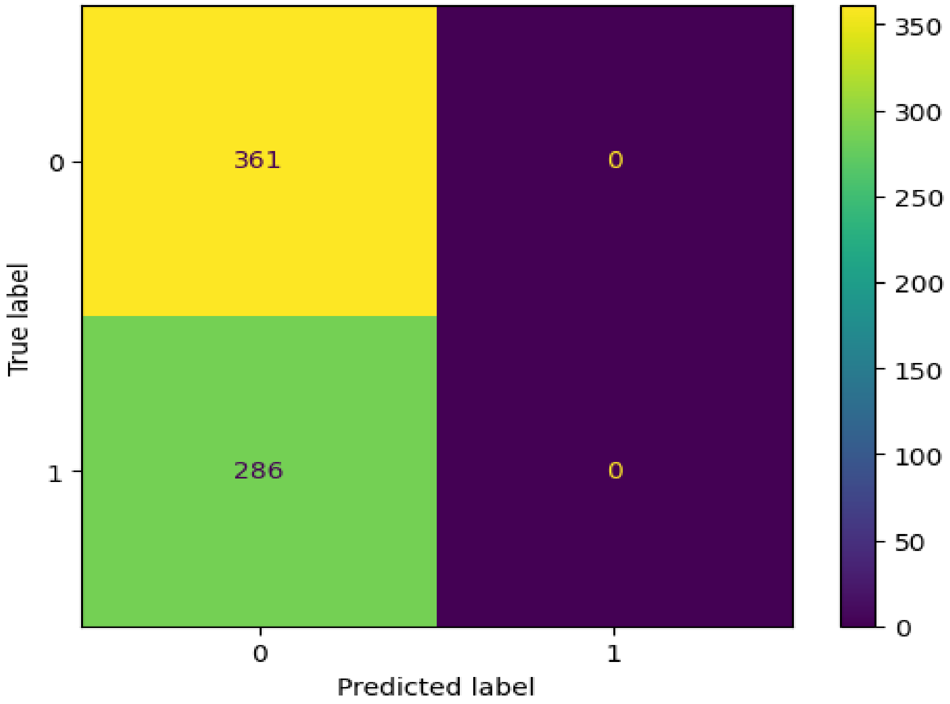
Methods results:

The model	**Accuracy**	**Error rate**	**F1-score**	**Precision**	**Recall**	**ROC (auc)**
Naïve bayes	97.21%	7.01	0.97	0.98	0.96	97.25%
Bayesian Belief Network	55.79%	286.0	0.71	0.55	1.0	50%
Decision Tree (Entropy)	96.29%	18	0.96	0.95	0.98	96.02%
Decision Tree (Error)	95.05%	15.02	0.95	0.95	0.95	95.02%
LDA	96.59%	14.01	0.96	0.96	0.97	96.44%
Neural network	94.89%	32	0.95	0.91	0.99	94.27%
KNN (Euclidean)	97.06%	7.01	0.97	0.98	0.96	97.11%
KNN(Minkowski)	97.06%	8.01	0.97	0.97	0.96	97.08%
KNN(Manhattan)	96.90%	8.01	0.97	0.97	0.97	96.94%

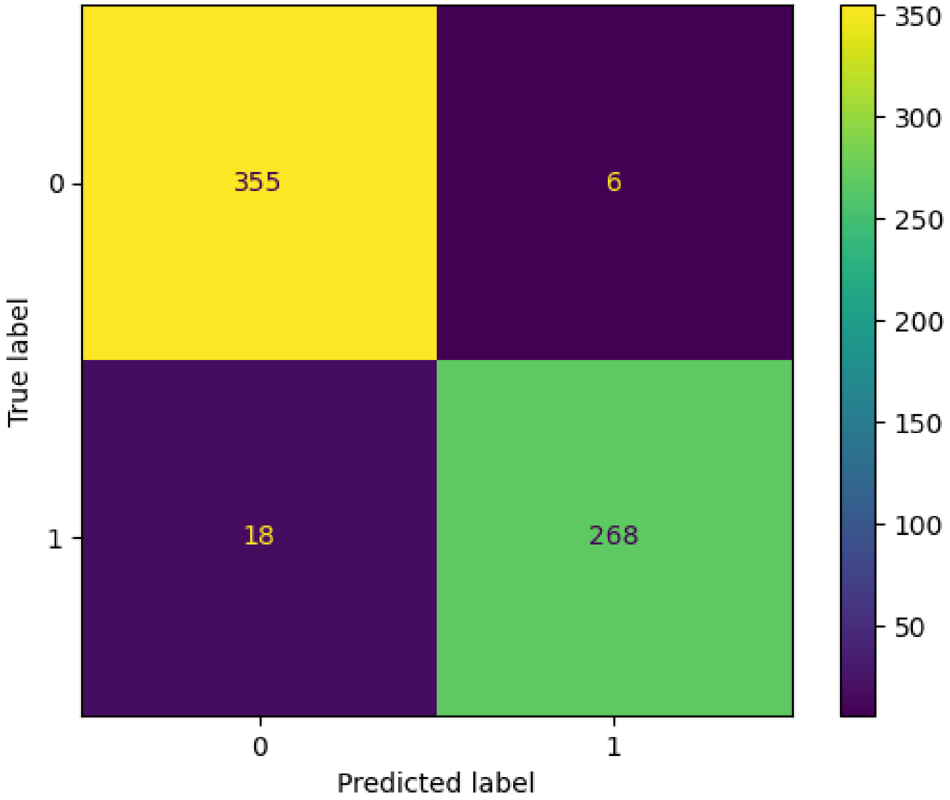
Naïve bayes confusion matrix:



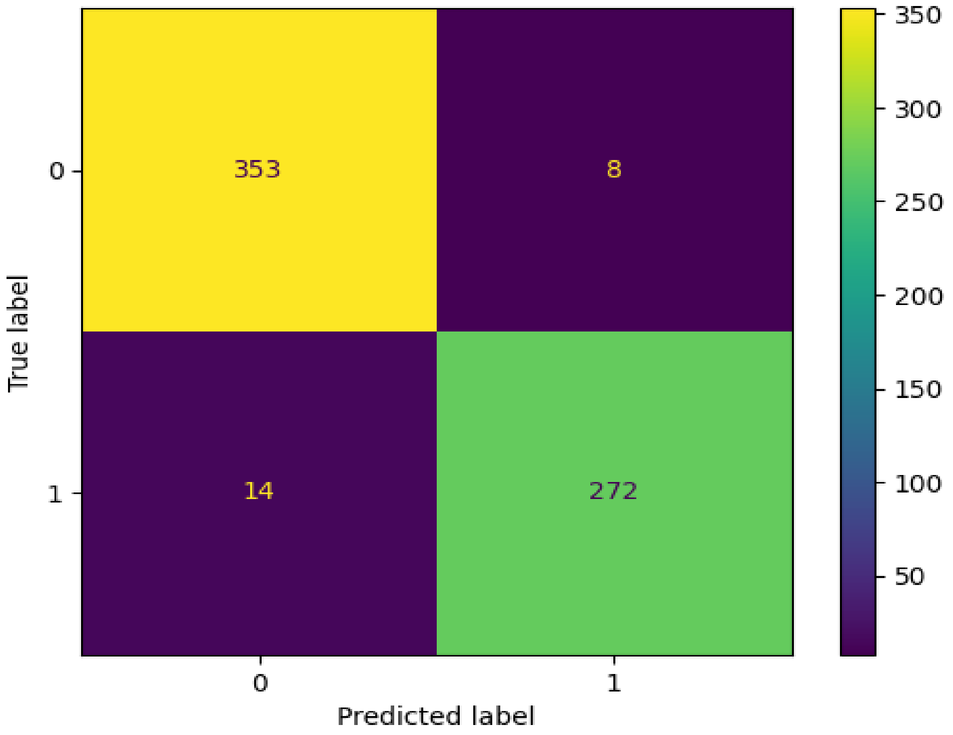
Bayes belief network:



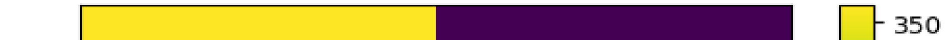
Decision tree:



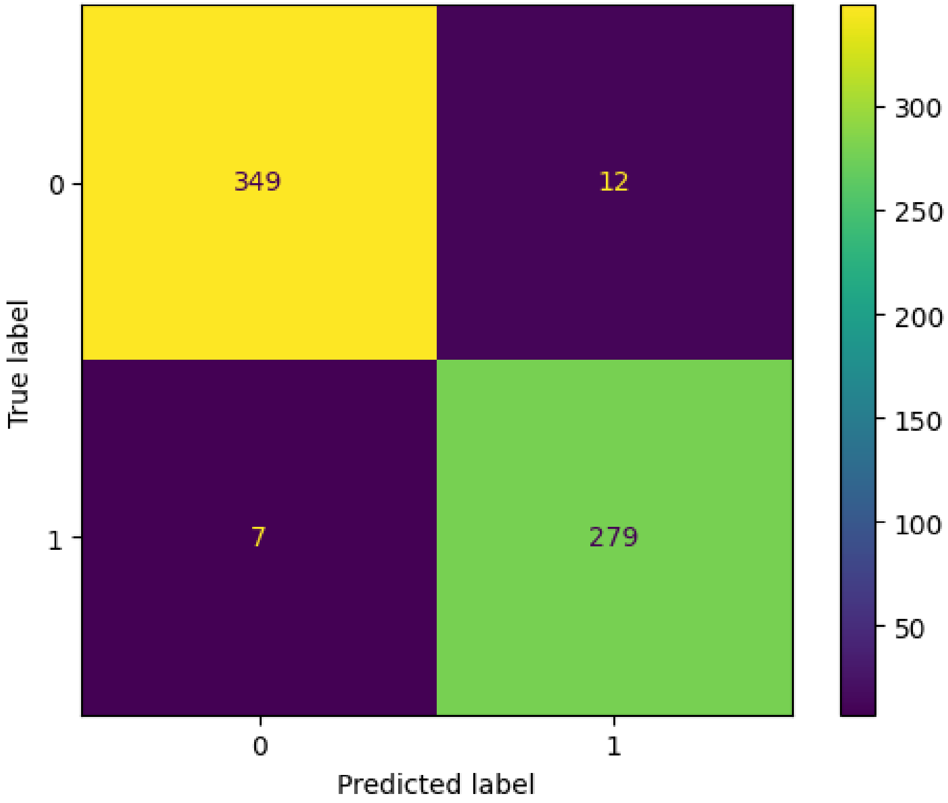
LDA confusion matrix:



Neural network confusion matrix:



KNN confusion matrix:



Conclusion and future work

In summary, the project was able to create an accurate gender identification model that relies on facial characteristics. This achievement is important for different aspects of society like promoting equal treatment between genders, bolstering security measures, and improving marketing analytics. To make this happen, various techniques such as data preprocessing and dimensionality reduction were employed along with implementing diverse machine learning algorithms alongside neural networks in a comprehensive approach.

The outcome emphasized the success of the Naive Bayes formula, which displayed a superior accuracy rate of 97.21% compared to other models. Similarly, k-Nearest Neighbors (KNN) using varying distance metrics also portrayed promising results with precision rates ranging between 96.90%-97.06%. The comparative evaluation delivered significant discernment into each technique's capabilities and limitations in gender prediction tasks, thereby facilitating informed selection processes for appropriate approaches to such undertakings.

After undergoing significant preprocessing, the dataset now comprises 5001 rows and 7 features. The thorough processing involved several steps such as data visualization, cleaning, and prediction of outliers. Additionally, feature reduction methods including Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), and Singular Value Decomposition (SVD) were employed to improve the performance efficiency gains for our model's optimization purposes.

The performance of the model was comprehensively evaluated through various metrics such as confusion matrix readings, accuracy, precision, recall and F1 score. The results indicate that gender prediction can be effectively achieved using Naive Bayes and k-Nearest Neighbors with different distance metrics which highlights their potential value in actual settings.

To advance the model performance, exploring various datasets and methods is necessary. Deep learning architectures, ensemble techniques, and alternative machine learning algorithms can be examined further to achieve better outcomes. Moreover, including wider data sets with varied characteristics may assist in improving robustness of the model for different facial features.

To sum up, this project establishes a foundation for progress in gender identification models that could be utilized in security measures, marketing strategies, and more. Further investigations into pioneering approaches will promote the continual enhancement of dependable and precise predictions on facial characteristics-based gender classification systems.

References

[1] Tabassum, H., Iqbal, M. M., Mahmood, Z., Parveen, M., & Ullah, I. (2023). Gender classification from anthropometric measurement by boosting decision tree: A novel machine learning approach. *Journal of the National Medical Association.*

[2] Raman, V., ELKarazle, K., & Then, P. (2023). Artificially Generated Facial Images for Gender Classification Using Deep Learning. Computer Systems Science and Engineering, 44(2), 1341-1355.

[3] Kumar, S., Gornale, S. S., Siddalingappa, R., & Mane, A. (2022). Gender classification based on online signature features using machine learning techniques. International Journal of Intelligent Systems and Applications in Engineering, 10(2), 260-268.

[4] Uddin, M. A., Hossain, M. S., Pathan, R. K., & Biswas, M. (2020, August). Gender recognition from human voice using multi-layer architecture. In 2020 International conference on innovations in intelligent systems and applications (INISTA) (pp. 1-7). IEEE.

[5] Jayakala, G. (2021). Gender classification based on fingerprint analysis. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(10), 1249-1256.

[6] Karkkainen, K., & Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1548-1558).

[7] Rwigema, J., Mfitumukiza, J., & Tae-Yong, K. (2021). A hybrid approach of neural networks for age and gender classification through decision fusion. Biomedical Signal Processing and Control, 66, 102459.

[8] Alashban, A. A., & Alotaibi, Y. A. (2021, July). Speaker gender classification in mono-language and cross-language using BLSTM network. In 2021 44th International Conference on Telecommunications and Signal Processing (TSP) (pp. 66-71). IEEE.

[9] Azhar, M., Ullah, S., Raees, M., Rahman, K. U., & Rehman, I. U. (2023). A real-time multi view gait-based automatic gender classification system using kinect sensor. Multimedia Tools and Applications, 82(8), 11993-12016.

[10] Almomani, A., Alweshah, M., Alomoush, W., Alauthman, M., Jabai, A., Abbass, A., ... & Gupta, B. B. (2023). Age and Gender Classification Using Backpropagation and Bagging Algorithms. Computers, Materials & Continua, 74(2).