ML Project- Unsupervised and Supervised Learning

Topic: Predicting Insurance Fund Performance Using Machine Learning

Course: Advanced Topics in Machine Learning

Lecturer: Dr. Chen Hajaj
Team Members: Itamar Melnik & Tomer Sabag

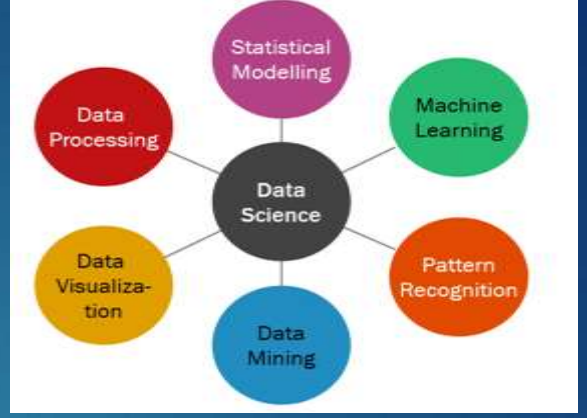
The Problem

- **Importance:** Evaluating the performance of insurance funds is critical in the financial sector, yet challenging due to the large volume of financial data.
- **Problem definition:** Investors struggle to identify high-performing funds among numerous options, which can lead to inefficient allocation of resources.
- **Objective:** Use machine learning techniques to predict the success of insurance funds and uncover patterns in fund behavior.

Project Goals

- **Main goal**: Predict and analyze the performance of insurance funds to empower stakeholders with actionable insights.
- **Sub-Goals:**
  - **Classification**: Predict fund success using supervised learning Models, achieved by evaluating the Alpha measure as a key performance metric.
  - **Clustering**: Use unsupervised learning to identify hidden patterns and groupings among insurance funds.
- **Unified Approach**: Both methods complement each other to achieve the main goal by combining predictive accuracy with deeper data-driven insights.

Methods



- **Data Processing**: Handled missing values using imputation techniques tailored to data distribution.
- **Feature Selection**: Identified and selected the most relevant features to enhance model performance.
- **Hyperparameter Tuning**: Optimized model parameters to improve accuracy and reliability.
- **Classification and Clustering**: Used classification models to predict Fund's success and clustering methods to identify patterns within the data groups.
- **Evaluation Metrics**: Evaluated models using metrics like Accuracy, Precision, Recall, F1-Score for classification, and Silhouette Score and Elbow Method for clustering.
- **Visualization**: Visualized results using bar plots to compare model evaluation metrics, scatter plots for clustering results, and heatmaps (PPS and Spearman matrices) to illustrate feature relationships and correlations.

Dataset Overview

- **Dataset Description:** The dataset, sourced from a governmental database, contains detailed financial and performance metrics for insurance funds in 2024. It includes features related to yield, management fees, market exposure, risk level and liquidity.

Feature	Description
ALPHA	Fund's excess return relative to market expectations (target variable).
AVG_ANNUAL_MANAGEMENT_FEE	Average yearly management fee
AVG_DEPOSIT_FEE	Average fee for deposits into the fund.
FOREIGN_CURRENCY_EXPOSURE	Exposure to foreign currencies
FUND_ID	Unique ID for the fund
FUND_NAME	Name of the insurance fund.
LIQUID_ASSETS_PERCENT	Percentage of liquid assets held by the fund.
MONTHLY_YIELD	Yield generated by the fund on a monthly basis.
PARENT_COMPANY_ID	Unique ID for the parent company.
PARENT_COMPANY_NAME	Name of the parent company.
REPORT_PERIOD	Period covered by the report.
SHARPE_RATIO	Risk-adjusted performance measure.
STANDARD_DEVIATION	Volatility of the fund's returns.
STOCK_MARKET_EXPOSURE	Proportion of investments in stock markets.
TOTAL_ASSETS	Total assets held by the fund.
YEAR_TO_DATE_YIELD	Year to date yield.
YIELD_TRAILING_3_YRS	Average annual yield over the past 3 years.
YIELD_TRAILING_5_YRS	Average annual yield over the past 5 years.

Clustering Algorithms:

- **K-Means**: Chosen for its simplicity and efficiency in identifying distinct fund groupings based on financial metrics.
- **Agglomerative Clustering**: Used to provide a hierarchical perspective and uncover nested relationships between funds.
- **DBSCAN**: Applied to detect outliers and explore clusters of varying shapes and densities without predefining the number of clusters.

Classification Models:

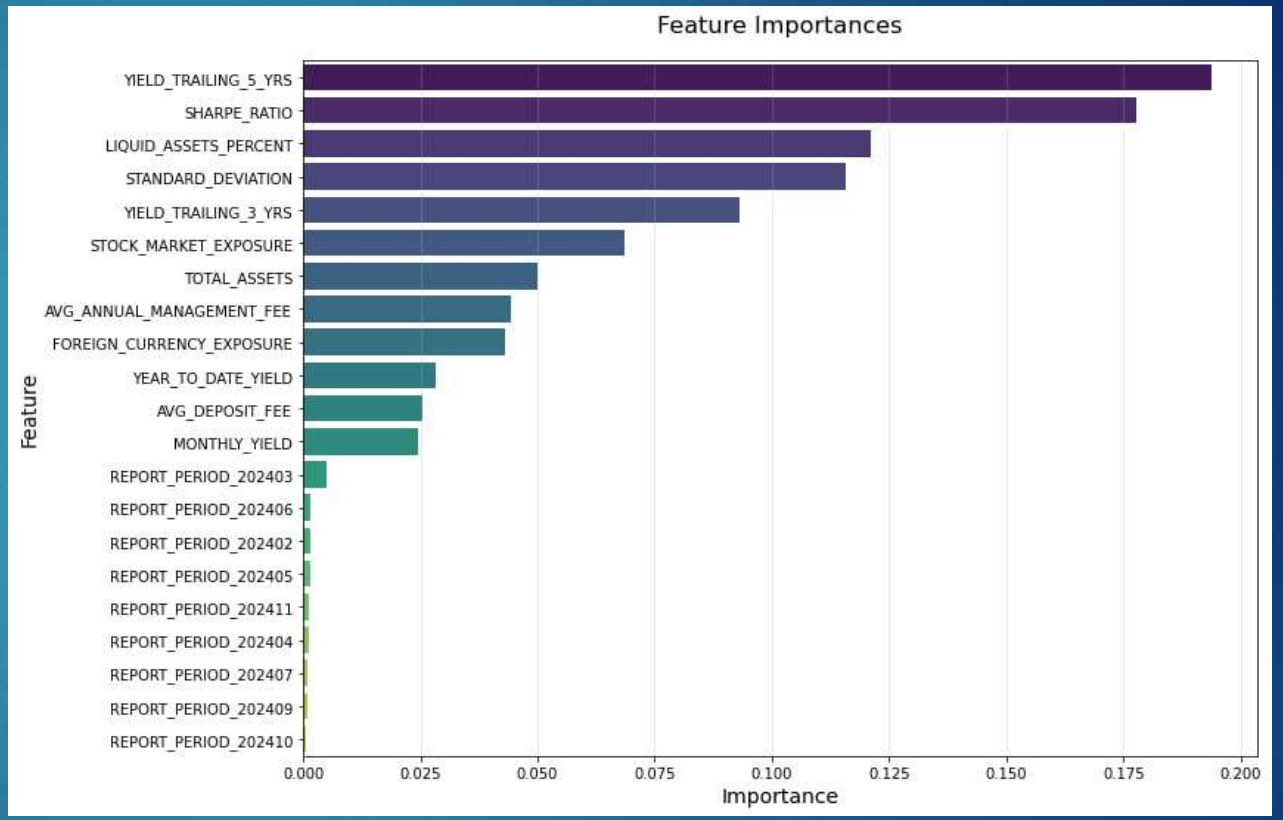
- **Random Forest**: Selected for its ensemble approach, which combines multiple decision trees to improve accuracy and robustness against overfitting.
- **Gradient Boosting**: Used for its iterative refinement process, delivering high accuracy in complex datasets.
- **XGBoost**: Chosen for its speed and ability to handle large datasets efficiently while reducing overfitting.
- **SVM**: Effective in capturing non-linear relationships and distinguishing overlapping classes.
- **k-Nearest Neighbors (kNN)**: Included for its simplicity and to evaluate how well more advanced models perform compared to a straightforward approach.

Experiments Overview

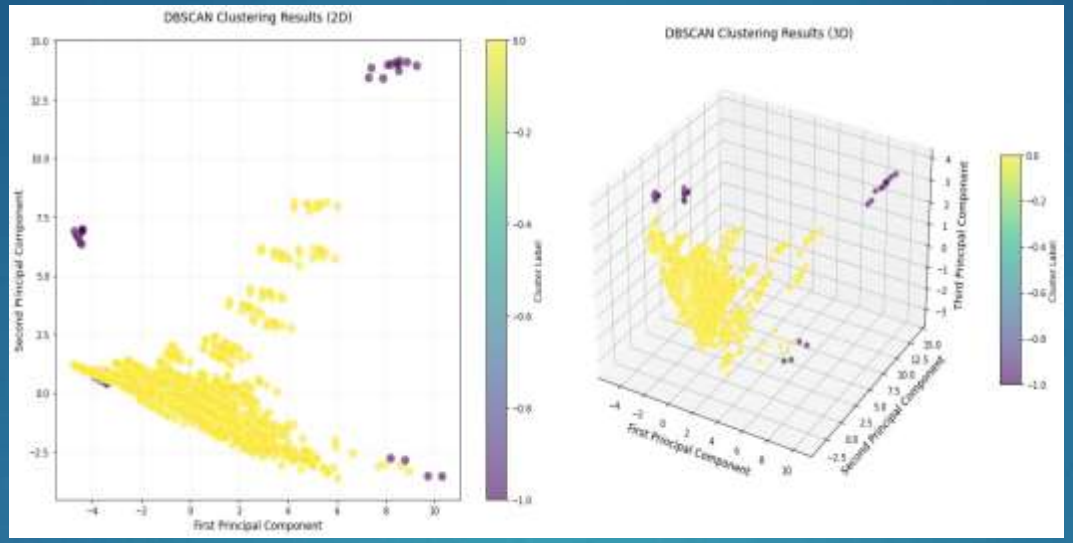
	Supervised	UnSupervised
Evaluation metrics	Accuracy, F1-Score, Precision, Recall, Cross-Validation	Silhouette Score, Elbow Method
Descripton	Models were trained to predict fund success (high/low performance) using Alpha as the target. Cross-validation ensured strenth and reliability of the models.	Clustering algorithms were used to group funds into natural clusters and identify patterns or outliers in the data.

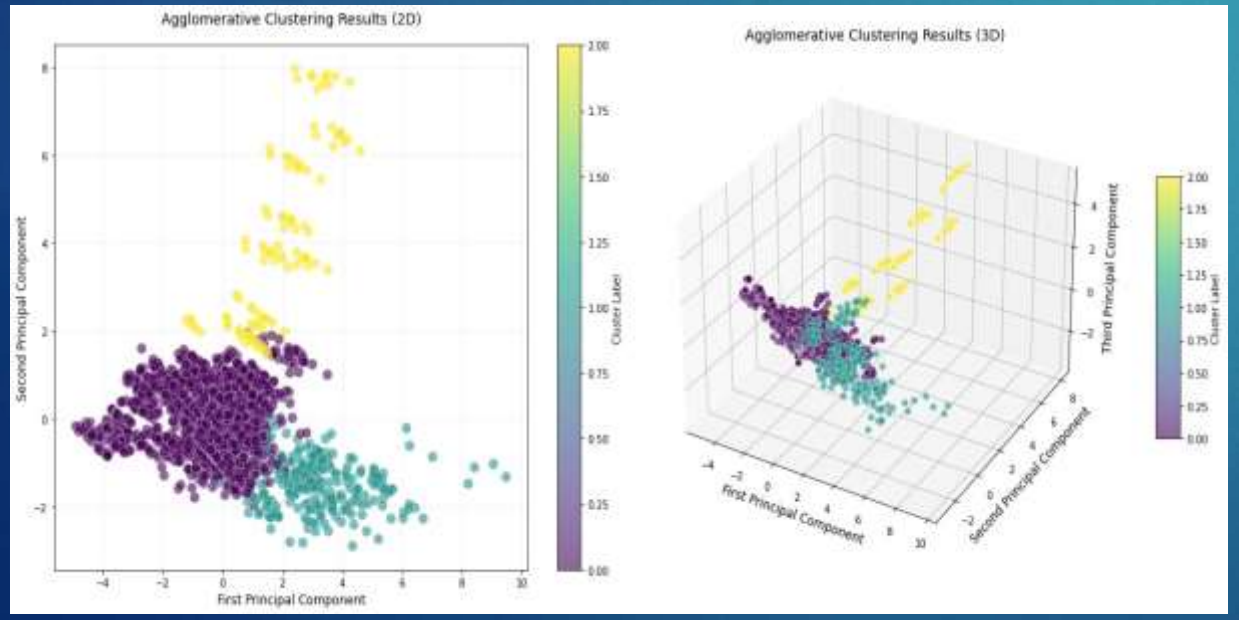
Classification Results

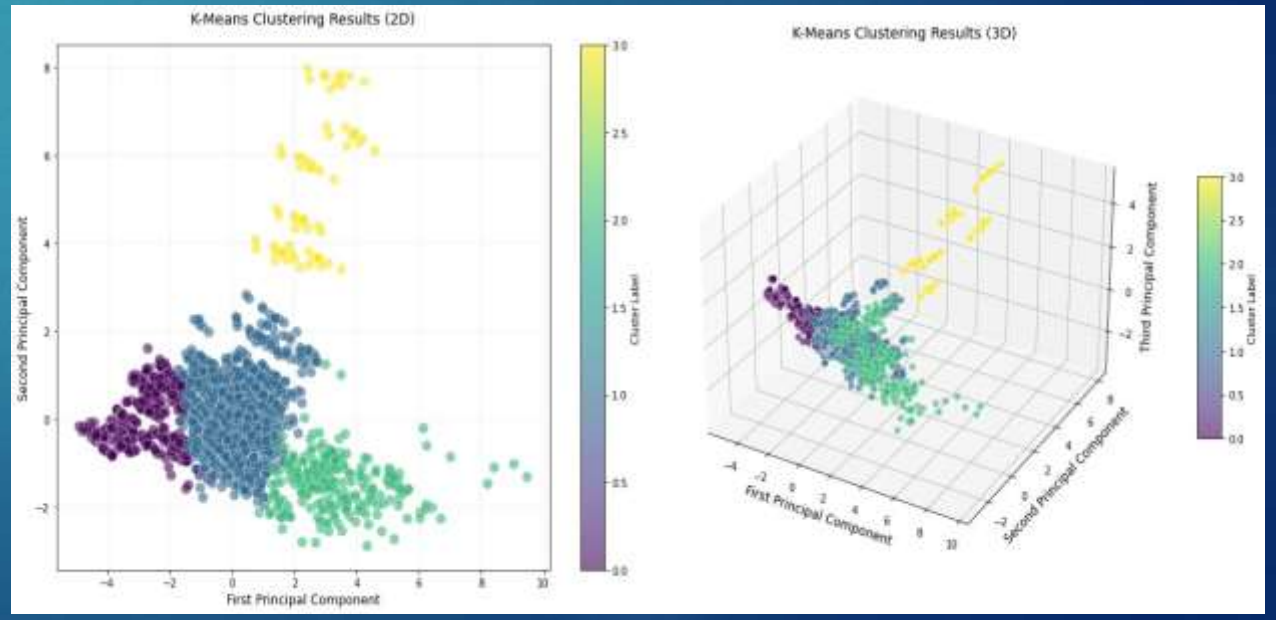
	Model	Accuracy	Precision	Recall	F1-Score
0	Random Forest	0.966292	0.968254	0.983871	0.976000
1	Gradient Boosting	0.951311	0.957672	0.973118	0.965333
2	XGBoost	0.955056	0.957895	0.978495	0.968085
3	SVM	0.943820	0.952381	0.967742	0.960000
4	KNN	0.932584	0.937500	0.967742	0.952381



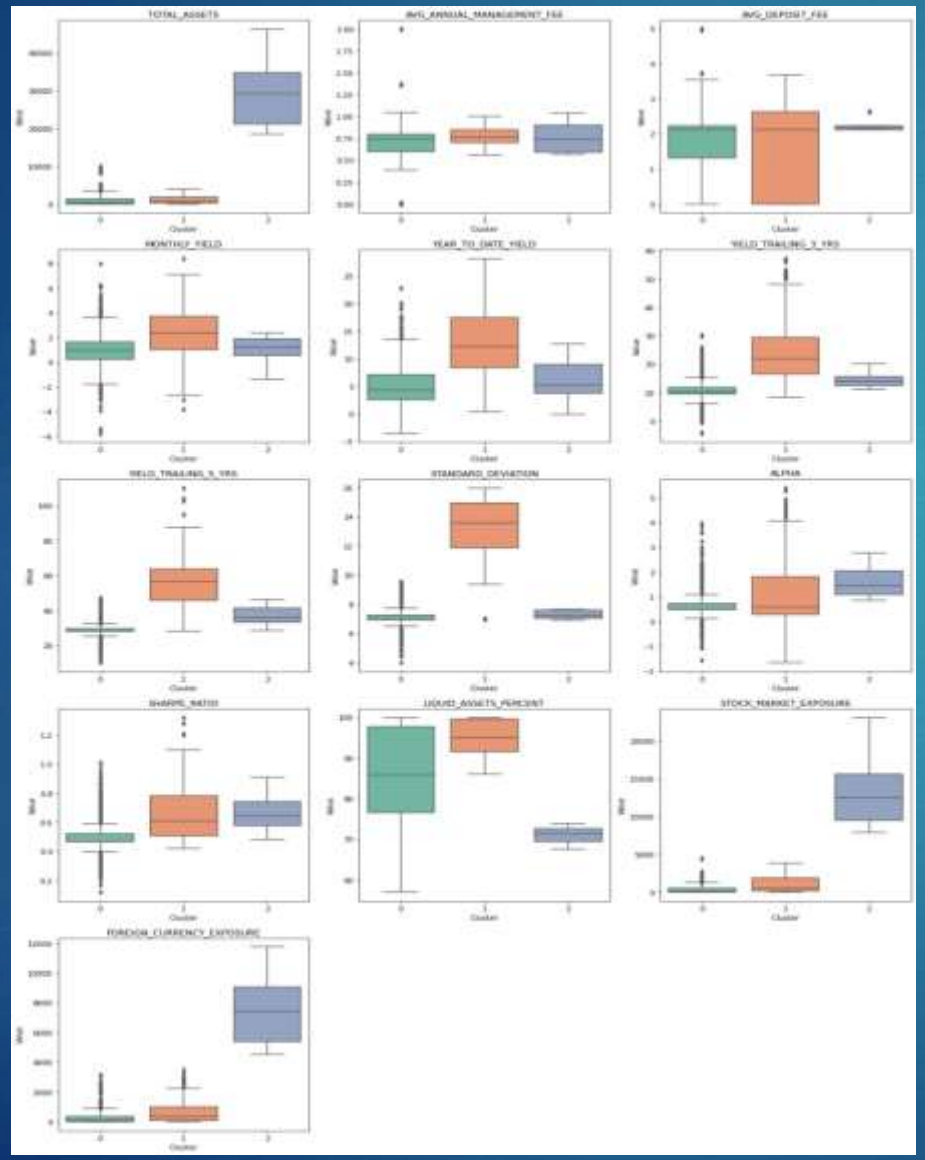
Clustering Results

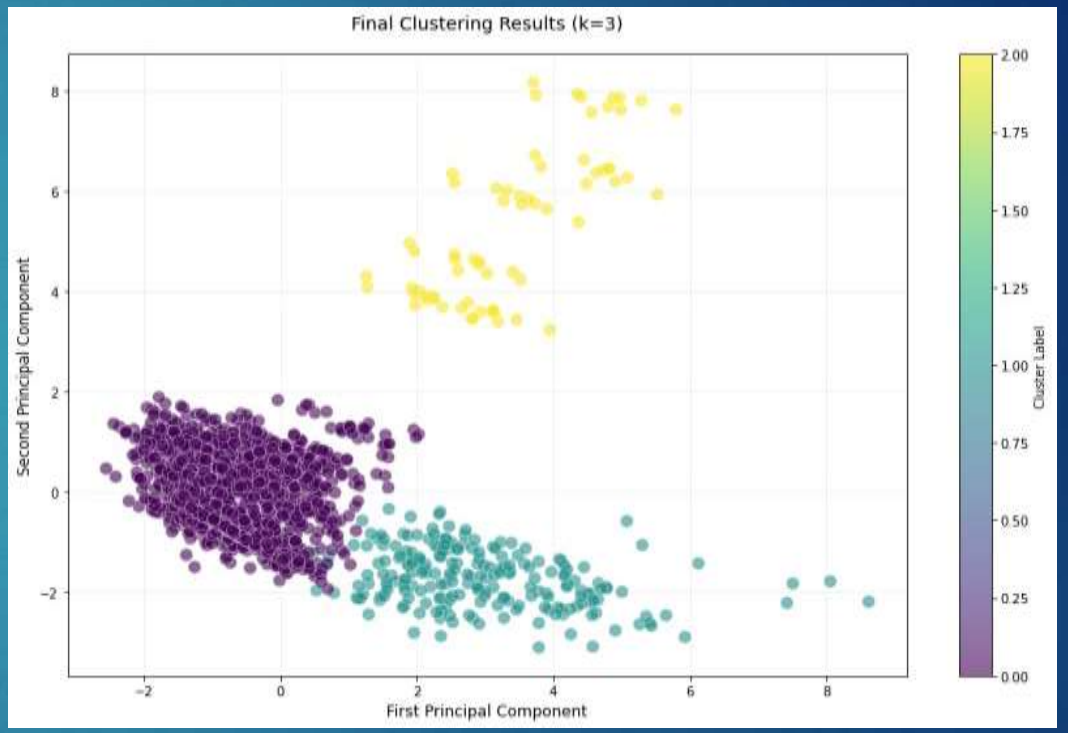






Clustering Results





Classification Models Comparison

Feature/Model	kNN	SVM	XGBoost	Gradient Boosting	Random Forest
Learning Method	Distance-based	Kernel-based	Advanced Boosting	Boosting (iterative error correction)	Bagging (ensemble of decision trees)
Non-linearity Handling	Limited	Excellent	Excellent	Excellent	Good
Robustness to Noise	Low	High	High	Moderate	High
Interpretability	High	Moderate	Low	Low	Moderate
Run Time	Fast	Slow	Fast	Moderate	Fast

Clustering Algorithms Comparison

Criterion	DBSCAN	Agglomerative Clustering	K-Means
Algorithm Type	Density-based	Hierarchical	Partition-based
Scalability	Moderate	Moderate(slow for large datasets)	Moderate (works best for medium-sized datasets)
Number of Clusters	Not required	Predefined	Predefined
Outlier Detection	High	Moderate	Low
Cluster Shape	Works with arbitrary shapes	Detects non-spherical clusters	Assumes spherical clusters
Computational Complexity	Moderate	High (due to dendrogram construction)	Low

Classification Conclusions:

- By analyzing fund performance using Random Forest, we successfully predicted fund success based on the Alpha measure, which we transformed into a binar target. The most important predictors were long-term yields (**YIELD_TRAILING_5_YRS**), risk-adjusted returns (**SHARPE_RATIO**), and liquidity (**LIQUID_ASSETS_PERCENT**), emphasizing their role in effective fund evaluation.

Clustering Conclusion:

- **Cluster 0**: Conservative and underperforming funds with low yields and minimal market exposure.
- **Cluster 1**: High-risk, high-yield funds characterized by high standard deviation, strong long-term yields, and high liquidity.
- **Cluster 2**: Well-diversified funds with significant ***TOTAL_ASSETS*** and ***STOCK_MARKET_EXPOSURE***, though not always achieving the highest yields
- **Key Insight:** We learned that high risk, as indicated by high standard deviation, and high liquidity lead to better performance compared to large asset holdings and high exposure to stocks and foreign currencies.

Unified Conclusion:

- To achieve optimal long-term performance, stakeholders should prioritize funds with strong long-term returns, high liquidity, and a well-managed risk profile. While higher risk (indicated by variability in returns) often correlates with better performance.