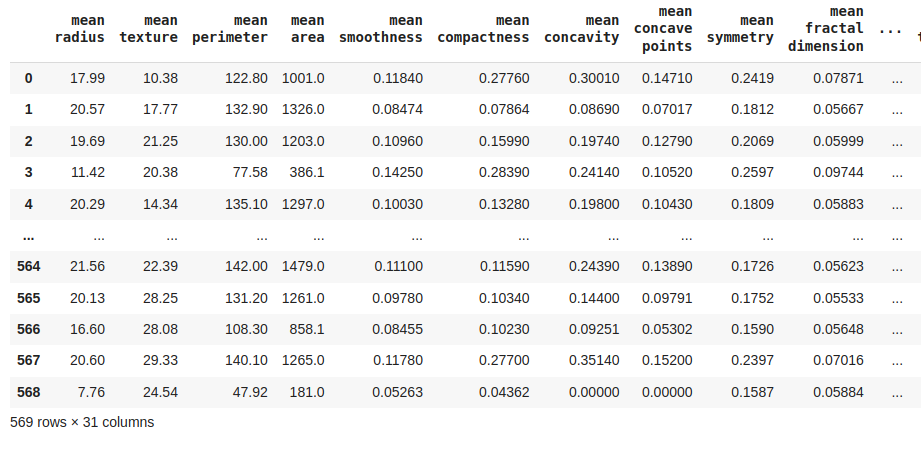
**Applying Principal Component Analysis (PCA) for Dimensionality Reduction and SVM for Classification in Breast Cancer Datasets**

**Abstract**

In this research I am going to explore the application of Principal Component Analysis for dimensionality reduction in breast cancer datasets. We know that PCA is a statistical technique mainly designed to help us simplify high dimensional data to low dimensional data by focusing and keeping the important informative features only. In this research we mainly focus on calculating the covariance matrix to understand feature relationships and then determining principal components based on the directions of maximum variance. We also calculate eigenvalues and eigenvectors of the given covariance matrix. After that we know that by selecting principal components associated with the highest variance, PCA helps us to reduce the number of features while preserving essential information. This method helps us to enhance the efficiency and accuracy of predictive models addressing issues of overfitting and optimizing feature selection in breast cancer diagnosis. In this research we also apply SVM for classification predictions a linear SVM was applied to classify the tumor as benign as non-cancerous and malignant as cancerous. This method helps us to effectively separate classes by identifying the optimal hyperplane in the feature space.

**1. Introduction to the dataset**

The dataset for breast cancer prediction includes a diagnosis label indicating whether the tumor is benign as non-cancerous and malignant as cancerous. In this dataset we have various features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The dataset computes ten real valued features for each cell nucleus, including radius (mean distance from center to perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (variation in radius lengths), compactness (perimeter^2/area - 1.0), concavity (severity of concave portions), concave points (number of concave portions), symmetry and fractal dimension ("coastline approximation" - 1). The mean, standard error, and "worst" values (mean of the three largest measurements) of these features were calculated for each image, resulting in a total of 30 features.These features together help in assessing and predicting the nature of tumors.



**2. Problem Statement and Challenges**

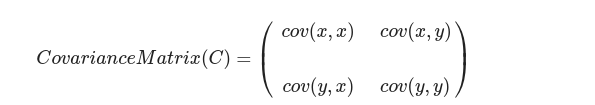
When examining the dataset, we find it has 30 features, which can be difficult to manage. This high number of features can cause problems like overfitting in machine learning models making it hard for these models to work well in the future. In the case of high dimensional datasets often include many features that are not very important for predicting the target outcome. These unnecessary features can hide important patterns and make it harder for models to work well with new data. Also, dealing with so many features can make the calculations more complex and can lead the model to focus on unimportant details instead of meaningful information. It's better to focus on the most important features.Breast cancer is the most common cancer in women and the second leading cause of cancer deaths. Diagnosing it usually starts with finding unusual changes through self-checks and imaging methods. However distinguishing between benign and malignant tumors based on numerous features is complex and models with too many features may struggle to provide accurate predictions.

**3. Objective of Dimensionality Reduction**

The main objective and goal of doing PCA on the given dataset is to reduce the dimensionality of breast cancer datasets by identifying and retaining the most informative features and applying SVM to classify the tumor as benign as non-cancerous and malignant as cancerous which helps us to make prediction in very short time and accurately. This method mainly aims to simplify the dataset by focusing on the principal components that capture the maximum variance. Also by reducing the number of features from the given dataset we know that PCA helps in mitigating the challenges associated with high dimensional data by reducing to lower dimensional data which also helps us to solve the problem like overfitting and computational complexity. This reduction enhances the efficiency and accuracy of various used machine learning models and also leads to improved diagnostic performance.

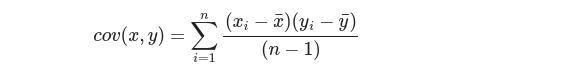
**4. Mathematics Behind Principal Component Analysis**

Lets we have two dimensions(features) dataset x and y. Since the dataset we took is 2-dimensional then we know that 2-dimensional data refers to the dataset having rows and columns, this will result in a 2x2 Covariance matrix.The Covariance matrix of these two features as follows:



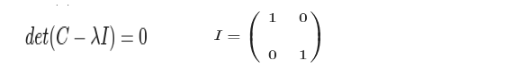
**Why we take covariance matrix is the main question:** we know covariance matrix is used to understand the relationships between the different features in the dataset. PCA aims to reduce the dimensionality of data while retaining as much variance as possible. Variance measures how much each feature varies from the mean. The covariance matrix helps to capture this variance in multiple dimensions.The covariance matrix also reveals how features are correlated with each other. If two features are highly correlated, they carry similar information.

Covariance matrix is a square matrix that measures the variations of a datapoint from the mean parameter values. Where covariance of each data point is calculated as follows:



In the next step we need to calculate the eigenvalues(λ) and eigenvectors(V) for the covariance matrix. Also we have an identity matrix(I) of the same dimension as C.

**Why we calculate the eigenvalues(λ) and eigenvectors(V)** **:** PCA uses this information to combine features into principal components.The principal components are the eigenvectors of the covariance matrix. They represent the directions of maximum variance in the given data. We know that the eigenvalues corresponding to these eigenvectors indicate the magnitude of variance in those directions. Higher eigenvalues mean more variance and therefore, more importance in the data’s structure.



Where,C is covariance matrix, λ is an eigenvalue for a matrix C , I is the identity matrix of the same dimension as C and ‘det’ is the determinant of the matrix.After solving above process we will get equation ,by solving this equation we will get different values of λ eigenvalue. For each eigenvalue(λ), a corresponding eigen-vector(V), can be found by solving the below equation.



After solving this equation we get eigen-vector(V) as below:



If we have a dataset with n features, then we have the corresponding n eigenvalues and eigenvectors. Next we form a feature vector which is a matrix of eigenvectors.



Finally we form the principal components by taking the transpose of the feature vector and multiplying it with the transpose of the scaled version of the original dataset.In this case transposing matrices ensures the correct alignment of dimensions for matrix multiplication. This alignment is crucial for accurately projecting the original data onto the principal components.



Finally we can calculate PC1 = λ1F1 + λ2F2 +...+ λnFn ,Similarly use the same formula for PC2 where values of λ1 , λ2 ,.., λn are different for PC1 , PC2 ... PCn because each principal component captures different aspects of the variance in the data.

**5. Mathematics Behind SVM**

We know that the general equation of a plane will be ax+by+cz+d=0. In this we are using a linear SVM algorithm to classify the data points. We know that support vector machines is a supervised learning technique mainly used for classification but it can also handle regression and outlier detection. The main objective of SVM is to identify the optimal hyperplane that best separates classes in an n-dimensional space. Minimize 1/2 ||W||^2 refers to the process of finding the optimal hyperplane that maximizes the margin between two classes like as below image source;

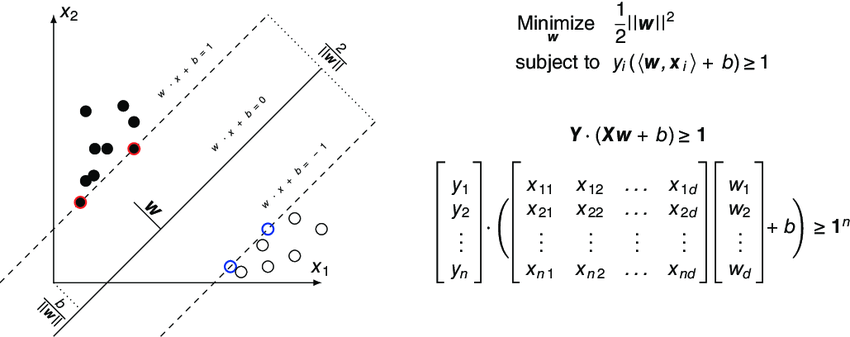
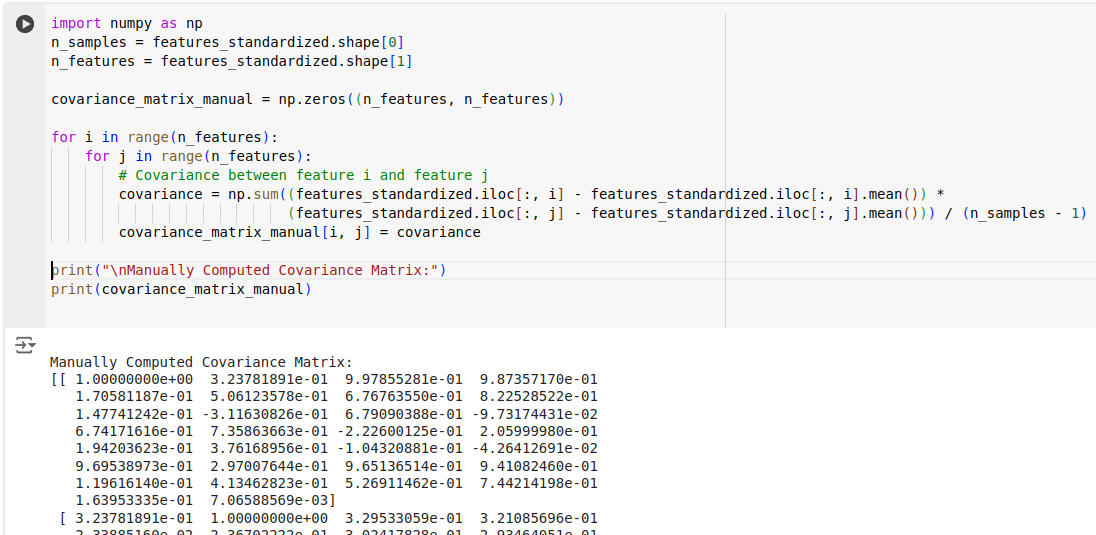


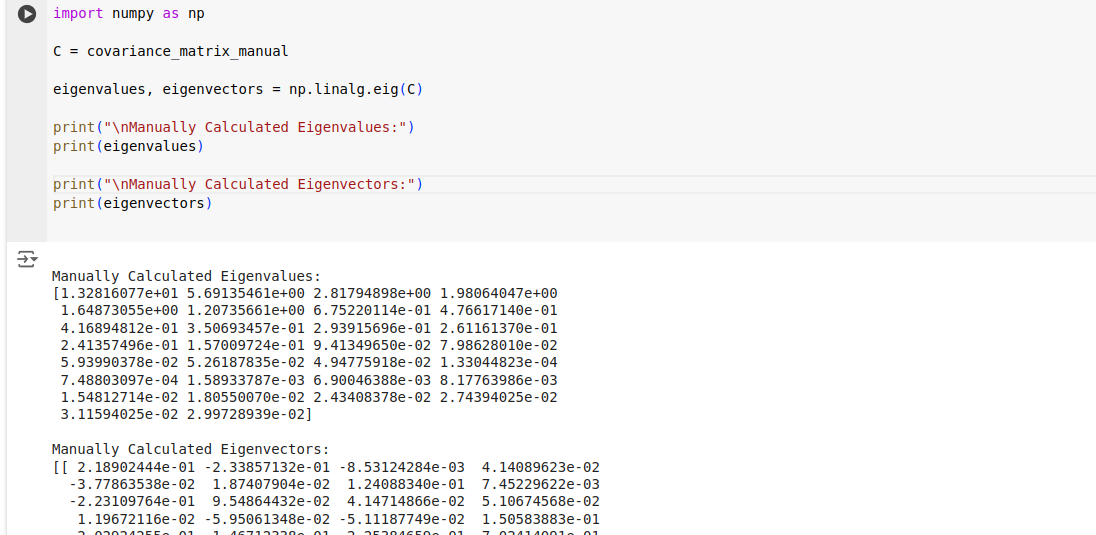
Image Source: https://www.researchgate.net

**6. Major Findings and Code Implementation in Datasets**

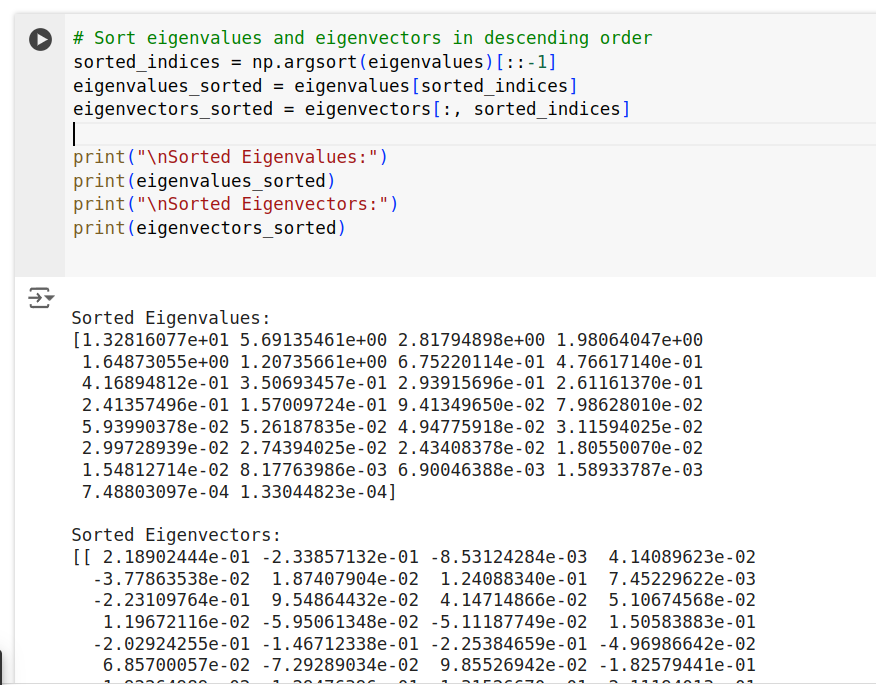
Computed Covariance Matrix between feature i and feature j as like below code;

****

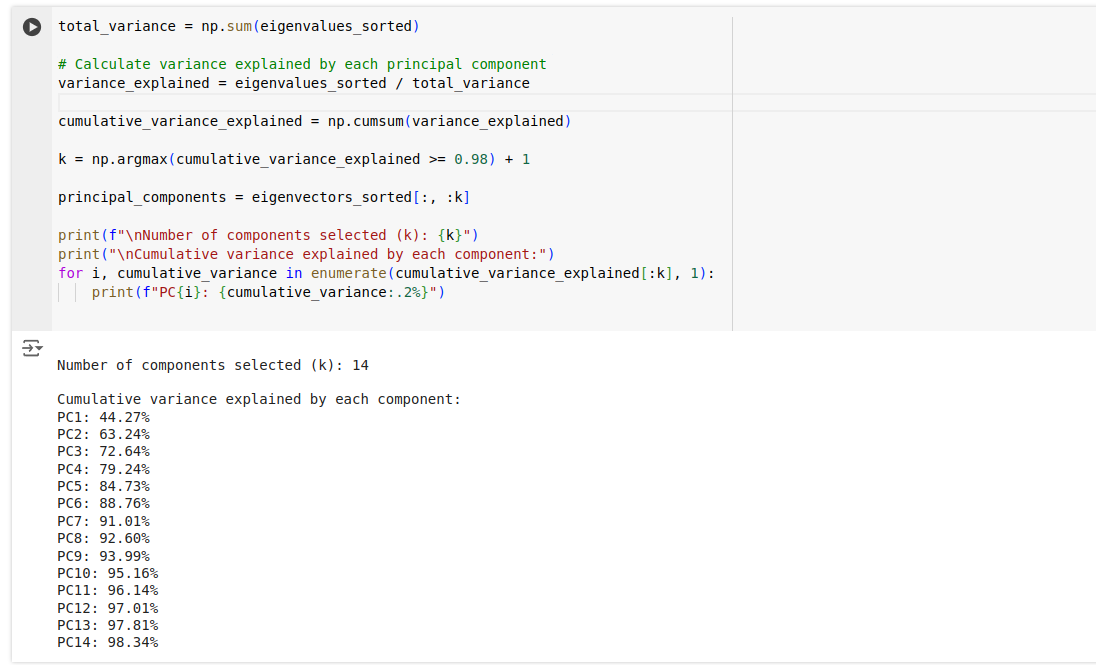
After that we need to calculate eigenvalues and eigenvectors. We know that the eigenvalues corresponding to these eigenvectors indicate the magnitude of variance in those directions calculated with the help of numpy as like below code;

****

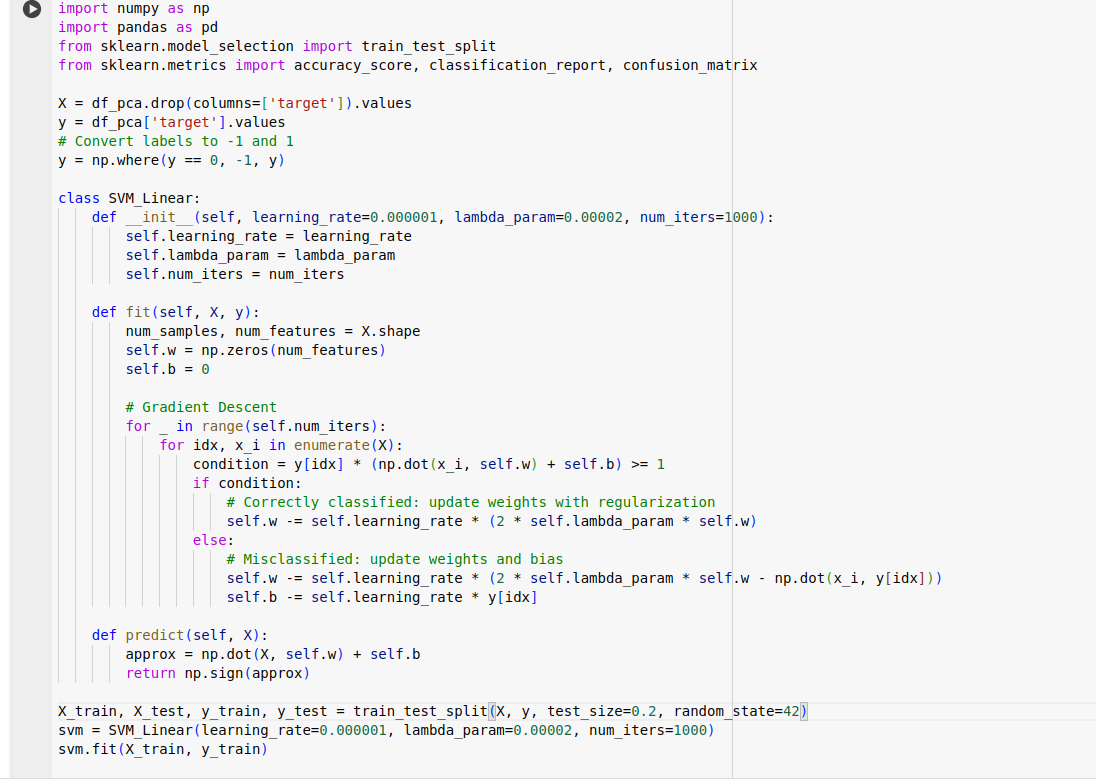
Sort eigenvalues and eigenvectors in descending order because it helps to prioritize the components that capture the most variance in the data as like below code;

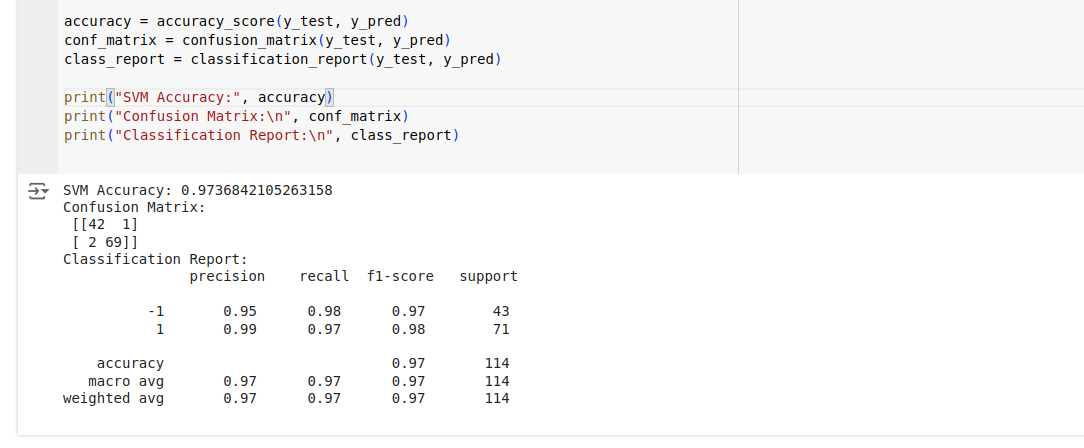
****

After that we need to calculate variance explained by each principal component and calculate cumulative variance to know how many principal components store information near about 100% , generally we take 98 to 99 % of the total variance like as below code i have implemented and reduce 30 features to 14 principal component;

****

Here we implement linear SVM. The main objective of SVM is to identify the optimal hyperplane that best separates classes in an n-dimensional space. Minimize 1/2 ||W||^2 refers to the process of finding the optimal hyperplane that maximizes the margin between two classes like as below code implementation;

****

****

#### 

After implementation of the SVM model we achieved an accuracy of 97.4%. As per the above confusion matrix which shows 42 true negatives, 1 false positive, 2 false negatives and 69 true positives. For class -1, the precision is 95% and recall is 98%, while for class 1, precision is 99% and recall is 97%. Finally after analyzing the model output I found that the model performs exceptionally well with high precision and recall across both classes.

**7. Conclusion**

The result after applying Principal Component Analysis(PCA) to breast cancer datasets effectively reduces dimensionality from 30 features to 14 principal components by taking 98 to 99 % of the total cumulative variance and enhances model performance by focusing on the most informative features. The use of a linear support vector machine (SVM) achieved a high accuracy of 97.4%, demonstrating strong classification capabilities. The model precision and recall scores indicate its effectiveness in distinguishing between cancerous and benign tumors.

**8. Reference:**

UCI Machine Learning Repository. (n.d.). *Breast cancer Wisconsin (Diagnostic) data set*.<https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>

DeCarlo, L. (2024, January 17). *Principal component analysis*. Visual Studio Magazine. <https://visualstudiomagazine.com/Articles/2024/01/17/principal-component-analysis.aspx>

*Implementing PCA from scratch*. Towards Data Science. <https://towardsdatascience.com/implementing-pca-from-scratch-fb434f1acbaa>

ResearchGate. (n.d.). *Support vector machine (SVM) formulation and matrix-vector multiplication*. Retrieved August 16, 2024, from <https://www.researchgate.net/figure/Support-vector-machine-SVM-formulation-and-matrix-vector-multiplication_fig6_320806445>