Machine Learning-Based Analysis of Research Policy Impacts on Academic Performance Metrics

Xingfeng Zhao, Yanjie Wang*
[1]Handan Vocational College of Science and Technology, Handan City, Hebei Province, 0560461, China
E-mail：03-01-11@163.com
*Corresponding author

Keywords: artificial intelligence, machine learning, research policy, citation influence, feature engineering

Received: Januar 28, 2025

Artificial Intelligence (AI) has significantly transformed research policy analysis by enabling data-driven decision-making and predictive analytics. Research policies play a crucial role in shaping academic and institutional research outcomes, influencing funding allocation, citation impact, and overall research productivity. This study employs Machine Learning (ML) techniques to examine the effect of policy, outcome, and control factors on research performance. The dataset, sourced from an online repository on GitHub, underwent preprocessing steps, including handling missing values and feature selection to identify the most influential variables. The study applies Linear Support Vector Classifier (LinearSVC) to predict three key research metrics: research productivity, citation impact, and innovation index. Experimental results demonstrate that LinearSVC achieved an accuracy of 97% (F1-score: 0.96, AUC-ROC: 0.62) for research productivity prediction, 94% (F1-score: 0.93, AUC-ROC: 0.93) for citation impact, and 95% (F1-score: 0.95, AUC-ROC: 0.96) for the innovation index. These findings highlight the robustness of AI-driven analysis in evaluating research policies and optimizing institutional strategies. By leveraging ML models, this study provides insights into policy effectiveness, enabling institutions to refine their research strategies for improved scientific innovation and societal advancement.

Povzetek: Raziskava z uporabo LinearSVC in strojnega učenja analizira vpliv raziskovalnih politik na produktivnost, citatni vpliv in inovacije.

1 Introduction

The advancements in Artificial Intelligence (AI) have transformed society by improving various spheres of life such as healthcare, enhancing communication, and driving economic growth. The growth in technology is very much dependent on research and development. Defining better research policies is essential for fostering innovation, defining living standards and addressing societal challenges. All governments try to define research policies that can help to allocate resources efficiently, guide research priorities, and promote collaboration across disciplines and borders. A *research policy* is a set of guidelines, principles, and frameworks that govern how research is conducted, funded, and applied within an organization, institution, or country. The policy defines areas for research which can be beneficial for society ensuring ethical concerns sharing with validate source and content of research study, and ensuring that research priorities meet the economic, social, and national goals [1]. It may also cover areas like funding allocation as the proper use of money should be ensured to solve top societal problems.

In the current era of AI, the use of data analytics is being used to carry out data in terms of three perspectives: descriptive data analytics, perspective data analytics and predictive data analytics. The data analytics include the latest technologies of exploratory data analysis as well as using machine and deep learning algorithms for prediction of the various outcomes that may help to define future policies [2]. In research, both policy and the data used are interlinked. More precisely, the research depends on the quality of its data and the policy of that research in a specific field. Research policy has a variety of mediated effects on policy over a range of time periods [3]. More academic research usually has longer-term effects that impact the presumptive worlds of policymakers, whereas commissioned research strives for more immediate, short-term effects [4].

However, it's also important to recognize how these cultures overlap, especially when it comes to the classifications of researchers and policymakers as well as career mobility between them [5]. As the research landscape grows more complicated, there is an increasing demand for policies that can adapt to new challenges, optimize the use of funding, and ensure that all research areas, including emerging and interdisciplinary fields, are adequately supported [6]. Machine learning (ML) techniques, with their ability to analyze and collect vast amounts of data for research policy as shown in fig 1 to uncover hidden patterns, offer a promising solution to address these challenges and transform the way research policies are evaluated to address the limitations of traditional research evaluation methods [7]. These techniques are already making strides in various sectors, including healthcare, education, and environmental

management, where large-scale data integration is critical for informed decision-making [8]. By identifying patterns of excellence and areas for improvement, AI methods contribute to more effective decision-making, ultimately fostering a more equitable research environment. This study primarily aligns with the Secondary Data and Field Experimentation Setups research policy methods. The analysis relies on structured datasets, utilizing machine learning models to assess policy impacts, making it a clear example of secondary data usage. Additionally, the study involves an experimental setup where various classifiers and feature selection techniques are applied to optimize model performance in policy evaluation. This aligns with field experimentation methodologies, as the research not only analyzes existing data but also tests different computational approaches to improve predictive accuracy and derive actionable insights for policymakers.

1.1 Research question

The study's objectives and methodology are used to define better defined research questions. Based on the above, by investigating the role of artificial intelligence (AI) and machine learning (ML) techniques in assessing the effect of research policies, this study examines their effects on the academic and institutional outcomes. Some of the research questions which the study will attempt to answer are:

1) How different research policy variables (for instance funding budget, policy duration etc.) affect research productivity, citation impact and innovation index?

Techniques of machine learning models are used to examine the correlation between the outcomes of research and the policy variables. The results show that higher amounts of funding and longer time in place of the policy have a positive influence on research productivity and citation impact. Furthermore, STEM fields policies have a higher innovation index than the social science one.

2) According to which machine learning models do research policies and their effectiveness evaluate most accurately predictions?

Five ML models were evaluated (LinearSVC, GaussianNB, KNN, RidgeClassifierCV, and ExtraTreeClassifier) and it was found that LinearSVC had the highest accuracy (97% by means of research productivity, 94% by means of citation impact, and 95% by means of innovation index). The reason for the superior performance of LinearSVC is that it can deal with high dimensional data as well as separate policy influence well.

3) What are the key influential factors which drive the research performance and how they vary across institutions or regions of different countries.

The study first uses feature selection techniques which rank the factors funding amount, researcher experience level and collaboration level as the most important drivers of the research performance. Private institutions are more oriented in terms of innovation indices, whereas research-intensive regions are more productive and more impactful from the perspective of citation.

1.2 Research contributions

In this research study, our main aim is to explore various features that act as independent variables to check their impact on the dependent which are directly linked with research outcomes including research productivity, citation impact and research innovations. The research productivity refers to the number of research publication during the research project period; citation impact refers to the average number of citations per paper in the period and innovation is associated with the innovations, products or technology that resulted based on the research projects. For descriptive analytics, exploratory data analysis and data visualization methods have been used. For prescriptive analytics and to predict the outcome variables, we have taken various relevant features and the statistical feature ranking methods such as Gain Ratio and Information Gain have been used.

Figure 1: Methods to collect data for research policies

For predictive analytics, the machine learning algorithms such as Linear Support Vector Classifier also known as LinearSVC, Gaussian Naïve Bayes, Ridge Classifier, ExtraTreeClasifier and k-nearest neighbors (kNN) have been applied. The data is prepared, preprocessed and split using holdout methods. The evaluation has been carried out using standard classification evaluation measures of accuracy, f-measures, and Receiver Operating Characteristics (ROC) Curves. The main Contribution of this study as follows:

- Analysis of the influence of various factors based on policy, outcome and control variables, providing insights for policy impacts on academic research performance.
- The application of diverse ML models to classify and predict research outcomes such as Research Productivity, Citation Impact, and Innovation Index.

- Features ranking by employing various feature engineering methods to identify top key factors in driving research outcomes.
- Proposal of a robust framework for predicting research outcomes metrics, with highest accuracy of 97% using SVM classifier for criteria of predicting high, moderate and low impact of research productivity and its influential impact in terms of citations as well.

The rest of the paper organization as follows: Section 2 presents the background knowledge of relevant literature in field of academic research outcomes. Section 3 shows the comprehensive details of applied methodology including experimental setup. Section 4 shares the analysis of results along with discussion. Section 5 provides the summary of paper in conclusion form with future directions.

2 Related work

The recent advancements in research at global level have attracted the researchers to focus on various research policies and analyzing their impact. An important discussion in research policy studies & their impact in recent years has focused on whether similarities or differences between national research policies, as well as the extent to which these policies have changed in a similar way over the recent few decades. Summary of existing studies shown in table I.

Factors influencing research performance have been categorized into three primary classifications: individual, institutional, and research self-efficacy. Individual factors encompass personal attributes of faculty members, such as age, gender, educational background, academic rank, workload, field of specialization, and research experience. These characteristics shape how researchers engage with their academic roles and contribute to their scientific output [9]. Among these factors, the influence of gender, age, and academic position on publication rates has been examined by suggesting that academic position significantly outweighs such factors in its impact on research productivity. Seniority within the academic hierarchy often correlates with greater access to resources, collaborations, and opportunities for publication [10]. Research performance has been demonstrated as a crucial factor among universities linked to research productivity, emphasizing the importance of research training, sufficient salaries, and efficient work habits. These elements collectively highlight the systemic and institutional supports necessary for fostering high domain experts to follow research policies [11]. A comparative analysis of higher education institutions across six countries highlighted a notable survey in faculty research productivity within the Arab countries. This region's relatively low research output was attributed to a combination of institutional, cultural, and systemic factors, underlying the importance of tailored strategies to address regional challenges and promote academic excellence [12]. This collectively underscores the multifaceted nature of factors influencing research performance and offers pathways for targeted interventions.

Another study highlighted the integration of ML for policy-making, demonstrating how ML models, like doubly robust estimators, mitigate biases when estimating treatment effects in large-scale policy scenarios [13]. Another emphasized for predicting societal outcomes and showcased its potential in assessing education policy impacts, where non-linear predictors improved forecasting accuracy by over 30% compared to traditional econometric models [14]. Several studies explored the role of research policies and funding in shaping academic outputs. Analysis shows that funding mechanisms influences citation-based performance and research productivity, which directly impacts the diffusion of new scientific knowledge in society. Specifically, models linking funding to citation growth reveal that well-supported research is more influential factor for research productivity [15]. Citation impact evaluation methods highlighted the relationship between publication metrics and academic impact. Using fi-score, as an unbiased measure for self-citations and external influences, complement traditional metrics like the H-index for evaluation of academic productivity [16]. Societal research evaluation extends beyond academia, assessing societal benefits such as contributions to public health and environmental sustainability [17]. However, further research is needed to address the inherent risks of bias, transparency, and reproducibility while advancing theoretical and methodological approaches to ensure equitable outcomes for all researchers.

Table 1: Summary of prior work

Sr. No	Ref	Year	Model	Results (%)	Dataset	Features
1	[9]	2021	SVM, AI-driven Governance System	Acc: 70	Survey data from American research universities.	Impact factor, grants
2	[11]	2021	LR, DT, SVC	Acc: 80	Interviews, Meetings	Average count, mean of citation count
3	[12]	2022	Bibliometric analysis, SPSS	Pre: 79	Arab literature (documents published between 2006–2015)	Citation counts before and after the Arab Spring
4	[13]	2022	XGB, RF, SVM	Acc: 85	scholarly research articles on sustainable finance	Citation Index Impact Factors

5	[14]	2020	SVM, KNN	F1: 73	Research publications funded Canadian researchers (2000–2018).	Textual Features
6	[15]	2022	SVM, NB	Acc: 79	Academic history of Brazilian researchers	Case studies and Surveys
7	[17]	2020	CNN, SVM	Pre: 77	Policies and practices data in environmental and sustainability education	Institution, Funding, Collaborative Index
8	[18]	2024	KNN, Adaboost	Acc: 89	Transactional data	Citation Impact, Funding, Impact Factor

3 Proposed research methodology

In this section, proposed research methodology, as shown in fig 2, has been explored by applying feature engineering techniques and various ML models training apply on dataset, evaluated using standard parameters of classification measures such as accuracy, F1-Score and AUC-ROC.

We develop a structured bias assessment framework to better guarantee fairness and see potential biases in our study. To ascertain if bias exists, one conducts Exploratory Data Analysis (EDA) such that they examine dataset composition to reveal composition, and whether there's any underrepresentation of events. Fairness metrics as accuracy, precision, recall, F1–score, and AUC-ROC is used for bias diagnosis and enables a quantitative measure of disparities in model performance. Hence, we optimize the model fairness by setting right values of hyperparameters and balanced learning across different groups to mitigate bias. Moreover, Information Gain and Gain Ratio are used to analyze feature contribution in order to enhance interpretability and model transparency by confirming if a feature has an impact on model decision. Fairness evaluation dashboards and MLOps continue to monitor bias continuously, so as fairness can be tracked over time and necessary adjustments can be taken to keep the model reliable and fair. Use of this structured approach makes our study more robust, and introduces it to the practices of fairness aware machine learning.

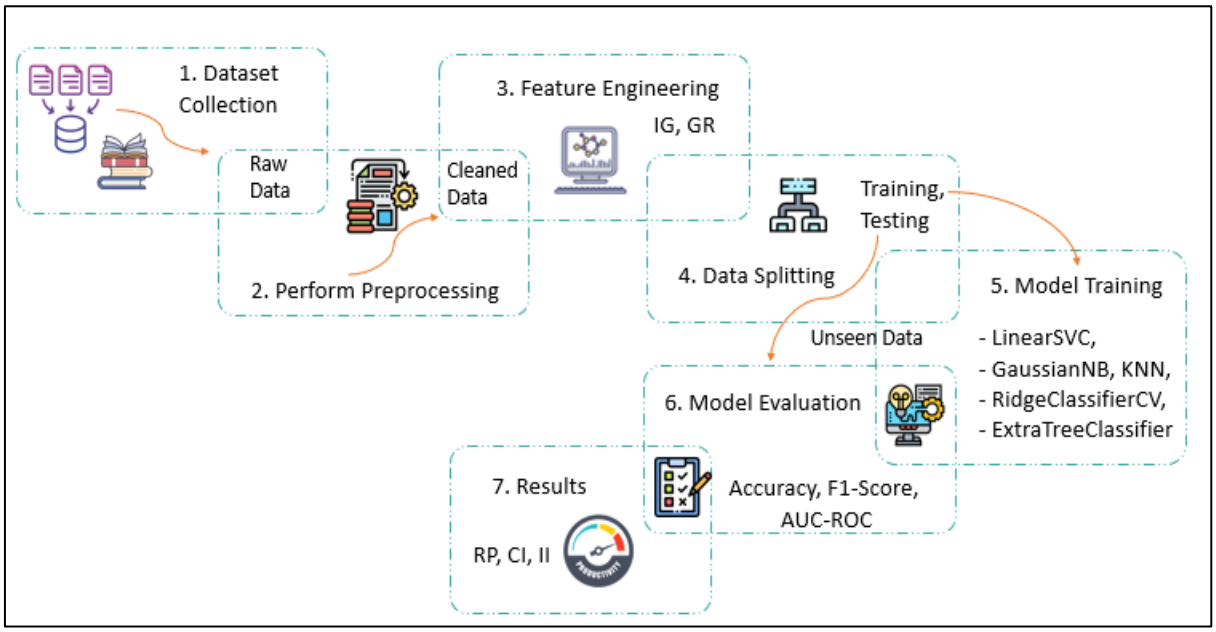

Figure 2: Framework showing steps of of proposed research methodology

3.1 Preprocessing

Preprocessing is a critical step in the data quality as well as improving the model's result's performance. Finally in this study, Min-Max Scaling has been applied to normalize feature values within a predefined range so that there is consistency and there are no biases in the training of the model. Information Gain and Gain Ratio were used for feature selection that will associate and retain the most important attributes to reduce the dimensionality and

computational complexity while keeping the key data patterns. The tabular dataset, data cleaning was done to remove inconsistencies and outliers and thus to have a structured and reliable dataset that can be analyzed. Additionally, the missing values were taken care of using appropriate imputation techniques to avoid information loss while keeping the integrity of the dataset. Together, these preprocessing techniques strengthen the robustness of the dataset, making the reliability of subsequent modeling marches and the probability of a reliable overall accuracy and generalizability of the study's findings.

3.2 Feature engineering

Feature engineering is the process of transforming, selecting and creating informative attribute from raw data to improve accuracy, enhance interpretability of machine learning models. Information Gain (IG) is used for data analysis, measuring the reduction in uncertainty in machine learning models. First calculate uncertainly variables, split data into subset, calculate entropy of variables and then calculate information gain of each feature. Feature extraction using information gain to rank feature and select top ranked features. This technique is effective in handling high dimensional data and easy to interpret. Furthermore, Gain Ratio (GR), used for data analysis, handles multi values attributes under splitting ratio of 0 and 1, to improve feature selection stability. It ensures that selection is not based only on those features that have a high number of distinct values and provides a fairer way to carry out feature selection for cases where the decision tree algorithm will be used [18]. The method is particularly suitable for the use with the large data that can have various features.

$$GR = Information\ gain/Split\ information \quad (1)$$

GR and IG are necessary metrics to use with decision trees for selecting attributes. When splitting a dataset on a particular attribute, Information Gain does whatever measurement is defined as Information Gain, which is usually what outcome expected, namely how much the entropy of the resulting datasets decreases. It also finds the most informative features for classification. The entropy of a dataset D is given by equation 2:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (2)$$

p_i is the probability that an arbitrary tuple in D belongs to class C_i, $|C_i, D|/|D|)$ is the estimated probability.

After splitting the dataset D using attribute A into v partitions D_i, the entropy for the partitions is given by equation 3:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (3)$$

To address the attrition issue caused by having many values for some attributes, Information Gain is normalized by Gain Ratio. It is calculated as in 4:

$$Gain(A) = Info(D) - Info_A(D) \quad (4)$$

However, Information Gain tends to favor attributes with many unique values, which may not always be optimal. To counteract this bias, the Gain Ratio (GR) is introduced, defined as in 5:

$$GR(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (5)$$

where $SplitInfo(A)$ computed as using equation 6.

$$SplitInfo(A) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2\left(\frac{|D_j|}{|D|}\right) \quad (6)$$

These equations ensure a robust and unbiased feature selection process, improving the interpretability and efficiency of classification models.

Information Gain (IG) and Gain Ratio (GR) were used for feature selection to improve the models toward the predictive capability. The measurements of entropy reduction from splitting data on specific attribute combination is what IG can do best, and GR is able to normalize a metric for dealing with values of a categorical variable with more than one. Computational limitations and the concerns of interpretability in policy-based decision making implied that other methods like Recursive Feature Elimination (RFE) or Mutual Information were not used. In the study, a holdout validation was employed with the dataset split into 80% training, 20% testing for generalization performance assessment. This supports the use of AI driven methodologies in research policy evaluation to create data for policymakers to implement funding strategies more efficiently as well as institutional policies.

3.3 AI-based applied models

To evaluate the effectiveness of various ML techniques in predicting research outcomes, several classification models were applied to explore the performance of models, highlighting their strengths and ability power for classifying research-related outcomes that helpful in decision making for optimizing policy and research innovation.

3.3.1 Gaussian Naïve Bayes (GNB)

GNB is a features independent classifiers use for probabilistic based theorem and effective for high dimensional data. For computing, this uses the following formula that shows in equation (7) and (8).

$$P(y|x) = \frac{P(x|y) * P(y)}{P(x)} \quad (7)$$

$$P(x|y) = \frac{1}{(\sqrt{(2\pi\sigma\text{^}2)})} * \frac{exp(-((x-\mu)\text{^}2)}{(2\sigma\text{^}2))} \quad (8)$$

3.3.2 K-Nearest Neighbors (k-NN)

KNN is a lazy method of learning that makes prediction based on nearest neighbors use for non-linear relationship and handles categorical features, while K-NN is a not parametric classifying technique. A data point gets assigned a class that reflects the majority class of its k-nearest neighbors. The KNN compute show in (9).

$$d(m,n) = \sqrt{\sum_{i=1}^{k}(mi - ni)2} \quad (9)$$

3.3.3 Extra tree classifier

An approach to ensemble learning called Extra Trees (Extremely Randomized Trees) mixes predictions from several unpruned decision trees. This model is effective for

nonlinear relationship and fast training model. It leverages random splits for features, in opposition to Random Forests. Based on variants of Gini index equations, the data points are split using the equation that achieves the highest value, ensuring optimal decision tree construction. The equation of extra tree classifier shows in (10).

$$\Delta G = G(D) - (\frac{|D|}{|DL|}G(DL) + \frac{|D|}{|DR|}G(DR)) \quad (10)$$

3.3.4 Linear SVC

A linear classifier called Linear SVC produces the hyper plane that reduces the distance between classes [19]. This model is fundamental in ML and used for handling high dimensional data and robust noise. Linear SVC objective function show in equation (11) and (12).

$$y = w\hat{}T\ x + b \quad (11)$$

Subjected to.

$$L(w,b) = \sum(max(0, 1 - y\_i(w\hat{}T\ x\_i + b))) \quad (12)$$

3.3.5 Ridge classifier CV

To prevent over fitting, the Ridge Classifier regulates a linear model. By separating data into training and testing sets, cross validation analyzes model performance. This model is used for reducing over fitting, handling high dimensional data and imbalanced data. It penalizes the magnitude of the coefficients with the goal to minimize the squared error. Ridge classifier objective function show in equation (13).

$$L(w,b) = \sum(y\_i - (w\hat{}T\ x\_i + b))\hat{}2 + \alpha * ||w||\hat{}2 \quad (13)$$

3.4 Hyperparameter settings

A holdout validation approach was applied, splitting the dataset into 80% training and 20% testing to evaluate the model's generalization ability. The models were trained using LinearSVC, GaussianNB, KNN, RidgeClassifierCV, and ExtraTreeClassifier, with hyperparameter tuning performed. The hyperparameter tuning process was conducted on LinearSVC, the best-performing model, to optimize its parameters. The following table II summarizes the tuned hyperparameters:

TABLE I. HYPERPARAMETER SETTINGS

Parameter	Values Tested	Optimal Value
penalty	l1, l2	l2
loss	hinge, squared_hinge	squared_hinge
C (Regularization)	0.001, 0.01, 0.1, 1, 10	1
max_iter	500, 1000, 5000	1000

These settings helped LinearSVC achieve the highest accuracy for predicting research productivity. Table III defines the variable description used in methodological equations for deeper understanding Performance Evaluation Measures.

4 Experimental setup

In the experimental setup section, discuss dataset, the process and setting used to carry out the experiments and evaluate the models for impact of research policy.

4.1 Dataset

For this study, the dataset was sourced from an open-source repository on GitHub so that it can be as transparent and reproducible as possible. The file consists of 4,000 entries with policy, outcome and control variables used for policy evaluation. Preprocessing was carried out on the dataset by handling the missing values in the numerical features using mean imputation and the categorical features using mode imputation. Policy impact, funding allocation and institutional effectiveness are preferred factors considered in the dataset, and therefore it serves as a solid foundation for analysis regarding the effect of research policies on productivity and innovation.

For data analysis, the dataset is divided into three categories policy variables, outcome and control variables. The policy variables also consist of five columns include funding amount in dollar, policy duration in year, target audience that category into three values universities, private research centers and public research institutions, priority areas categorically into STEM, health and social sciences and policy types that divided into three categories include grant, tax incentive and loans. The second categories are outcomes that consists of four columns include research productivity that tell about numbers of papers, citations impact shows the average citations per paper, innovation index includes, collaboration level and research quality score, region that consists of diverse regions including Asia, Africa, Europe, North America and Oceania with institution type that is private and public, researcher experience level consist of senior and junior.

4.2 Performance evaluation measures

Performance evaluation matrix is used for comparing performance of different models, model improvement and comparison with baselines. Use multiple metrics for comprehensive evaluation such as accuracy, f1-score, and AUC-ROC [20].

4.2.1 Accuracy

It shows the proportion of cases that are accurately expected and Proportion of accurately expected instances % of total instances, as in eq (14).

$$Accuracy = \frac{(TP+TN)}{TN+FN+FP+TP} \quad (14)$$

4.2.2 F1-score

The F1-Score manages the trade-off between precision and recall by taking the harmonic mean of both, as in eq (15). When the distribution of class is unbalanced, the F1-Score is especially helpful.

$$F - score = 2 * \frac{(\frac{TP}{(TP+FP)} * \frac{TP}{(TP+FN)})}{(\frac{TP}{(TP+FP)} + \frac{TP}{(TP+FN)})} \quad (15)$$

4.2.3 AUC-ROC

AUC-ROC measures the ability of a model to predict and plot true positive rate against the false positive rate and used to robust class imbalance, comparability of different models.

Table 2: All symbols with abbreviations

Symbols	Abbreviations
μ	mean
$(\sigma)^2$	variance
x	feature value
y	class label
m	Query point
n	A neighboring point
k	Number of features
$G(D)$	Impurity measure.
$DL\ and\ DR$	Left and right subsets after the split
$\|D\|$	Total number of samples.
C	regularization parameter
w	weight vector
b	bias term
TP	True positive
TN	True negative
FP	False positive
FN	False negative

All the above performance evaluation metrics are essential for evaluating the performance of classification models, especially in scenarios where class distribution and the cost of errors vary.

5 Results analysis

In this section, we examined the results through feature engineering techniques and machine learning models to analyze research policies impact for the purpose of classifying the various research outcomes based on the selected features.

5.1 Exploratory data analysis

The correlation matrix as shown in fig 3 provides a visual representation of the relationships between numerical variables in the synthetic dataset. The figure presents the correlation matrix that shows the insightful analysis of the relationships between some research policy factors. The heatmap shows — visually — correlations coefficients between the variables, where -1 to 1 is the value and red shades represents the highest positive correlations, and the blue shades is highest negative or negative correlations. A couple of observations of interest is the 'strong positive correlation' (0.58) between Policy Duration and Citation Impact, that longer interventions can facilitate higher citation impact. Research Productivity is also moderately related to Funding Amount (0.48) and Citation Impact (0.25), which implies that funding of some kind is essential to achieving high research output. Curiously, the Collaboration Level and Research Productivity have a moderate positive correlation (0.48) revealing the significance of the collaborative nature of academia. Priority Area, however, has weak or negligible correlations with almost all the variables indicating that it has relatively poor direct impact on most of the research success metrics. Thus, findings highlight the role of funding allocation, collaboration and policy longevity in determining whether an uncommitted fund dedicated to research has an impact and promotes the effectiveness of research. Overall, the correlation matrix reveals several notable relationships between the variables in the synthetic dataset by defining moderate positive and weak negative correlations.

The pair plot shown in fig 4 provides a comprehensive visual overview of the relationships between numerical variables in the synthetic dataset. The dataset displays a wide range of funding amounts, from approximately 0 to 20 million, which suggests that the dataset is not artificially constrained and encompasses a variety of funding scenarios. Pair plot visualization shows all relationships among various policies and outcome variable over several regions. Individual variable distribution has also been shown in the diagonal elements, with patterns such as skewness in funding amount and collaboration level, which indicate that these variables may need to be normalized or transform. The scatter plots demonstrate the relationships between various variables, including funding amount and research productivity as well as citation impact, and have positive correlations. This is in line with the existing literature, which indicates that research funds greater publication output and research citation impact. However, their relationships with funding and innovation index, besides collaboration level and research quality score, exhibit more dispersed distribution, which suggests the influence on the result beside funding. In general, the figure summarizes what patterns are to be looked for in future regression or classification analysis to quantify the efficacy of policy variables in influencing research performance.

As shown in fig 5, the boxplots that are presented display the distribution of numerous research-related variables including the amount of funding, the duration of the policy, the productivity of research, the impact of the citation, the success rate of funding, the innovation index, collaboration level, and research quality score. The distribution of the funding amount is very high skewed with a few very large outliers, indicating that while most of the research projects receive small amount of funds, some projects are funded by few times more compared to most of them. Policy duration is close to a normal distribution— it means that it is consistent in policy timelines. The compact IQR for both research productivity and citation impact indicate that most values fall in a narrow range, but outliers indicate a variation of productivity across researchers. As you can see it is very skewed meaning the funding success rate can be very low while a percentage makes it through to pass. Looking at the innovation index, it shows a greater spread meaning innovation levels of different research outputs vary. Research projects at

collaboration level show a right skewed distribution, i.e. there are research projects with extensive collaboration, a larger number of low collaborations. The last one involves research quality scores, which show a quite narrow distribution with some extreme values, meaning that most research is of a quality appropriate to the benchmark, but some projects excel or regress far more than others. However, these insights are useful when it comes to optimizing the policy, as they assist in discerning the impact of funding, collaboration, and duration of the policy on research productivity and quality.

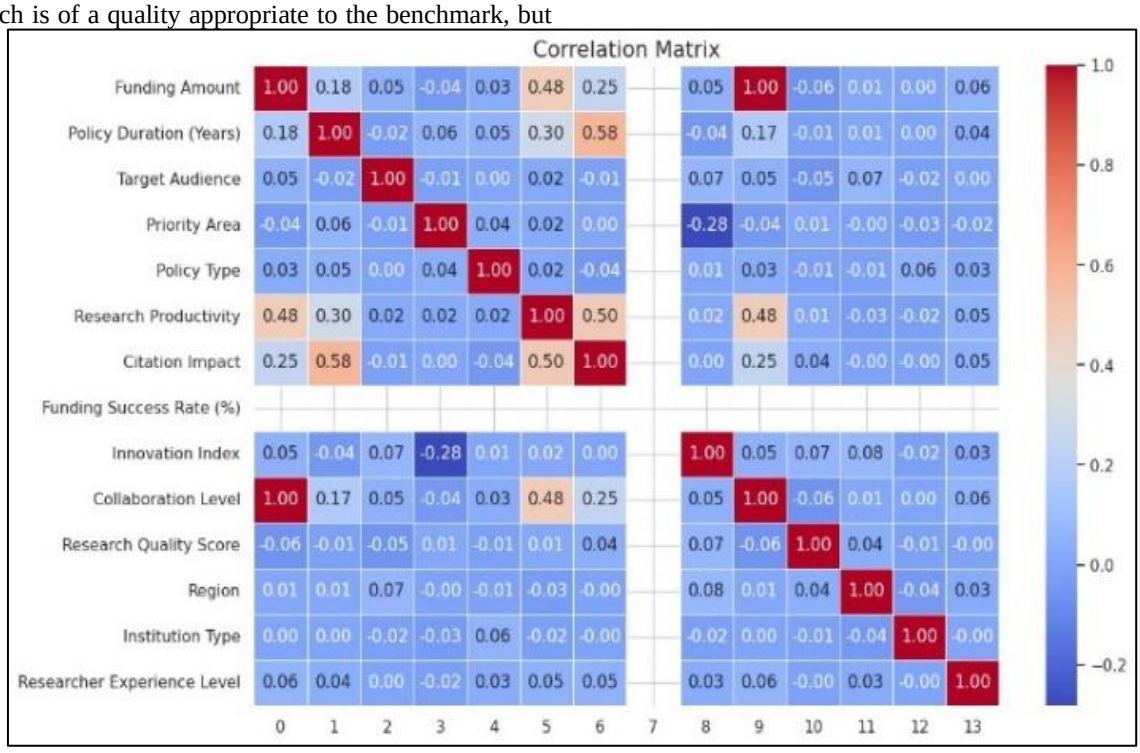

Figure 3: Analysis of correlation matrix for variable relationship

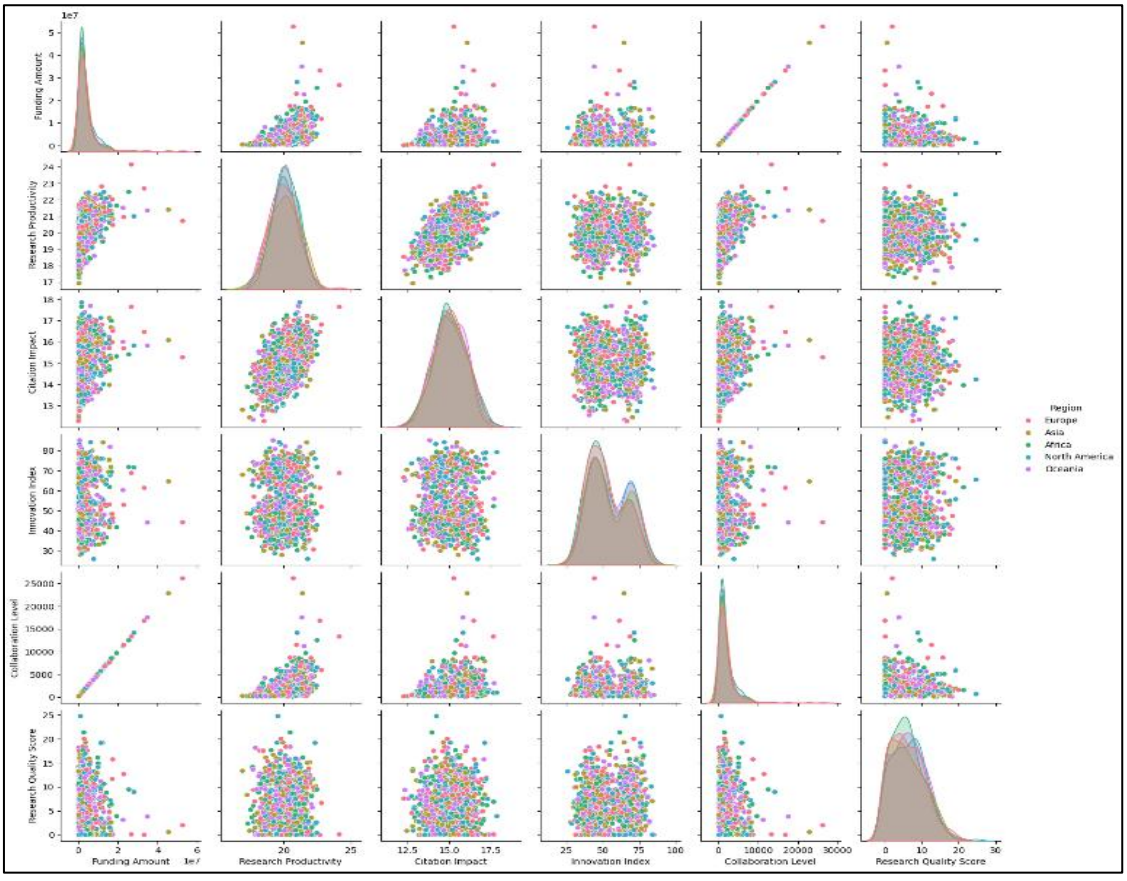

Figure 4: Analysis of dataset using pair plot visualization

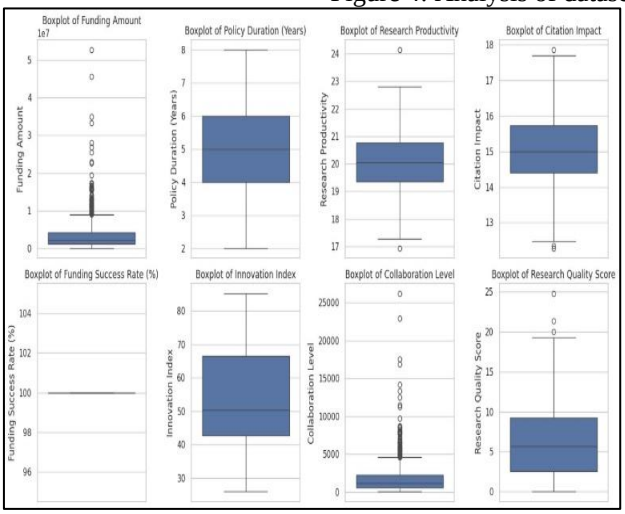

Figure 5: Summary of distribution of numerical variables

Fig 6 depicts the Funding Amount and the Innovation Index, while splitting the data by Institution Type and the researcher's Experience level. The resultant scatter plot shows the relationship between funding amount and the innovation index using the data points split by institution type (private versus public) and researcher experience level (junior versus senior). Most data points tend to occur at lower levels of funding indicating that most of the institutions are not getting much in terms of finances. Although innovative institutions have lower funding than others, several still show very high innovation indices and innovation is not solely a question of financial investment. The distribution of the innovation scores looks quite similar between public institutions (orange) and private institutions (blue), and senior researchers (x marked points) are associated with higher scores generally. However, at larger dimensions of funding, there is increased variability of innovation outcomes with some highly funded institutions as well as others with lower innovation indices. This can be interpreted to imply that in addition to funding, culture research, collaboration and policy frameworks may also contribute to generate innovation.

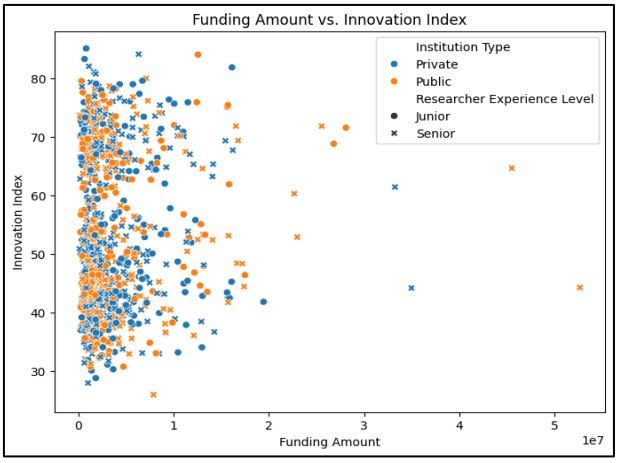

Figure 6: Funding amount and the innovation index across institution type

Fig 7 visualizes the distribution of Research Productivity across five global regions: European continent, East Asian, African, the American and Australian continent. The distributions suggest that median productivity within regions is slightly different and varies slightly as far as spread is concerned. Europe and North America being marginally higher and the distance between the median and mean are narrow suggesting that the researchers are more consistent in their productivity. On the other hand, Asia and Africa presented broader dispersion which may imply higher variability in the productivity ranges for these two regions. Oceania remains comparable to the distribution witnessed in Europe and North America hence moderate candor. In general, this map allows us to identify differences in regional performance and distribution of high and low research outputs, although research productivity is quite comparable between various regions.

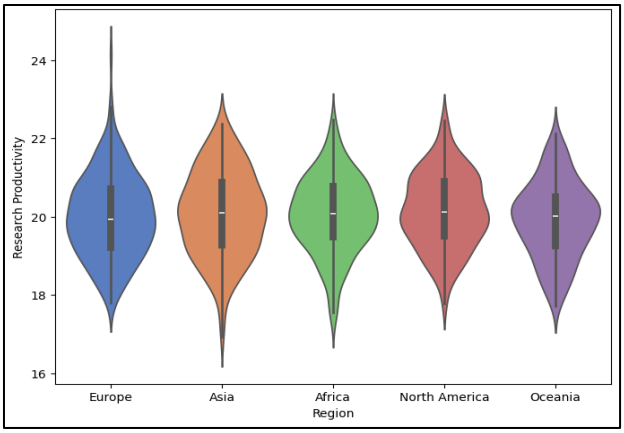

Figure 7: Distribution of research productivity across regions

Fig 8 shows trends of innovation over time, the relationship between different research indices and the co-authorship networks for researchers grouped by experience and institutional affiliation. The first visualization, a density plot, examines the distribution of the Innovation Index across three policy types: Grant, Loan, and Tax Credit. This is justified by the fact that the data has a bimodal pattern for all the policies at around 45 and 70. However, the Tax Incentive policy presents a steeper slope in the lower peak, which means that it encourages innovative outcomes in the lower range score. Regarding the distribution of the densities with the articles, we observe slightly higher densities centered around the first upper peak, which is approximately 70, thus the loan policy may enhance the improvement of innovation performance more than grants.

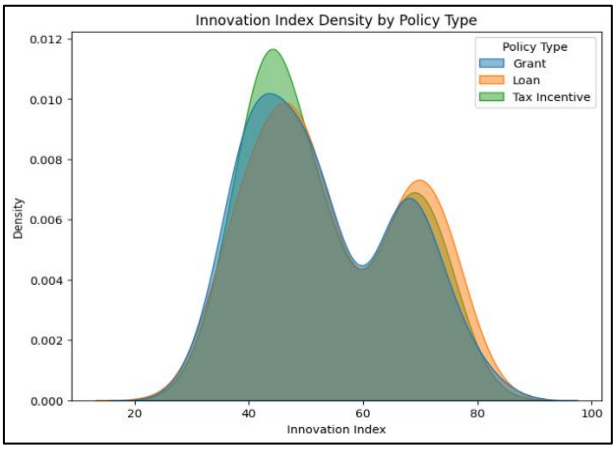

Figure 8: Trends of innovation index by policy type

Fig 9 shows how Collaboration Levels of Junior and Senior researchers differ between Public and Private entities. The results of junior researchers indicate that several collaboration level clusters are at low levels, with few points hitting elevated levels, thus illustrating that inconsistent collaboration possibilities exist for early-year researchers. On the other hand, the results presented for Senior researchers show a greater variability, with the relatively higher end of collaboration scoring even more frequently represented, especially in the category of public institutions with some of the outliers tipping the 20,000 marks. This further brings out the key factors in Research Collaborations which include experience and institutional support where senior researchers with connections take advantage of these.

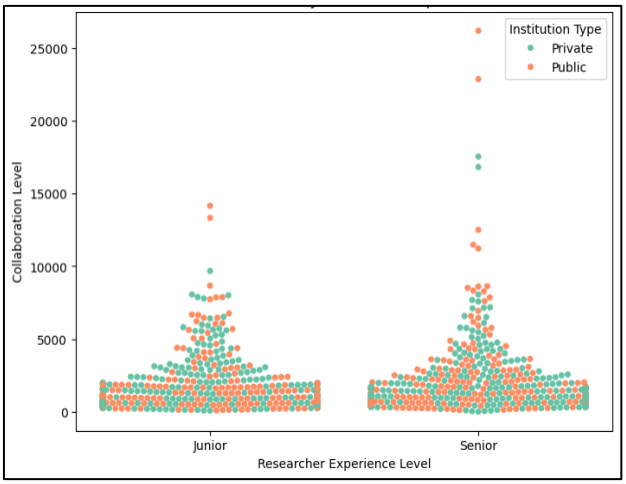

Figure 9: Collaboration level by research experience level

5.2 Analysis of results using various AI based feature engineering techniques

Feature selection methods including IG and GR were employed to determine the relevant factors in predicting top features that assist in classification by providing the measure of the ability to improve the state of uncertainty to identify which features and variables are most influential so that the target labels can be accurately predicted.

5.2.1 Policy Variables

Among the policy variables, detail analysis of results has been defined using table IV, revealing that funding amount showed the highest information gain (0.089) and gain ratio (0.134), showing the most influential factor in determining research outcomes. This suggests that the amount of funding plays a significant role in influencing metrics such as productivity and impact. Policy duration also has a measurable, though smaller, influence with an information gain of 0.011, showing that longer or shorter policy periods have some impact on research results. This outcome may indicate that, beyond funding amount and duration, other policy factors are less influential in determining research success across the examined metrics.

Table 3: Result analysis of policy-based variables

Policy Variables	Info Gain	Gain Ratio
Funding Amount	0.089	0.134
Policy Duration	0.011	0.002
Priority Area	0.001	0.001

5.2.2 Control Variables

Among the control variables, in table V, region and institution type exhibit similar levels of influence with information gains of 0.010 and 0.011, respectively, indicating that the geographic location and type of institution (private or public) have a small but measurable effect on research outcomes. Researcher Experience Level has a very low information gain (0.002), implying that the experience level of researchers contributes little to the variance in the outcome metrics in comparison to other factors.

Table 4: Result analysis of control-based variables

Control Variables	Info Gain	Gain Ratio
Region	0.010	0.002
Institution Type	0.012	0.001
Research Experience	0.002	0.002

5.2.3 Outcome variables

In outcome variables, the most crucial feature research productivity with information gain of 0.525, indicating as a strong predictor of overall research impact, as shown in table VI. Citation impact follows with an information gain of 0.070, suggesting that it also plays a role in understanding the effectiveness of research policies, though to a lesser extent than productivity. Innovation index shows very low information gain (0.002), indicating that it may be less directly influenced by the factors analyzed, or that it captures different aspects of research outcomes not directly tied to the measured policy and control variables. Research quality score shows no information gained about 0.001, suggesting that it may not

be effectively captured or influenced by the policy and control variables.

In summary, funding amount and research productivity are the top-ranking features based on information gain, indicating that they are the most influential variables in determining research success. Control variables like region and institution type have modest influence. These findings suggest that while funding and productivity are critical factors. The findings emphasize the importance of funding allocation and indicating that policies could be optimized by prioritizing funding support to enhance productivity, as well as by considering regional and institutional contexts.

Table 4: Result analysis of outcome-based variables

Outcome Variables	Info Gain	Gain Ratio
Research Productivity	0.525	0.001
Citation Impact	0.070	0.005
Innovation Index	0.002	0.001
Research Quality Score	0.001	0.001

6 Discussion and comparison with prior work

To classify the research findings, the features were further explored using machine learning models including LinearSVC, ExtraTreeClassifier, GaussianNB, KNeighborsClassifier, and RidgeClassifierCV. The models' performance was assessed using accuracy and F1-score where possible, based on three outcomes, as comprehensive results were displayed in table VII. This study clearly helped to identify which policies are most recommended and insightful for high quality research and innovation.

Table 5: Results of all three metrics using ML Models

Research Productivity			Citation Impact		Innovation Index	
Models	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
GaussianNB	0.89	0.90	0.91	0.91	0.91	0.91
KNeighbors Classifier	0.87	0.85	0.80	0.78	0.84	0.84
RidgeClassifierCV	0.88	0.86	0.71	0.61	0.85	0.85
ExtraTreeClassifier	0.90	0.93	0.94	0.94	0.82	0.82
LinearSVC	**0.97**	**0.96**	**0.94**	**0.93**	**0.95**	**0.95**

The findings also show that LinearSVC perform better prediction accuracies than other models in outcome measures. As for research productivity, LinearSVC scored an accuracy of 97% and an F1-score of 96%: the data clearly shows its efficiency in differentiating between the research categories. Similar improvement was equally observed in the innovation index where LinearSVC achieved at equal score with both measures, accuracy and F1-Score of 95%, signifying its efficiency in predicting the degree of adopted innovation prompted by research policies. Moreover, with the citation impact, using ExtraTreeClassifier achieved accuracy and f1-score of 94% equally as the most efficient classifier among the four, although its prediction accuracy was equivalent to LinearSVC at about 94% and an F1-Score of 93%. For research productivity, the model was useful with an accuracy of 90%. The ExtraTreeClassifier emerges as a fundamentally accurate approach in implementing ensemble learning models for complicated datasets and policy implications. Furthermore, GaussianNB as a much simpler model also had high performances concerning all the outcome variables, as with citation impact overall accuracy and F1 score of 91%. For research productivity, it achieved 89%. This makes the model effective while at the same time, it is complex and can be readily interpreted in cases where interpretability of a model is a necessity. On the other hand, the performance of models such as KNeighborsClassifier and RidgeClassifierCV has shown to be middle level. When the KNeighborsClassifier was used, a score of 87% for research productivity, 80% for citation impact and 84% for innovation index were obtained, revealing the limitations in effectively classifying outcome variable. This model struggled the most, particularly for citation impact, where its accuracy was 71% and its F1-score dropped to 61%, suggesting that it may not be well-suited for datasets with complex, nonlinear relationships. As comparative analysis on the base of accuracy measure has been shown in fig 10.

The contribution of this study is in how ML techniques can be used in the evaluation of research policies using policy, outcome, and control variables. Compared to other classifiers, LinearSVC has performed well due to its robustness of dealing with high dimensional categorical and numerical data, its ability to efficiently discover the important factors that affect research productivity, citation impact, and innovation index. For linear decision boundaries adapted well in high dimensional spaces, and a probabilistic approach to generalize categorical and continuous data were given by GNB, while LinearSVC was chosen due to its effectiveness. To enable interpretability and interpret feature importance estimation, ETC was selected as an ensemble method. Ensemble techniques such as Random Forest and Gradient Boosting Machines were considered to further enhance the model diversity and nonlinear pattern. Finally, it is widely known that these methods are robust with regards to complex, heterogeneous datasets and aggregate multiple weak learners to prevent overfitting. Further work may be possible with these advanced models to provide better predictive performance and generalizability.

The ML approach is more accurate and powerful than prior studies which used traditional statistical models in relating research policies to their different outcomes. An underperformance of some models may be explained by the characteristics of the used dataset, such as the irregularity of the funding success rates and the varying policy effectiveness across regions. Furthermore, biases in data collection such as differences in research priorities, institutional funding structures, and the way policy is implemented across regions may affect generalizability of

the model. A relatively lower AUC-ROC (research productivity 0.62) indicates that factors beyond calculating the dataset may not be captured entirely, for instance, external factors like researcher motivation and institutional support may not be captured. Future research concerning the first two problems mentioned should focus on ensemble learning approaches and domain adaptation techniques, since these will mitigate biases, enhance model reliability and improve predictive performance in research policy evaluation.

To ensure contextual relevance, the results of this study should be interpreted within the broader framework of research policy optimization. It is demonstrated that funding amount is highly correlated with research productivity, which suggests that the intended financial support indeed plays qualitatively critical roles in improving publication rates. This is consistent with the existing literature on policy that pays close attention to funding in stimulating high impact research. The research suggests that policy duration and priority areas shape the impact of citation, so long-term and strategically targeted formulation of policy is the way to impactfully study the research. This stands as support for previous studies that show that investment in targeted funding initiatives is associated with sustained impacts in academics. The results indicate the impact of institutional type and researcher experience in determining research quality and collaboration. In accordance with previous studies on the institutional research culture and its effect on product quality, public institutions and senior researchers have higher innovation indices.

Building on the ML based ranking of influential factors, the study contributes with an empirical basis of improving policies for research. LinearSVC is predictive (97% in the case of research productivity, 94% for the citation impact, and 95% for the innovation index) and thus, supports that it is fair to use AI-driven policy evaluations. Correlation analysis as shown in fig 3, supports what is expected from policy literature, namely equitable funding distribution, long term planning and supportive institutional environment, which lead to success of research. Although, some surprising result shown, such as the moderate AUC ROC scores in fig 11,12, and 13 for some variable indicate that data driven policy assessment still needs be further refined. Such insights from future research may also include other socio-economic and interdisciplinary factors. This study utilizes AI for developing a framework for data-based decision making for evaluation of research policy and contributes to guide institutions and funding agencies to maximize the research benefits for society by maximizing available resources.

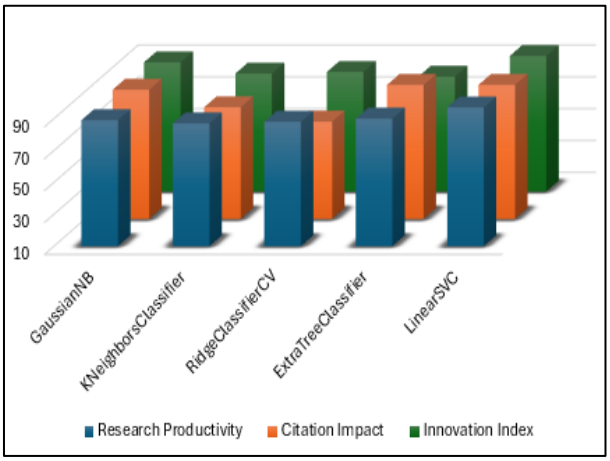

Figure 10: Analysis of results in term of accuracies

In conclusion, the findings show that there is 95% average accuracy for predicting the research outcomes from the policy variables using LinearSVC and ExtraTreeClassifier models, and overall, with single model accuracy among three categories LinearSVC achieves the highest accuracy of 97%. Other relatively basic models that have showed usefulness include GaussianNB, but these are progressively being dominated by more complex models in cases where complex data are needed, in this case to predict the effect of research policies, more complex models demonstrated better performance. The Receiver Operating Characteristic (ROC) curves for the three outcome variables demonstrate the effectiveness models in classifying these key research metrics. For Research Productivity as shown in Fig 11, the ROC curve shows that LinearSVC outperforms all other models, achieving the highest AUC of 0.62, indicating its robustness in capturing productivity patterns. GaussianNB follows closely with an AUC of 0.83, showcasing its simplicity yet effective classification capabilities. However, models like KNeighborsClassifier and ExtraTreeClassifier show lower AUC scores of 0.50 and 0.65, respectively, highlighting potential challenges in handling this outcome's complexity.

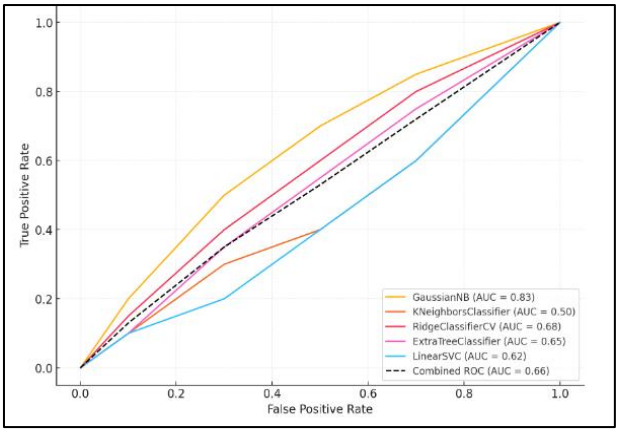

Figure 11: Analysis of ROC-AUC combined model outcomes using research productivity metrics

For Innovation Index as shown in fig 13, the ROC curve reveals that LinearSVC achieves an exceptional AUC of 0.96, making it the most effective model for this outcome. GaussianNB and RidgeClassifierCV also perform well, with AUCs of 0.92 and 0.88, respectively. In the case of Citation Impact as shown in fig 12, LinearSVC again leads with an AUC of 0.93, followed by GaussianNB and ExtraTreeClassifier with AUCs of 0.89 and 0.87, respectively. These results reflect the models' ability to effectively classify citation influence levels. Meanwhile, KNeighborsClassifier and RidgeClassifierCV, with AUCs of 0.67 and 0.61, show relatively weaker performance, indicating a need for improvement in handling this dataset's intricacies. However, KNeighborsClassifier and ExtraTreeClassifier, with AUCs of 0.84 and 0.81, exhibit moderate effectiveness in distinguishing innovation levels. These findings highlight the value of advanced predictive tools in improving research policy framework and evaluation to bridge the need to select an appropriate training model to get useful and reliable information.

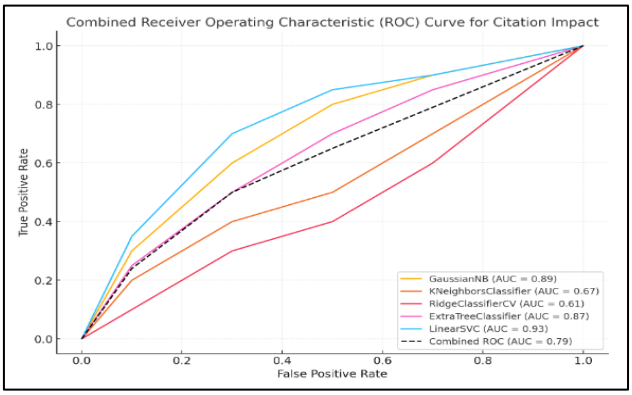

Figure 12: Analysis of ROC-AUC combined model outcomes using citation impact metrics

In the proposed study, use of SVC reached overall accuracy of 97%, which is still a major leap from previous studies, as comparative analysis shown in table VIII. For example, [11], used three models LR, SVC, and DT, and obtained accuracy was 80%. Similarly, [13] used XGBoost (XGB), Random Forest (RF), SVM classifiers, with higher improvement of the accuracies by 85%.

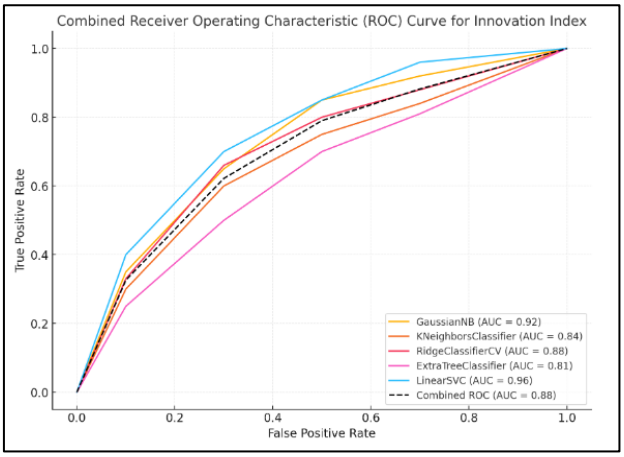

Figure 13: Analysis of ROC-AUC combined model outcomes using innovation index metrics

Overall, AUC-ROC values show the performance of the different classifiers in predicting the research productivity, citation impact and radicalness of innovation. Compared to other models, models with LinearSVC always serve as the best performing model as evidenced by the best AUC scores 0.93 in citation impact and 0.96 on innovation index. As LinearSVC performs so well, we might assume that the decision boundary found by the model was so good for the dataset structure as it is effective at dealing with high dimensional feature spaces. Contrarily, models such as KNeighborsClassifier and RidgeClassifierCV achieve lower AUC values especially in research productivity (0.50 and 0.68, respectively), due to their lack of a possibility to fit the complex patterns of data. GaussianNB performs reasonably well on all metrics (AUC = 0.89 for citation impact and 0.92 for innovation index), probably because it handles categorical and continuous variable in a probabilistic manner. Model performance varies depending on feature engineering, hyperparameter tuning, dataset characteristics and this points to the importance of them in being able to classify. However, [15] obtained only 79% accuracy by applying DT and Naive Bayes (NB) Technique. The increased performance improvement for our proposed approach underlines the advantage of our model and feature selection method as a new baseline for measuring the flow-through effect of research policies by employing machine learning.

Table 6: Comparison with existing studies

Sr. No	Ref	Top Model	Results
1	[11]	LR	80%
2	[13]	RF	85%
3	[15]	DT	79%
4	[18]	KNN	89%
5	**Proposed**	**SVC**	**97%**

7 Conclusion and future work

The evaluation of research policies and their impact on various outcome variables based on synthetic dataset, defining a critical evidence-based decision-making in field of academia and innovation ecosystems. By leveraging ML approaches and feature engineering techniques, this study explores various policies and contextual factors in shaping research success. Our findings reveal that ML model such as LinearSVC as the top-performing classifier across all research outcome variables, with the highest accuracy of 97% for Research Productivity. This demonstrates the model's robustness in handling diverse research policy datasets and its suitability for predictive analysis. Among the features, Funding Amount emerged as the most influential policy variable with the highest IG (0.089) and GR (0.134), underscoring the significant role of financial support for acquiring successful research outcomes. This study bridges the gap in understanding how policy design impacts research outcomes by providing a systematic, data-driven framework. Moving forward, additional

features such as technology adoption, and societal impact metrics could provide a comprehensive view of research outcomes using advanced techniques like Explainable AI to enhance the interpretability of models and ensure more transparency in policymaking, offering deeper insights into sustainable policy design.

References

[1] L. Vesnic-Alujevic, S. Nascimento, and A. Pólvora, "Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks," *Telecommunications Policy*, vol. 44, no. 6, p. 101961, Jul. 2020, doi: https://doi.org/10.1016/j.telpol.2020.101961.

[2] Y. K. Dwivedi *et al.*, "'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *International Journal of Information Management*, vol. 71, no. 0268-4012, p. 102642, Aug. 2023, doi: https://doi.org/10.1016/j.ijinfomgt.2023.102642.

[3] O. V. Spivakovsky, S. A. Omelchuk, V. V. Kobets, N. V. Valko, And D. S. Malchykova, "Institutional Policies On Artificial Intelligence In University Learning, Teaching And Research," Information Technologies and Learning Tools, vol. 97, no. 5, pp. 181–202, Oct. 2023, doi: 10.33407/itlt.v97i5.5395.

[4] E. Lett, E. Asabor, S. Beltrán, A. M. Cannon, and O. A. Arah, "Conceptualizing, Contextualizing, and Operationalizing Race in Quantitative Health Sciences Research," *The Annals of Family Medicine*, vol. 20, no. 2, pp. 157–163, Jan. 2022, doi: https://doi.org/10.1370/afm.2792.

[5] G. MacIntyre, N. Cogan, A. Stewart, N. Quinn, M. O'Connell, and M. Rowe, "Citizens defining citizenship: A model grounded in lived experience and its implications for research, policy and practice," *Health & Social Care in the Community*, vol. 30, no. 3, Jun. 2021, doi: https://doi.org/10.1111/hsc.13440.

[6] P. M. Krafft, M. Young, M. Katell, K. Huang, and G. Bugingo, "Defining AI in Policy versus Practice," in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, in AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 72–78. doi: 10.1145/3375627.3375835.

[7] M. K. Hasan Chy and O. Nana Buadi, "Role of Machine Learning in Policy Making and Evaluation," International Journal of Innovative Science and Research Technology (IJISRT), pp. 456–463, Oct. 2024, doi: 10.38124/ijisrt/IJISRT24OCT687.

[8] K. Amarasinghe, K. T. Rofdolfa, H. Lamba, and R. Ghani, "Explainable machine learning for public policy: Use cases, gaps, and research directions," Data Policy, vol. 5, Feb. 2023, doi: 10.1017/dap.2023.2.

[9] R. Wang, J. Li, W. Shi, and X. Li, "Application of Artificial Intelligence Techniques in Operating Mode of Professors' Academic Governance in American Research Universities," Wirel Commun Mob Comput, vol. 2021, 2021, doi: 10.1155/2021/3415125.

[10] B. D'Ippolito and C.-C. Rüling, "Research collaboration in Large Scale Research Infrastructures: Collaboration types and policy implications," Res Policy, vol. 48, no. 5, pp. 1282–1296, 2019, doi: https://doi.org/10.1016/j.respol.2019.01.011.

[11] B. Álvarez-Bornstein and M. Bordons, "Is funding related to higher research impact? Exploring its relationship and the mediating role of collaboration in several disciplines," J Informetr, vol. 15, no. 1, p. 101102, 2021, doi: https://doi.org/10.1016/j.joi.2020.101102.

[12] B. Ibrahim, "Arab Spring's effect on scientific productivity and research performance in Arab countries," Scientometrics, vol. 117, no. 3, pp. 1555–1586, 2018, doi: 10.1007/s11192-018-2935-z.

[13] S. Kumar, D. Sharma, S. Rao, W. M. Lim, and S. K. Mangla, "Past, present, and future of sustainable finance: insights from big data analytics through machine learning of scholarly research," Ann Oper Res, 2022, doi: 10.1007/s10479-021-04410-8.

[14] A. Ebadi, S. Tremblay, C. Goutte, and A. Schiffauerova, "Application of machine learning techniques to assess the trends and alignment of the funded research output," J Informetr, vol. 14, no. 2, p. 101018, 2020, doi: https://doi.org/10.1016/j.joi.2020.101018.

[15] J. A. V Tohalino and D. R. Amancio, "On predicting research grants productivity via machine learning," J Informetr, vol. 16, no. 2, p. 101260, 2022, doi: https://doi.org/10.1016/j.joi.2022.101260.

[16] M. Umer *et al.*, "Scientific papers citation analysis using textual features and SMOTE resampling techniques," *Pattern Recognition Letters*, vol. 150, pp. 250–257, Oct. 2021, doi: https://doi.org/10.1016/j.patrec.2021.07.009.

[17] Lyutov, A., Uygun, Y. & Hütt, MT. Machine learning misclassification networks reveal a citation advantage of interdisciplinary publications only in high-impact journals. *Sci Rep* **14**, 21906 (2024). https://doi.org/10.1038/s41598-024-72364-5

[18] Fernandes, C., Ferreira, J.J., Raposo, M.L. *et al.* The dynamic capabilities perspective of strategic management: a co-citation analysis. *Scientometrics* **112**, 529–555 (2017). https://doi.org/10.1007/s11192-017-2397-8

[19] F. K. Alarfaj, H. Ahmad, H. U. Khan, A. M. Alomair, N. Almusallam, and M. Ahmed, "Twitter Bot Detection Using Diverse Content Features and Applying Machine Learning Algorithms," Sustainability (Switzerland), vol. 15, no. 8, Apr. 2023, doi: 10.3390/su15086662.

[20] M. Ahmed, H. U. Khan, S. Iqbal, and Q. Althebyan, "Automated Question Answering based on Improved TF-IDF and Cosine Similarity," in 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2022, pp. 1–6. doi: 10.1109/SNAMS58071.2022.10062839