Predictive Analytics - Final Assessment

by PC Lekgwara

Submission date: 25-Oct-2025 02:00PM (UTC+0200)

Submission ID: 2792107373

File name: 114743_PC_Lekgwara_Predictive_Analytics_-_Final_Assessment_1523335_1152674049.pdf (608.91K)

Word count: 4840

Character count: 24790

PREDICTIVE ANALYTICS FINAL ASSESSMENT

OCTOBER 2025

MASTERS OF APPLIED DATA SCIENCE

PREDICTIVE ANALYTICS

Due 25 October 2025 @ 23:59

NAME AND SURNAME: Piletsi Claude Lekgwara

STUDENT NUMBER: 216009458

Table Of Contents

DEEP LEARNING AND UNDERSTANDING DROPOUT RATE

Abstract

As of 2025 in South Africa the current youth unemployment rate stands at 46.1%. The ages range from 15 to 34 years. There are social and economic factors that are the cause of this, but the most evident factor is the dropout of children between primary and high school. It is evident in society that without an education the chance of getting a job greatly decreases. There are evident factors such as parental backgrounds and household income, areas and age that lead to a child dropping out of school. This report aims to learn, understand and define patterns of relevant factors that lead to a child dropping out of school. The definition of these patterns will be interpreted and communicated in an understandable manner to educational regulatory bodies. Their understanding of the underlying patterns and relationships related to dropout will help them develop strategies that can reduce the dropout rate and decrease the youth unemployment rate. We developed a deep learning model to understand and interpret the underlying relationships of the features that can lead to a child dropping out of school. The deep learning model learns to predict if the child will dropout = 1, or not dropout = 0. We train the model on seen data and then we test it on unseen data and validate its performance to see how it would generalize on real world data. We then applied a model agnostic technique called Shapley Additive exPlanations(SHAP). This technique communicates the score of a factor. The score indicates the contribution of the dropout binary class ultimately helping in why the model made a specific prediction. The results from the SHAP algorithm are what will be used to communicate back to the educational regulatory bodies in an interoperable and understandable way. Running Kfolds validation techniques we discovered that the model learns well and after running the SHAP algorithm we determined the top 3 factors that influence a child dropping out is their attendance, living standards whether it be urban or rural and their age. This information has pinpointed a critical reason for children dropping out and can now develop strategies to reduce the dropout of children in primary and high school by now being able to attend to their needs.

Introduction

At the beginning of the first quarter of 2019 the youth unemployment rate for age groups 15-24 and 25-34 was 55.2% and 34.2%. Since then, it has shown a gradual increase in these rates and now stands at 59.6% and 39.4% in the final quarter of 2024. (Statista Research Department, 2025). The youth unemployment rates had greatly decreased in the year the second quarter of 2020 for all age groups. In the third quarter of the same year the unemployment rates increased greater than in 2019 and continued increasing till it reached a peak in the third and fourth quarter of 2021 and began to decrease again. These two age groups rank number 1 and 2 amongst other age groups in the unemployment rate

The most obvious case in the labour field is that the level of education and employment opportunities has a direct relationship. This means that the higher an individual climbs the educational ladder he or she creates

more opportunities of employment for themselves. Education provides the basic and all the way up to advanced theoretical framework and thinking to perform various tasks. Lack of this knowledge results in the inability to perform various tasks at various levels. This incapability results in the distrust of employers, and this distrust leads to a decrease in the chance of employability which ultimately results in accumulating to the unemployment rate.

Education levels rank from primary to tertiary and to get to the next rank the current level needs to be achieved. The tertiary education level focuses on industry knowledge and standards. Success in the tertiary education level opens the most doors to employment. An individual dropping out at any of the levels of education hinders the GDP of the country but dropping out at a tertiary level is less harmful to an individual who dropped out during primary school or high school. A primary or high school dropout moves along with the unemployment rate of the country and spends the most time in the unemployment bin. To overcome the unemployment rate challenge in the country, long-term role players need to be assessed. This assessment will provide details of the factors leading to dropout.

According to Archambault, Janosz, Dupéré, Brault and Andrew (2017), dropout can result from configurations of individuals, social and family risk factors. These configurations may also be shaped by larger social and cultural contexts. Sparks (2022) then identified that students who come from a background where there is low income and those with disabilities have been the group with the greatest increase in dropouts. Individuals from different backgrounds experience personal daily situations from a social and economic standpoint. These experiences tend to creep into their educational lives and can negatively influence their performance and willingness to learn. This can lead to the individual dropping out. This provides evidence that primary and high school dropouts predominantly have to do with social and economic factors. The currency is that there is no solution that converges social and economic factors together that can detect students that are at risk of dropping out in the primary and high school education levels.

The objective is to develop a deep learning model that will not only understand patterns and relationships between social and economic factors but will also predict students in primary and high schools that are at risk of dropping out. The deep learning model will serve as an input for the model agnostic technique, which will help in communicating the underlying patterns and relationships to schools and educational bodies. SHAP will also help communicate the patterns and relationships in a less complex manner so schools and regulatory bodies can then develop strategies catered to different schools whether it be in a rural or urban area. The understandable information communicated to the audience will facilitate in the intervention of students that are at risk of dropping out. Regulatory bodies can use this information to develop strategies that can help schools with the infrastructure needed to prevent students from dropping out of critical educational levels. Regulatory bodies will now be a catalyst in the reduction of the youth unemployment rate.

Literature Review

“The Secretary of State for Work and Pensions, Liz Kendall(Lab, Leicester West) said that almost 1 million young people were not in education, employment or training, which was terrible for their living standards, their future job prospects and their health” (Martin, et al., 2025). According to Stats sa (2025) education plays a pivotal role in the determination of young individual’s employment prospects, they further go on to say that individual without a matric qualification face the steepest uphill climb with an unemployment rate of 51.6%. These references provide evidence on how the lack of education has a negative relationship with employment prospects. Negative employment prospects have social and economic implications for the individual and the productivity of a country. The increased rate of dropouts directly affects unemployment rate. “When students drop out of high school, this usually has a negative impact both on their personal development and on society both on their personal development and on society, as they exposed to an increased risk of unemployment, health problems and social integration difficulties” (CRET, 2025). Relevant studies provide evidence that the cause of dropping out of school isn’t an occurrence that happens now, it’s a series of occurrences that accumulate over time. The series of occurrences predominantly have to do with social factors regarding the individual’s personal wellbeing and wellbeing at school. Dropping out leads to uncertainty in the job market, distrust from banks, the lack of ability to acquire home loans and financing for automobiles. It can lead to an individual’s decrease in social status which can lead to self-isolation. Dropout rates is an important factor to understand unemployment rate because it is measured as a factor in unemployment rate. Understanding the complex social and economic convergence of individuals dropping out will help with the detection of individuals who are at risk of dropping out.

There are social, demographic and economic factors that contribute to individuals dropping out. From a social perspective. Negative social status hinders the performance of an individual during class hours which will decrease success rate of the individual. Unrest at home negatively leaks into the performance of an individual during educational hours. Parents’ attitude to an education leaks into individual’s attitude to education. This leak is directly proportional to the overall performance of an individual at school. From an economic perspective, the total household income. Household income encompasses living standards and access to basic utilities. This affects an individual’s view on education. This view not only leaks into the individual’s performance but also the attendance of the individual. Attendance and dropout have an inversely proportional relationship. This means that the more an individual attends school the chance of dropping out decreases. The demographics perspective focuses on age and area. On average students in the secondary level of education tend to drop out more. This means that age and the risk of dropping out have a positive relationship, the older the individual the greater the chance of that individual dropping out. Area focuses on whether it’s rural or urban. Urban and rural demographics are exposed to different situations. The dropout rate in rural areas is generally higher than in urban areas. There are other various factors that lead to dropouts that are imposed by regulatory bodies that can influence individuals dropping out of school for

example, quality of educators, lack of compassion from educators towards learners and some nuance instances such as bullying.

“Design a model with 90% reliability, which helps in the prediction of school dropouts in higher and secondary education institutions, implementing machine learning techniques.” (Jiménez-Gutiérrez, et al., 2024). This is evidence that machine learning models can be trusted to understand and interpret the underlying relationships between economic and social factors that can help in predicting students that are at risk of dropping out. The primary goal of machine learning models is to recognize complex patterns within datasets. The complex patterns in this case exist because of the convergence of social and economic factors. The model will learn these patterns and then be evaluated on unseen data and see how well it is at predicting the individuals who are at risk of dropping out. A major benefit of machine learning models is that they are robust and learn overtime with new data. This constant learning helps schools and regulatory bodies adapt their strategies to consistently changing patterns and relationships. Machine learning models remove human error from tasks. Interpretable machine learning models are designed to simplify the complex patterns that the machine learned to predict or classify a target. They are important because they ensure that there’s trust and transparency by showing how predictions are reached. They also provide information on the most important features. “SHAP offers a model-agnostic and consistent method for elucidating the contribution of each predictor variable to predictions made” (Sani, et al., 2025). SHAP assigns a score to a feature ranking from highest to lowest score. The feature with the highest score is the most important feature.

Building an interpretable deep learning model is important because the interpretations are what we feed the regulatory bodies and schools. The interpretations serve as information for the audience. They need this information to build or adapt current strategies that will result in the reduction of primary and high school dropouts. Regulatory bodies and schools lacked knowledge of the complexities of the social and economic factor convergence. They had a surface level view of the causes of dropouts. This deep learning model directs the audience to which features have the biggest impact resulting in dropping out of school. The information provided by the interpretable model will be the foundation of building strategies that reduce unemployment rate as well as the number of individuals who suffer from the social and economic implications of dropping out of school. This will reduce the number of individuals that move along with the youth unemployment rate curve and even reduce the amount time spent on the unemployment rate curve.

Methodology

The dataset has a mix of numerical and categorical columns. The numerical columns are age, household_income and highest_grade. The categorical columns are gender, province, urban_rural, parent_education, attendance and dropout. The target for the dataset is the dropout column. If dropout=1 the individual is at risk of dropping out and if dropout=0 then the individual is not at risk of dropping out. Age refers to the individual’s age, gender refers to whether the individual is male or female, province refers to the province in which the student resides in and urban_rural refers to whether the individual lives in an urban or rural area. Household_income is the sum of income earned in the individual’s household. Parent_education

refers to the parent with the highest education level. Attendance refers to whether the child attends classes at school. Highest_grade is the highest grade the student achieved. After analysis I split the data into df_train and df_test.

The train data set has 1600 rows with missing rows only in the parent_education column and no duplicates. After careful analysis of the train data set we detected inconsistencies in the attendance field whereby there are students who have a dropout value of 1 but they are attending. This is messy data as a student cannot be a dropout but still attending. There are also students who are in very low grades for their age, so we created an age map that assumes that we follow normal age to grade distinctions in South Africa for example age = 15, highest_grade = 8. This projects a real-life scenario. We then check if there are any duplicates in the dataset and if there are any, we remove them because the model has a row that it can learn from. This step is the dataset check that will be used for seen and unseen data this is the Machine Learning OPS stage called data drift monitoring stage. Which looks at historical data and then flags rows that do not comply with historical data standards and makes the relevant changes to the unseen data. The data is still unseen by the model.

Exploratory data analysis was done on the training set; Table 1.1 summarizes what was learned about our data. This data also helps with building more rules for the data drift monitoring stage.

Numerical Features	**Summary**
Age	Majority of the training set falls between 12 and 19 years. There are outliers in the age column but the median is about 15 years old Bimodal shape
Household_income	Most of the dataset falls under the first bin meaning that a lot of incomes earn close to 0. There are outliers in household income with 1000000 being an extreme case. The shape of distribution is positively skewed.
Highest_grade	Most of the dataset falls within ranges of grade to grade 11. There can't be outliers in highest_grade because it was accounted for in the data drift monitoring phase. The shape is negatively skewed.
Categorical Columns	
Gender	Female is the value that appears the most
Province	Northern Cape is the most frequent value
Urban_rural	Urban areas is the most frequent database

Parent_education	Secondary education is the most frequent amongst parents
Attendance	Yes – Is the most frequent value

Table 1.1 – summary of features and their summary

After exploratory data analysis, the data drift monitoring stage gets updated with a solution for new data coming into the destination (Data Science Warehouse). The updated stage now compensates for all columns in the data set. The stage imputes the categorical columns by the mode, imputes missing values in the parent_education column with "Missing', which is described by a parent with no education level. This stage will also indicate which column has been imputated or if no columns needed imputation. This leads to running isolation forest with a contamination of 0.05. Isolation forest is used to detect outliers from the outlier columns which is age and household_income. Isolation forest is a machine learning algorithm which falls under the tree modules. The use of isolation forest and outlier columns helps us detect data points deemed as outliers the rows are not removed but are flagged. A binary column is created for this flag and now acts as a feature for our dataset. The feature will be interpreted by its importance and will determine its weight with regards to prediction of the binary class. The train dataset now gets split into features and targets.

A numerical pipeline is created whereby it imputes missing values using random forest regressor and then after imputation it scales all numerical features. A categorical pipeline is also created whereby all categorical features are one_hot encoded. We create a preprocessor pipeline that fits and transforms all columns in the training feature dataset.

To validate if the model is ready, the Kfolds validation method is used. Kfolds validation uses the training data. It splits the training data into training data and test data. The training data is split into the desired number of folds, which in this case is 4. Kfolds trains and evaluates the model based on the desired model we'll be building later the data test data is known as the validation data and that is the data that it is evaluated on. Figure 1.1 depicts the results of the Kfolds validation algorithm.

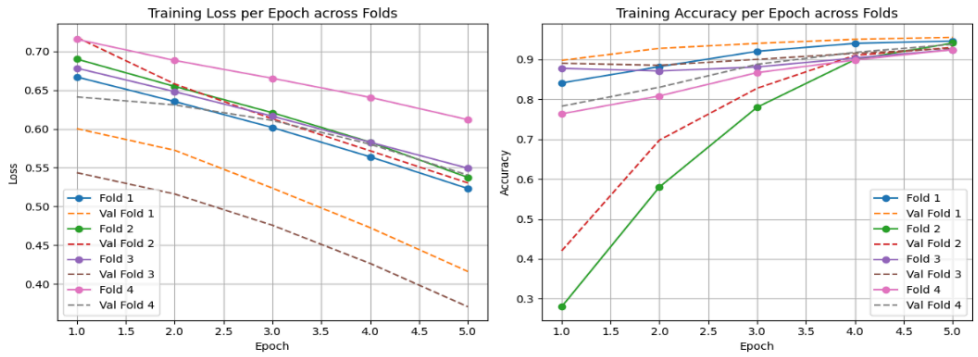


Figure 1.1 - Kfolds validation summary

The left plot shows the Training loss per epoch as well as the validation. All folds are decreasing for training and validation data. The vertical distance of the corresponding training and validation folds are remarkably similar This suggests that the model is improving its fit to both training and unseen data. The accuracy training folds generally start lower than the validation folds, but they all converge around the same place. This validates that the model can now be trained because there is no sign of overfitting, the model's learning stability is consistent across all folds and there's a consistent generalization to the unseen data.

The Kfolds validation technique used a dense binary neural network for classification, therefore it is the desired deep learning model that is going to use the full training features and target. The model is sequential. This means that layers are stacked, the purpose of this is so that the output of one layer is the input for the next layer. The model has 3 hidden layers starting from 32 neurons all the way up to 8, with an activation of relu. The output layer has an activation of sigmoid which outputs a probability between 0 and 1. To compile the model the Adam() optimizer is used and a loss function of binary crossentropy is used. This is because the model is predicting a binary class. The metric used to measure the success rate of the model is accuracy. The accuracy metric will measure how the number of times predicted values were correct. The training features and target have now been fitted to the model and are ready for evaluation. The 5 cross validation was used to evaluate the model. The metrics are accuracy, precision, recall, f1-score and ROC_AUC. A score will be assigned to the metrics and interpreted. Accuracy measures if the model predicted the correct class, Precision measures, when it predicted the positive class how many of the predictions were correct. Recall measures the number of positives that model predicted against the total sample of positives. F1_score measures recall and precision together, it measures the duality of the model for example it wants to catch all classes not just one to solve a small business problem. ROC_AUC refers to the model's ability to distinguish the two classes.

The final phase of the model is to integrate it with model agnostic technique SHAP. SHAP uses feature names, therefore feature names need to be extracted from the preprocessing pipeline. SHAP needs to know what predictive model was used which in this case was a dense neural network. The deep explainer function was passed to the SHAP function. The previous step now helps SHAP get the values that are needed to present feature importance and how the model reached a certain prediction.

Application

To evaluate the performance of unseen data, the data had to go through the data drift monitoring stage as it has to pass a certain criterion to be deemed usable. The data has not been seen. The data drift stage is a machine learning operations process before it gets loaded into the data science warehouse. After this pass the predicted proportions are calculated and then the threshold was determined. A threshold greater than 0.5 tells it to only bring in values where the probability is greater than 50%. The predicted data is now evaluated against the actual target values in the test data set. Results are validated using the 5 cross validation, which is a great metric to use for classification. This table summarizes the results and implications of the results

Metric	Implication
Accuracy	0.93
Precision	1.0
Recall	0.55
F1_Score	0.71
ROC_AUC	0.99

***Table** 1.2 – 5 Cross Validation Metrics for Binary Classification*

The model did well on unseen data. 93% of the predictions were correct when evaluated against true target values. The model got 100% of all the positive predictions correct that it made. The recall value is 55% This means that out of the total positive sample the model got just over half of them. F1_score was 71% therefore the model does well even when a class imbalance is evident which is true in this case. The ROC_AUC score is 99%, this means that it is almost perfect at distinguishing the different classes. This is evident that the model didn't memorize during the training stage and that it learned the complex patterns and relationships in the dataset and performs well on unseen data.

SHAP's foundation is the model's ability to learn complex relationships. SHAP just reports back the complex relationships back in an interpretable way. It is the heart of understanding complex business problems and pin pointing starting points to solve the business problems. This study aims to help detect students in primary and high school that are at risk of dropping out. Through integration of SHAP and the neural network, features that influence dropping out in school have been detected. These findings can now be communicated. The first image depicts the features that heavily influence the dropping out of learner.

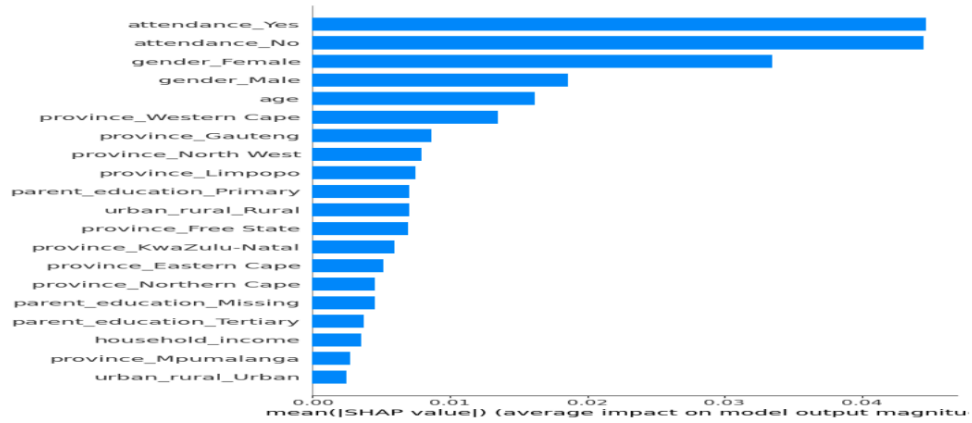


Figure 1.2 – Feature importance when predicting output

Features names are encoded but they are still relatively close to each other. Attendance is the feature that is the most important when trying to understand feature importance. Gender is second and age is third and provinces are fourth and fifth is parent education level. It is evident that majority of the factors that contribute to school dropouts are demographic factors. SHAP also provides more detail with regards to how each feature contributes to the prediction of an outcome. The image below aids in this.

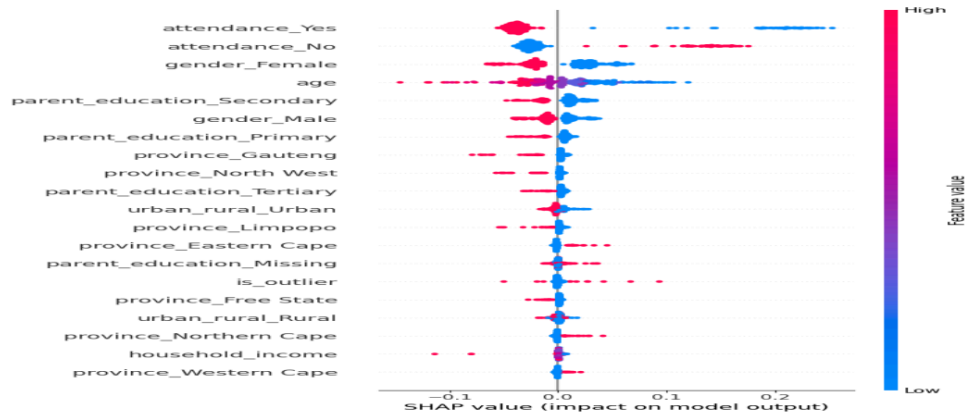


Figure 1.3 – SHAP summary plot

The red color of this plot shows how high the actual value is in the dataset. It works the same for categorical features because they typically range from 0-1, but 1 is the highest value and for numerical features, they follow the natural order for example 1500 > 1000. The features are also still ordered from most important to least important. This plot depicts the direction of the prediction to the right is dropout=1 and to the left dropout =0. The plot provides evidence to the audience that when attendance=1 the chances of the student dropping out are slim, the attendance the latter also provides evidence of this, when attendance = No the right shifts to the right therefore this means that the individual is at risk of dropping out. Features with least importance their SHAP values are all clustered in the middle. Take for instance household_income there are rare cases whereby the higher the household_income value the chances of dropping out are slim, but majority of the values are clustered in the middle. This also provides evidence that dropping out of school does not have to do with monetary perspectives.

Discussion

This section is going to focus on why certain features are important in dropout rates. Attendance is directly proportional to academic performance. The lack of attendance can lead to failing more than once. Constant failure leads to negative mental implications which can lead to distrust in oneself or the education system and lead to dropping out of school. Female pregnancies in highs school is a dominant factor in dropping out. Age is a dynamic one because some individual's drop out because they can start working at the age of 16 or some don't drop out because they feel like they are closer to the finish line, that is why the splits look very

similar. It is evident that dropout likelihood increases in different provinces of the country. This means that education regulatory bodies need to develop strategies that focus on social factors that address issues with attendance, gender and age and cater for the dynamic nature of the different provinces. They also need to restructure their current policies so that they can adapt to the newly developed strategies. All levels of education need psychologists so they can have a deeper understanding of mental implications these features have on learners. This understanding will also be beneficial in the strategies they need to develop in order to reduce dropout rates and unemployment rates.

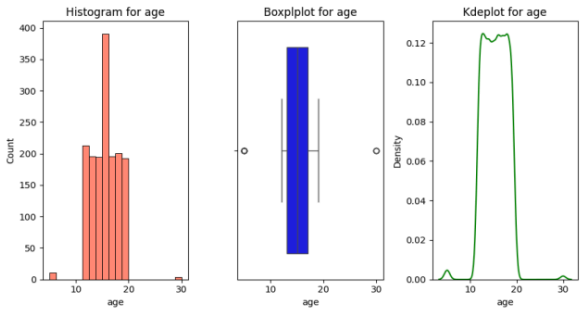
The limitations of this study are that the dataset was small, and the sample size is not enough to solve such a large problem in this country. Data should be collected for different income groups to determine mental implications based on their demographics. The outlier flag also doesn't help that much in terms of model generalization it did add noise, to improve, these outliers should be detected and removed. There was also a feature imbalance, an improvement would've been to add a class_weight function that can help the model learn better and detect the minority class better.

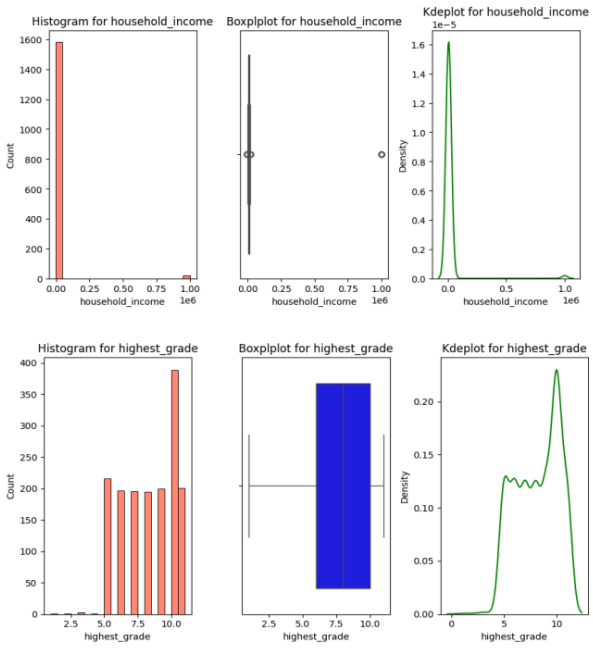
Conclusion

Neural networks integrated with model agnostic techniques close the gap between data and business problems. It provides a deeper understanding of global issues as well as the underlying reasons for global issues. Dropping out of school has psychological implications for learners as well as monetary implications on the country. The synergy between neural networks and model agnostic techniques provides information not only governing bodies but to the population. Reduction in dropout rates not only benefits future leaders but it benefits the labour force of the country and will in turn contribute to the growth of the productivity of the country.

Appendix:

A) – EDA for numerical and columns





References

Anon., 2025. *South Africa's Youth in the Labour Market: A Decade in Review*. [Online]
Available at: https://www.statssa.gov.za/?p=18398

Archambault, I. et al., 2017. Individual, social, and family factors associated with high school dropout among low- SES youth: Differential effects as a function of immigrant status.. *British Journal of Educational Psychology*.

CRET, C., 2025. School dropout among adolescents- School satisfaction and self efficacy. *Journal Plus Education,* Volume 35.

Jiménez-Gutiérrez, A. L., Mota-Hernández, C. I., Mezura-Montes, E. & Alvarado-Corona, R., 2024. Application of the performance of machine learning techniques as support in the prediction of school dropout. *Scientific Reports*.

Martin, A., Foody, E., Curtis , C. & Shanker, B., 2025. *NEET young people*. s.l., House of Commons.

Sani, J., Halane, S., Ahmed, A. M. & Ahmed, M. M., 2025. Application of machine learning algorithms and SHAP explanations to predict fertility preference among reproductive women in Somalia. *Scientific Reports*.

Sparks, S. D., 2022. Cascade of Factors Drive Down Graduation Rates.. *Education Week*.

Statista Research Department, 2025. *Unemployment rate in South Africa from Q1 2019 to Q4 2024, by age group*. [Online]
Available at: https://www.statista.com/statistics/1129482/unemployment-rate-by-age-group-in-south-africa/

Predictive Analytics - Final Assessment

ORIGINALITY REPORT

4%	3%	2%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.graaffreinetadvertiser.com Internet Source	1%
2	Bui Thanh Hung, M. Sekar, Ayhan Esi, R. Senthil Kumar. "Applications of Mathematics in Science and Technology - International Conference on Mathematical Appplications in Science and Technology", CRC Press, 2025 Publication	<1%
3	www.coursehero.com Internet Source	<1%
4	eug.petroleum.gov.eg Internet Source	<1%
5	openhub.spu.ac.za Internet Source	<1%
6	everant.org Internet Source	<1%
7	Edward M. Sosu, Pichyada Pheunpha. "Trajectory of University Dropout: Investigating the Cumulative Effect of Academic Vulnerability and Proximity to Family Support", Frontiers in Education, 2019 Publication	<1%
8	Swapna Donepudi, Rajeswari Nakka, Krishna Kishore Thota, Mohan Ajmeera, S. Phani Praveen, S. Sindhura. "Chapter 7 Enhancing Person-Centric Health Care for Diabetes Prediction: A Comparative Study of LightGBM, XGBoost, and Hybrid LIGB Model", Springer Science and Business Media LLC, 2025 Publication	<1%

9	github.com Internet Source	<1%
10	www.researchwithnj.com Internet Source	<1%
11	bspace.buid.ac.ae Internet Source	<1%
12	peerj.com Internet Source	<1%
13	worldwidescience.org Internet Source	<1%
14	www.slideshare.net Internet Source	<1%
15	Samuel Ntsanwisi. "Societal factors impacting young Black men in South Africa", Journal of Humanities and Applied Social Sciences, 2024 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		