informatics

MDPI

Article

ACME: A Classification Model for Explaining the Risk of Preeclampsia Based on Bayesian Network Classifiers and a Non-Redundant Feature Selection Approach

Franklin Parrales-Bravo [1,*,†], Rosangela Caicedo-Quiroz [2,†], Elianne Rodríguez-Larraburu [3,†] and Julio Barzola-Monteses [1,4,†]

1 Grupo de Investigación en Inteligencia Artificial, Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Guayaquil 090514, Ecuador; jjbarzolam@ube.edu.ec
2 Centro de Estudios para el Cuidado Integral y la Promoción de la Salud, Universidad Bolivariana del Ecuador, Km 5 ½ vía Durán—Yaguachi, Durán 092405, Ecuador; rcaicedoq@ube.edu.ec
3 Facultad de Salud y Servicios Sociales, Instituto Superior Universitario Bolivariano de Tecnología, Guayaquil 090313, Ecuador; erodriguez@bolivariano.edu.ec
4 Centro de Estudios en Tecnologías Aplicadas, Universidad Bolivariana del Ecuador, Km 5 ½ vía Durán—Yaguachi, Durán 092405, Ecuador
* Correspondence: franklin.parralesb@ug.edu.ec
† These authors contributed equally to this work.

Abstract: While preeclampsia is the leading cause of maternal death in Guayas province (Ecuador), its causes have not yet been studied in depth. The objective of this research is to build a Bayesian network classifier to diagnose cases of preeclampsia while facilitating the understanding of the causes that generate this disease. Data for the years 2017 through 2023 were gathered retrospectively from medical histories of patients treated at "IESS Los Ceibos" hospital in Guayaquil, Ecuador. Naïve Bayes (NB), The Chow–Liu Tree-Augmented Naïve Bayes (TAN_{cl}), and Semi Naïve Bayes (FSSJ) algorithms have been considered for building explainable classification models. A proposed Non-Redundant Feature Selection approach (NoReFS) is proposed to perform the feature selection task. The model trained with the TAN_{cl} and NoReFS was the best of them, with an accuracy close to 90%. According to the best model, patients whose age is above 35 years, have a severe vaginal infection, live in a rural area, use tobacco, have a family history of diabetes, and have had a personal history of hypertension are those with a high risk of developing preeclampsia.

Keywords: preeclampsia; Bayesian networks; feature subset selection; machine learning; explainable AI

Citation: Parrales-Bravo, F.; Caicedo-Quiroz, R.; Rodríguez-Larraburu, E.; Barzola-Monteses, J. ACME: A Classification Model for Explaining the Risk of Preeclampsia Based on Bayesian Network Classifiers and a Non-Redundant Feature Selection Approach. *Informatics* **2024**, *11*, 31. https://doi.org/10.3390/informatics11020031

Academic Editor: Pengyu Hong

Received: 22 January 2024
Revised: 8 May 2024
Accepted: 12 May 2024
Published: 17 May 2024

1. Introduction

Nowadays, preeclampsia, a placentally derived illness during pregnancy, is responsible for 10–15% of all maternal deaths worldwide [1,2]. Preeclampsia is a progressive multisystem disorder that typically occurs after 20 weeks of gestation or postpartum in a woman with previously normal blood pressure [1]. It is characterized by the onset of newly developed arterial hypertension, which may include proteinuria and multiple organ dysfunction, such as hematological abnormalities, alterations in biochemical markers of coagulation, and hepatic function. Moreover, it can be accompanied by neurological complications or evidence of uteroplacental dysfunction, such as fetal growth restriction [3].

Although delivery is the only definitive treatment for preeclampsia, clinical management involves finding a balance between reducing the risk for the pregnant woman and the risk of prematurity for the fetus. For this purpose, emphasis is placed on the importance of early detection of pregnant women at higher risk of developing diseases and implementing preventive measures and interventions in the management of pre-existing medical conditions [4].

It is difficult to determine which patients will suffer preeclampsia because it can appear without a spike in blood pressure or the presence of protein in the urine [5]. However, several risk factors have been identified for the condition [6]. According to the American College of Obstetricians and Gynecologists (ACOG), previous preeclampsia, chronic hypertension, diabetes mellitus, chronic kidney disease, autoimmune diseases, and multifetal pregnancies are the main "high" risk factors for the development of preeclampsia [7,8]. Other risk factors, classified as "moderate", include nulliparity, advanced maternal age, maternal obesity, and family history, among others [8].

In the Ecuadorian context, according to the 2020 Ministry of Public Health report, hypertensive disorder due to severe preeclampsia is the leading cause of maternal death in Guayas province (Ecuador) [9]. Thus, the identification and timely treatment of this disease can lead to a significant improvement in maternal and perinatal outcomes [10]. In this sense, machine learning (ML) techniques can be useful to build early predictive models of preeclampsia. For example, the assessment of the immune system in early pregnancy can be useful in predicting the risk of preeclampsia in asymptomatic pregnant women [11]. However, that goal becomes a real challenge because the causes of this disease are still poorly understood [11,12]. Furthermore, the existence of several risk factors and probable existence of multiple pathogenic phenotypes of preeclampsia make the construction of predictive models even more difficult [13].

It is common knowledge that, to apply ML models, the data must be preprocessed [14]. There are some techniques to achieve this preprocessing; in this study, the Feature Subset Selection (FSS) task is applied. According to Gopika [15], a major problem in datasets is the presence of irrelevant or redundant features, which leads to statistical correlation problems between them. An alternative to reduce the dimension of these features is to apply different FSS approaches such as filters, wrappers, and embedded methods used in the ML field. However, FSS techniques do not completely remove these redundant or irrelevant features [16].

All in all, this work analyzes the performance of some Bayesian network classifiers induced from medical records collected at the gynecology unit of the "IESS Los Ceibos" hospital. To build the models, a combination of filter and wrapper approaches is considered for the feature selection task. This new approach has been reported and has presented better performance than existing FSS techniques [2,14].

The remaining sections of the paper are structured as follows. In Section 2, prior research relevant to this investigation is provided. Section 3 offers a concise overview of the methodology to train and evaluate Bayesian network classifiers to predict preeclampsia. The results and discussion of the outcomes are detailed in Section 4. Lastly, Section 5 encompasses concluding remarks.

2. Related Work

2.1. Bayesian Networks for Detecting Preeclampsia

Several ML and deep learning (DL) techniques, such as classification trees, Bayesian networks classifiers (BNCs), neural networks, or random forests, among others, have successfully predicted preeclampsia [17]. Among them, BNCs have the unique ability to help decision-makers determine cause and effect under conditions of uncertainty [18,19]. Therefore, BNCs are considered in this work for representing our classification model.

At this point, it is important to differentiate what a Bayesian network (BN) is from what a Bayesian network classifier (BNC) is. BN is a general-purpose generative model that is trained regardless of the goal for which it is intended. A BNC is a type of BN in which the topology is selected to maximize a single criterion (predictive accuracy) and to perform a single job (classification). In the present work, we will focus on training BNCs to classify the risk of suffering preeclampsia.

In recent years, different BNs have been proposed for analyzing preeclampsia. In Table 1, relevant previous works are listed, describing the authors, the objectives, the techniques considered to build both the network structure and the parameters (conditional

and marginal probabilities), the clinical features considered as nodes in the BN structure, and the aspects that the authors regarded as future work.

Table 1. Review of previous studies related to BN for analyzing preeclampsia, ordered by year of publication.

Authors/Year	Goal	Techniques/Tools	Clinical Features	Future Work
McLachlan, S., et al. [20] 2024	To model a broadly accurate BN model, capable of describing diagnosis and treatment outcomes using expert clinical knowledge, large publicly available privacy-preserving datasets, and published statistics for a given population.	Combines expert elicitation with knowledge from the literature to build the structure and parameters of the BN model. Further investigation was conducted in cases where the model parameters did not match the expert estimates.	Maternal age, deprivation, multiplicity, parity, ethnicity, BMI, glucose, diabetes, gestation, maternal outcome.	Not mentioned in the manuscript.
Amiri, M. [21] 2023	To build a BN for determining the association between diverse influential factors associated with maternal vitamin D and mode of delivery, generating maternal complications such as preeclampsia and preterm delivery, leading to a higher probability of cesarean section.	Authors employed the Hill–Climbing (HC) algorithm to select the best Directed Acyclic Graph (DAG) corresponding to their data. To fit the parameters of BN, they used Bayesian estimation.	Education, age at first pregnancy, husband's education, husband's job, husband's smoking, age at current pregnancy, residence type, number of children, job, city, vitamin D status, vitamin D at delivery, intervention, type of delivery, preeclampsia, reason for cesarean, birth weight.	Not mentioned in the manuscript.
Moreira, M.W., et al. [5] 2016	To build a system to support intelligent decision, applied to the diagnosis of preeclampsia using a BN.	The network structure and its conditional and marginal probabilities were obtained from medical experts (in the literature).	Headache, epigastric pain, nausea/vomiting, blurring of vision, giddiness, hyperflexia, edema, oliguria, hypertension, proteinuria.	This has only been preliminary work. Future work will consider the evaluation of the network using real cases and the corresponding expert evaluation.
Moreira, M.W., et al. [22] 2016	To build a mobile solution for high-risk pregnancy monitoring using sensor networks. It uses a Naïve Bayes classifier to better identify the severity of hypertension, helping experts in the decision-making process.	The authors modeled the BNC structure manually. In it, they considered two attributes to classify the severity of preeclampsia, namely, blood pressure and proteinuria (in 24 h).	There are three different classifications of blood pressure: normal (less than 139 mmHg systolic), high (140–179 mmHg systolic), and extremely high (180 mmHg systolic). Additionally, there are three categories for proteinuria (within 24 h): absent (no protein in the urine), traces (between 0.3 and 1 g/24 h), and severe (more than 3.5 g/24 h).	Authors propose to carry out a comparative study between the different BN classifiers using discrete variables. They also propose using a larger dataset to improve the performance of the system by increasing precision and sensitivity.
van Meurs, A., et al. [23] 2014	To build a BNC for predicting preeclampsia, by looking at the predictions for normal and preeclamptic pregnancies.	Preeclampsia risk estimation from the model for each patient was compared to preeclampsia development. When data are available in early pregnancy, the model can distinguish between preeclampsia and non-preeclampsia pregnant women and is able to predict a higher risk for the diagnosed patients.	Risk factors: age, BMI, smoking, parity, twin pregnancy, family history of preeclampsia, previous history of preeclampsia, pre-existing vascular disease, preexisting renal disease, antiphospholipid syndrome, diabetes mellitus. Medication and measurements: blood pressure, protein-to-creatinine ratio, serum creatinine, and hemoglobin and medication.	To perform a randomized controlled trial with a larger number of patients to establish this cut-off curve with more accuracy, and to validate the preeclampsia model prospectively. Once validated, the model can assist in early preeclampsia diagnosis and thus allow early treatment of it.
Velikova, M., et al. [24] 2014	To forecast the progression of the preeclampsia disease through a dynamic BN model.	An exploration of the basic causal mechanisms and known interactions of preeclampsia has been carried out to acquire the structure and probabilistic parameters of its proposed dynamic BN.	Antiphospholipid syndrome, parity and history of preeclampsia, chronic hypertension, renal disease, diabetes, family history of preeclampsia, family history of hypertension, family history of diabetes, multiple pregnancy, obesity, maternal age, smoking, hemoglobin, creatinine, blood pressure.	To study the effect of treatment per pregnancy week on the model's performance in relationship to actual patients. To study whether different treatment scenarios can help prevent worsening of the patient's condition.

Table 1. *Cont.*

Authors/Year	Goal	Techniques/Tools	Clinical Features	Future Work
Velikova, M., et al. [25] 2011	To build a dynamic BN model for the at-home time-related development of preeclampsia.	The network structure and its conditional and marginal probabilities were obtained from medical experts (in the literature). The model includes the risk factors and laboratory measurements taken during 10 checkups at 12, 16, 20, 24, 28, 32, 36, 38, 40, and 42 weeks of pregnancy.	Age, smoking, obese, chronic HT, parity–historyPE, hemoglobin, creatinine.	This has only been preliminary work. Clinical data will be used to tune the probability distribution of the model such that it reflects current clinical practice in their location.

From Table 1, we can extract some guidelines to carry out our present work:

- As proposed by Amiri, M., et al. [21], we learn the structure and parameters of the BNC from the data collected, using Bayesian estimation to fit the parameters;
- Following future work outlined by Moreira, M. W., et al. [22], we consider some algorithms and not only Naïve Bayes for learning the BNC structure;
- As in the works listed in Table 1, we consider clinical data related to the following points: personal and family health history, age, and cultural and demographic characteristics;
- Following future work outlined by Velikova, M., et al. [25], clinical data are used to build the BNC structure and parameters so that it reflects current clinical practice in our local context. More specifically, we use clinical data from patients who have received care in the gynecology unit of the "IESS Los Ceibos Hospital" in the city of Guayaquil.

All in all, our work carries out both the learning of the structure and the parameters of BNCs from our data, which were collected retrospectively. These data refer to clinical features related to the following points: personal and family health history, age, and cultural and demographic features. For learning the BNC structure and parameters, we use the "bnclassify" package in R [26]. In addition, because of the promising results achieved in [27–30], our work considers the following algorithms for building the structure of BNCs: "Naïve Bayes" (BN), "Tree Augmented Naïve Bayes Chow-Liu" approach (TAN_{cl}) (both of which were adapted by [31]), and "Forward Sequential Selection and Joining" (FSSJ) (which is a Semi Naïve Bayes with a forward approach to constructive induction [32]). To fit the BNC parameters, we use the Bayesian estimation included by default in the "bnclassify" package [26].

2.2. Pre-Processing of the Data

Filter, wrapper, and embedded methods are the widely used approaches in FSS tasks. The first approach is characterized by assigning values to each feature using a function that generates a ranking. The features that head this ranking are selected and used as inputs to the ML model. This approach presents the lowest computational cost compared to the others [15].

On the other hand, the second approach uses classification algorithms to identify the efficiency of the features by considering performance metrics analyzed during model training such as accuracy. This procedure is performed on a subset of features with value assignment, and then the best subset of features is selected [15].

Finally, the last approach is characterized by the use of learning algorithms to search for the optimal subset of features [33]. But, as previously mentioned, FSS approaches do not eliminate superfluous or unnecessary characteristics from datasets; hence, new approaches to FSS should be researched to address this issue [15,16].

3. Materials and Methods

3.1. Clinical Data

Retrospective medical data have been collected from medical histories of patients treated during the period 2017–2023 at "IESS Los Ceibos" hospital in Guayaquil, Ecuador. Due to its retrospective, non-interventional nature with the use of anonymized data, the requirement of informed consent was waived.

The number of medical records collected from "IESS Los Ceibos" hospital was 1467. A total of 64 baseline features have been categorized. Collected medical features were related to the following points: personal or family health history, age, and cultural and demographic features. All the features of our dataset are categorical type.

The feature to be classified (class) is "Disease1". It has two categorical values: positive or negative, where positive refers to patients that suffer from the disease while negative refers to the opposite case. The distribution of positive–negative values was obtained over the records of the medical dataset, as shown in Table 2. The baseline accuracy percentage is close to 76% when classifying all the records as negative cases.

Table 2. Distribution of positive–negative cases of preeclampsia.

Case	Number of Records
Positive	351
Negative	1116

A very popular metric used to measure performance of classification models is the accuracy [34]. It measures the ratio of properly classified samples to total samples. However, accuracy loses its reliability when the dataset is imbalanced (i.e., there are significantly more samples in one class than in the other classes), since this leads to an overly optimistic assessment of the classifier's capacity on the majority class [34].

According to Table 2, we see that our data are imbalanced, having 76% of records of patients who have not presented preeclampsia. Thus, to face this class imbalance issue, this work also considers the use of the "F1 score" metric when evaluating the performance of classification models because it remains one of the most widespread metrics among researchers for avoiding an overly optimistic assessment of the performance of classification models [34].

3.2. Methodology

Figure 1 shows the steps to be considered in the methodological design, which begins with data processing and ends with the deployment of the BNC that predicts preeclampsia. Each phase of this design is described in detail below.

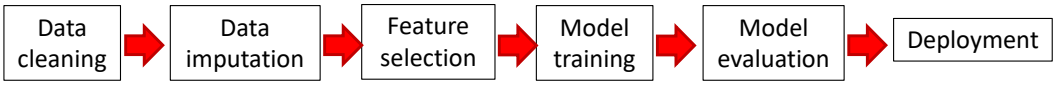

Figure 1. Diagram of the proposed methodology for building a preeclampsia classification model.

3.2.1. Data Cleaning

In this step, the existence of duplicate columns in the dataset is checked. Duplicate columns can be removed. Columns with a unique value are removed too. Both operations have been considered following the recommendations exposed in [35,36].

3.2.2. Data Imputation

In this step, the existence of missing values in the dataset is handled. Missing values can be replaced when applying any technique of data imputation. The majority of imputation techniques are restricted to one class of variables, either categorical or continuous. The various kinds are usually handled independently when dealing with mixed-type data. As a result, these techniques disregard potential connections between different variable types.

Due to that, the present work considers "MissForest", a non-parametric technique that handles several kinds of variables at once [37]. In addition, in contrast with other techniques used for dealing with multiple imputations in electronic health-record data, the machine learning approach called "RandomForest" imputation does not need the specification of a specific regression model and may handle interactions and nonlinearities [38].

All in all, we selected the "MissForest" method and the "RandomForest" algorithm to predict missing values because they admit mixed data, achieving improved performance in classification models according to [39–41].

3.2.3. Feature Selection with NoReFS

A new approach proposed in [14] has been considered to perform the FSS task. We will call it NoReFS: Non-Redundant Feature Selection. It is a combination of filters and wrapper FSS approaches. These two methods were considered in combination since better performances have been reported during the training of classification models in the medical context [42,43]. The filter methods selected are chi-square (chi2), mutual information, and ANOVA F-value classification (F-classif). On the other hand, the selected wrapper method is the Linear Forward Selection since the resulting performance when selecting features can be improved impressively [44].

We proceed to carry out the selection of variables with each of the three aforementioned filter methods. This step produces three subsets, as can be seen in the diagram in Figure 2.

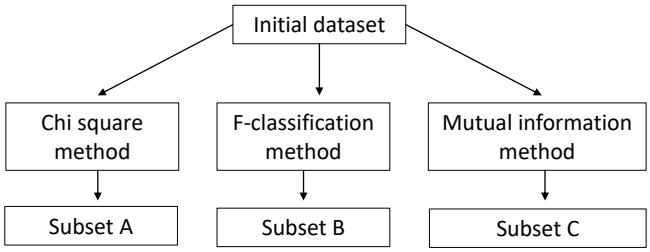

Figure 2. Feature selection with filter methods.

The next step is to extract the best features from each of these subsets (A, B, or C) using Linear Forward Selection (wrapper approach). Through this process, the subsets A', B', and C' are obtained. Then, all the selected variables are joined and compared, and removal of redundant features is applied. Finally, a dataset without redundant variables is achieved. This process is presented in Figure 3.

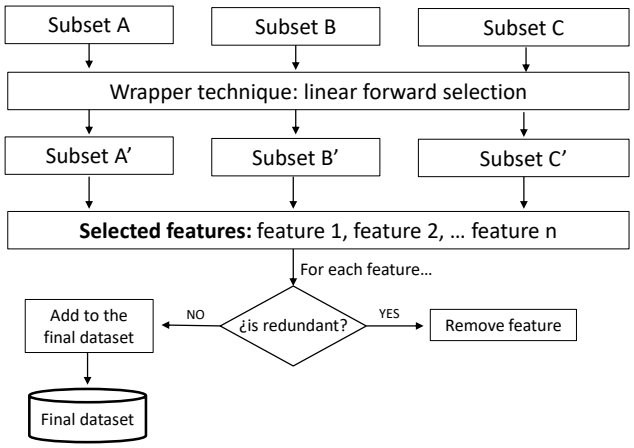

Figure 3. Extraction of the best features and comparison of redundant features.

As presented in [14], using this methodology (NoReFS) produces comparable outcomes to models trained with other FSS techniques, with the added benefit of eliminating redundant features from the dataset.

3.2.4. Model Training

The algorithms selected for building BNCs are "Naïve Bayes" (NB), the Chow–Liu approach of "Tree Augmented Naïve Bayes" (TAN_{cl}) adopted by [31], and a Semi Naïve Bayes with a forward approach to constructive induction that is called "Forward Sequential Selection and Joining" (FSSJ), as defined by [32]. Those algorithms have been selected because of their good performances achieved in [27–30].

Regarding the percentage of data used for training and testing, it is recommended to increase the percentage to 80% or 90% for training when one has larger datasets. The 80:20 training/test split ratio (80% for training and 20% for testing) is recommended, especially for larger datasets, to provide enough training samples [45].

On the other hand, by decreasing the split ratio from 80:20 to 60:40, the amount of data used for training is reduced. Moreover, using a ratio of 60:40 or 70:30 allows us to have a clearer idea of the performance of the classification model on the test data, even if the training task is carried out on a smaller number of samples. However, in the work presented by [45], a significant difference between the 60:40 and 70:30 split ratios is not observed.

In this work, the entire dataset is divided into train and test sets when following the 70:30 ratio. This implies that the first 70% of all data is assigned as the training set and the remaining 30% as the test set. Those percentages were selected following the recommendations presented in [46–49].

3.2.5. Model Evaluation

Once the preeclampsia classification models have been trained with each algorithm presented in the previous step, the performances of the models are compared. To carry out this comparison, we consider the metrics of accuracy, specificity, and F1 Score since they have been widely adopted to measure the performance in binary classification tasks [34]. The sensitivity measures the fraction of positive cases that are classified as positive, while the specificity measures the fraction of negative cases classified as negative. In our case, the positive values are the patients who have a positive case of preeclampsia, while the negative cases are the ones without that disease.

To carry out an honest estimation of the performance, a 10-fold cross-validation will be considered on the training set. Moreover, to compare the performance of classification models on unseen data, the test set will be utilized. Between them we will select the one with the highest accuracy.

3.2.6. Deployment

The BNC obtained in the previous step was exported to a DSC format file to be opened with "Bayesfusion" [50] Version 3.0.6518.0 (also called "GeNIe Academic"), with the purpose of facilitating the visualization and management of the network by the doctors who work at the "IESS Los Ceibos" hospital. Bayesfusion has been selected because this tool is useful for performing inferences in BNs and other types of graphical probabilistic models. Moreover, models created in it can be easily shared and used on mobile devices or through a web browser [51]. Finally, the process of setting evidence in the BN to obtain the classification in a BNC model is quite simple: it is only necessary to click on the corresponding category value of each node.

4. Results and Analysis

4.1. Data Cleaning, Imputation, and Feature Selection

In this step, we have proceeded to perform the data cleaning and data imputation following the steps described in Sections 3.2.1 and 3.2.2. After that, we proceeded to

perform the FSS task with the NoReFS methodology presented in Section 3.2.3. Table 3 shows the selected features when applying the NoReFS method.

Table 3. Selected features when applying the NoReFS approach.

Feature	Description	Labels
"Hypertensionpersonalhistory"	Hypertension personal history	yes/no
"Parity"	The number of times the fetus has reached a viable gestational age	1/2/3/4/5/6/7 or more
"Gravidity"	The number of times the woman has been pregnant	1/2/3/4/5/6/7 or more
"Fetalstatus"	Previous fetal status at birth	born alive/stillborn/NA
"Tobaccouse"	Tobacco use	yes/no
"Diabetesfamilyhistory"	Existence of relatives with diabetes	yes/no
"Nupucells1"	Patient vaginal infection	mild/moderate/severe
"Maternalage-categorized"	Maternal age by ranges	State0: <35/State1: $\geq$35
"Education_Level"	Education level	primary/secondary/tertiary
"Specificplacearealivedincountyof"	Area where the patient resides	urban/rural

In order to verify if the selected features are in accordance with expert knowledge, we comment on what the medical literature published to date has found about each of them.

The "Hypertensionpersonalhistory" feature is strongly related to the disease, due to the fact that the increase in blood pressure during pregnancy causes an increased risk of developing the disease [52]. In addition, the "Parity" and "Gravidity" features are also marked as risk factors in [10,53] because having a high number of these implies having a greater chance of developing preeclampsia. Concerning the state of the previous fetus, or "fetalstatus", the medical literature mentions that having a previous stillbirth episode may increase the risk of preterm delivery [54,55]. Also, having previous babies with low birth weight has a higher risk of developing the disease [56]. About tobacco use or "tobaccouse", the medical literature associates it with a higher risk of developing complications during pregnancy, including preeclampsia [57]. The history of diabetes or "diabetesfamilyhistory" is of great importance in the incidence of the disease because women with a history of diabetes have a higher probability of developing the disease [58]. In addition, the patient's vaginal infection or "nupucells1" can provoke an abnormal immune response in the mother, increasing the risk of developing preeclampsia [59,60].

In addition to the aforementioned medical features, age, cultural, and demographic features are also important factors in preeclampsia risk prediction. It is essential to know the maternal age or "maternalage-categorized" of the patient who is pregnant because, when the patient is older than 35 years, they have a greater risk of developing the disease [61]. Moreover, women with a low educational level or "Education_Level" have a higher risk of developing the disease because it is correlated with having less access to medical care and health services, which increases the risk of developing the disease [62]. Finally, the place where the patient lives or "specificplacearealivedincountyof" can greatly influence the disease since, if the patient lives in a rural area, they have less access to medical care and health services. This can increase the risk of developing the disease [63].

4.2. Model Training

As mentioned in Section 3.2.4, the algorithms selected to built our BNCs are: "Naïve Bayes" (NB), "Tree Augmented Naïve Bayes" (TAN_{cl}), and "Semi Naïve Bayes" (FSSJ). Moreover, we will consider "Disease1" as the class to predict. In addition, we have followed the train-to-test ratio of 70:30, as recommended in [47]. For comparison purposes, we carry out model training with and without the selected features in Table 3.

Below are the networks built with NB, TAN_{cl} and FSSJ. Since the NB and TAN_{cl} algorithms make use of all the features of the dataset to build their node connections, making it difficult to visualize their models, we only present those models trained with the selected features from Table 3 in Figures 4–6. In addition, we present in Figure 7 the network obtained with the FSSJ but without applying our proposed NoReFS approach (Section 3.2.3).

In Figure 4, we show that the model trained with the NB algorithm connects the feature to be classified (Disease1) with all the other medical features without taking into account the relationships that may exist between those other features. This is due to the NB assumption of conditional independence given the class. While this assumption of independence is often violated in practice, the NB nevertheless often offers competitive classification accuracy [29].

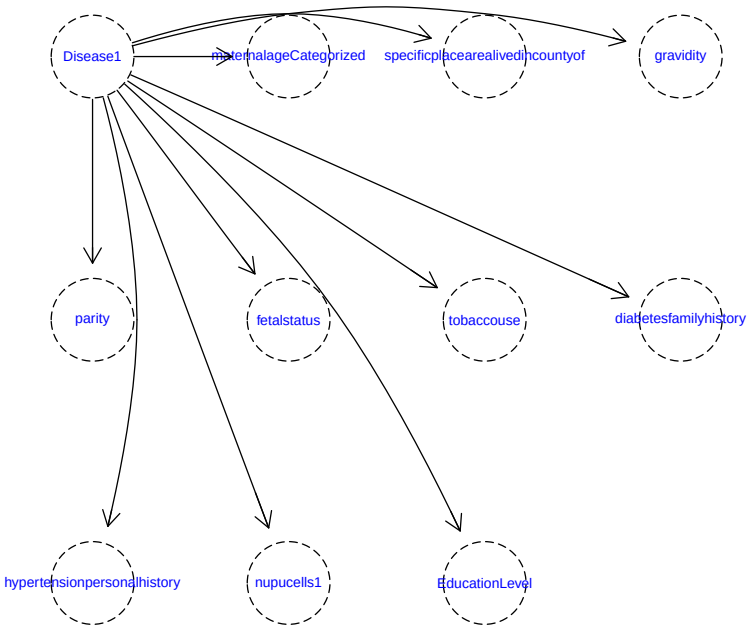

Figure 4. Model trained with NB and the selected features from Table 3.

In Figure 5, the model trained with TAN_{cl} has learned the relationships between the features, representing them with connections in the form of a tree. Some medical features are not connected but only with the class (Disease1). These are "specificplacearealivedincountyof", "diabetesfamilyhistory", and "nupucells1". Regarding the connected features, we can review whether their causal influences are in accordance with medical knowledge. For example, we show a causal influence between "maternalage-categorized" and "gravity". This may be because a higher age of pregnancy usually implies a higher number of previous pregnancies. There is also a causal influence between "gravity" and "Education_Level". A probable explanation for this relationship is that having a large number of children could imply a lack of knowledge of contraceptive methods and other aspects related to human sexuality (aspects that are learned in educational units). Additionally, a causal influence between "Education_Level" and "hypertensionpersonalhistory" is represented in the model. This could imply that a low educational level usually generates a lack of knowledge about healthcare, with hypertension as one of the consequences, along with other ailments. Furthermore, a causal influence between "gravity" and "parity" is described in the model. This may be because a high number of pregnancies often implies a high number of births. Finally, the model describes a causal influence between "parity" and "fetalstatus". We consider that this relationship is obvious since birth implies by itself that the baby is born alive.

It is important to note that we do not see any medical explanation for the causal influence between "hypertensionpersonalhistory" and "tobaccouse" represented in Figure 5 since the first does not cause the second.

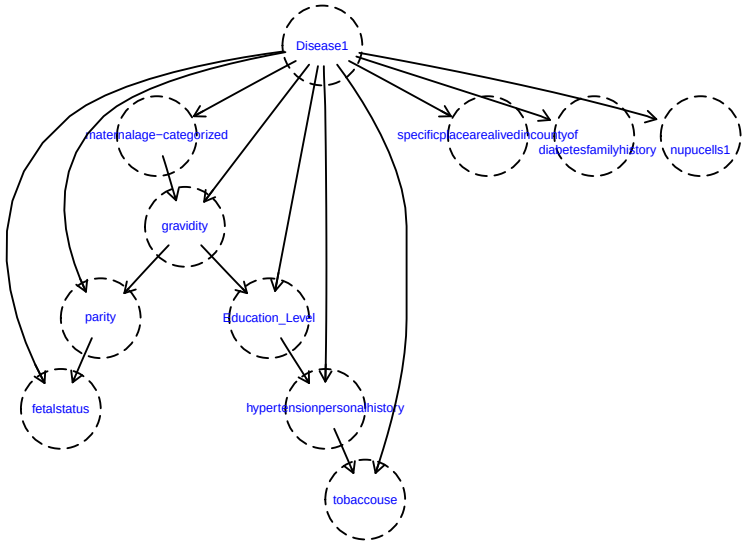

Figure 5. Model trained with TAN_{cl} and the selected features from Table 3.

In Figure 6, the model trained with the selected features of Table 3 and the FSSJ algorithm has learned only one relationship between "hypertensionpersonalhistory" and the class (Disease1). This unique relationship is because the FSSJ algorithm performs a feature selection itself, as described in [32]. In this, the author mentions that the FSSJ algorithm initializes the set of features to be used by the classifier for the empty set. Next, two operators are used to add features to the model until no improvement is found. The first operation consists of adding each feature not used by the current model representation. The second operation consists of joining each feature not used by the current model with each feature currently used.

The only relationship represented in the model is the causal influence between "Disease1" and "hypertensionpersonalhistory". That causal relationship is abundantly documented in the medical literature [52], with hypertension being a risk factor for developing preeclampsia, which, according to its definition, is a hypertensive disorder developed in pregnancy.

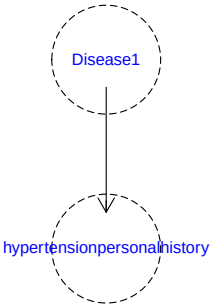

Figure 6. Model trained with FSSJ and the selected features of Table 3.

Because FSSJ intrinsically applies a selection of features, Figure 7 is presented. In it, a BNC has been trained when using the FSSJ algorithm but without applying the previous selection of features presented in Table 3.

Exploring the existing relationships in the model in Figure 7, we see that some of its medical features were not selected when applying our NoReFS approach. These are age at first marriage ("ageatfirstmarriage-categorized"), hemoglobin level on admission ("hemoglobinlevelonadmissionfordel-categorized2"), and the time trimester of first antenatal care visit ("timetrimesteroffirstancvisit"). Those medical features have also been pointed out as risk predictors by the medical literature. For example, it is important to consider the age at first marriage because uterine immaturity in very young teenagers is likely a major

cause of defective deep placentation and adverse reproductive outcomes such as adolescent preeclampsia [64]. In addition, a lower hemoglobin level on admission is described as one of the symptoms of severe preeclampsia by [65]. Finally, the time trimester of the first antenatal care visit is important to consider because, in [66], authors found that women who worried about poor health in early pregnancy, controlling for other relevant health indicators, were 2 to 3 times more likely to develop preeclampsia.

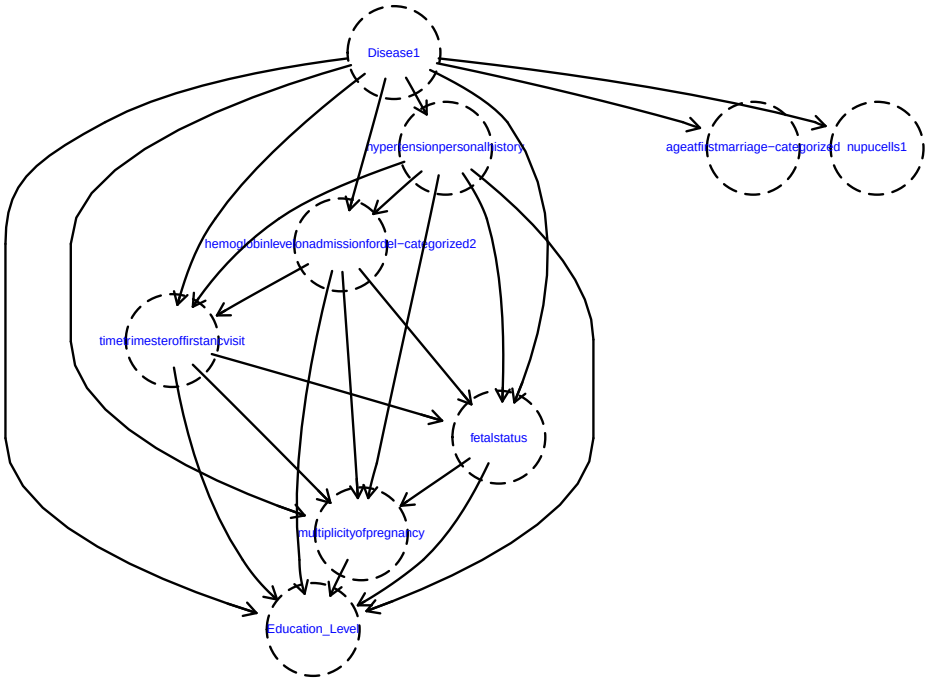

Figure 7. Model trained with FSSJ without applying a previous FSS task.

All in all, some BNCs have been trained. Almost all the causal relationships existing in these models are corroborated by the medical literature. To determine the goodness of the models presented in this section and select the one with the best performance, it is necessary to use the test set and an honest estimation of the performance metrics, aspects that are described in the next section.

4.3. Model Evaluation

4.3.1. Honest evaluation of classification models

With the purpose of having an honest estimation of performance, a 10-fold cross-validation on training dataset was considered. The performance of BNC models when applying our proposed FSS approach and without it are shown in Table 4.

Table 4. Performance of BNCs when using 10-fold cross-validation and with/without our proposed FSS approach. The best results are highlighted in bold.

NoReFS	Algorithm	Performance Values (Mean ± Std Deviation)			
		Accuracy	**Sensitivity**	**Specificity**	**F1 Score**
NO	NB	76.12% ± 3.74%	0.72 ± 0.07	0.79 ± 0.04	74.05% ± 2.10%
	TAN_{cl}	82.62% ± 3.52%	0.79 ± 0.04	0.87 ± 0.04	77.32% ± 2.12%
	FSSJ	80.31% ± 2.07%	0.78 ± 0.05	0.83 ± 0.02	78.29% ± 1.37%
YES	NB	78.24% ± 4.16%	0.74 ± 0.02	0.85 ± 0.02	75.19% ± 3.25%
	TAN_{cl}	**89.64% ± 3.78%**	**0.87 ± 0.02**	**0.91 ± 0.03**	**84.45% ± 1.92%**
	FSSJ	86.12% ± 3.52%	0.83 ± 0.04	0.88 ± 0.04	82.17% ± 2.12%

As can be shown in Table 4, the NoReFS approach has slightly improved the performance in comparison with the achieved without considering the NoReFS method. This can be observed by comparing the accuracy, specificity, and F1 score metrics. Moreover, the BNC trained with the TAN_{cl} algorithm has achieved the best results, with an accuracy close to 90%. In addition, its high sensitivity and specificity values indicate that this model achieves good predictions of positive and negative cases of preeclampsia.

4.3.2. Performance of classification models on unseen data

For comparing the goodness of trained models on unseen data, the test set was used, and the metrics mentioned in Section 3.2.5 are considered. The performances that BNCs have achieved on the test set are presented in Table 5.

Table 5. Performance of BNCs when using the test dataset. The best results are highlighted in bold.

NoReFS	Algorithm	Performance Values			
		Accuracy	Sensitivity	Specificity	F1 Score
NO	NB	75.72%	0.75	0.77	73.17%
	TAN_{cl}	82.41%	0.78	0.85	79.64%
	FSSJ	79.52%	0.76	0.82	76.18%
YES	NB	77.59%	0.73	0.84	74.38%
	TAN_{cl}	**88.71%**	**0.86**	**0.92**	**83.17%**
	FSSJ	85.43%	0.82	0.86	81.60%

According to the results in Table 5, models that considered the NoReFS approach outperform the results achieved by the models that do not consider it. This can be observed by comparing the accuracy, specificity, and F1 score metrics. In addition, the BNC trained with the TAN_{cl} algorithm achieved the best performance, with an accuracy close to 90%. Moreover, its high sensitivity and specificity values indicate that this model achieves good predictions of positive and negative cases of preeclampsia.

All in all, the BNC trained with the TAN_{cl} algorithm has been proven to obtain the best performance, in both Tables 4 and 5, when using the NoReFS approach. TAN_{cl} has also achieved outstanding results when training classification models in [27,67–69].

4.4. Deployment

Due to the high performance achieved when comparing BNCs in the previous section, the model trained with TAN_{cl} was selected to be deployed through the Bayesfusion software, as mentioned in Section 3.2.6. It is depicted in Figure 8.

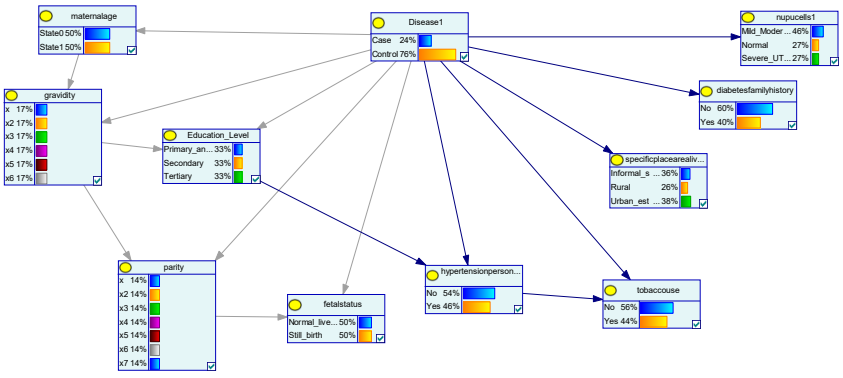

Figure 8. BNC trained with TAN_{cl} and the selected features from Table 3.

In Figure 8, we observe that the marginal probability distribution of each feature has similar probabilities for their categories, with the exceptions of "hypertensionpersonalhistory", "tobaccouse", "diabetesfamilyhistory", "nupucells1", and "specificplacearelivealivedincountyof", which have a strong influence on the classification of the risk of preeclampsia. Thus, the other features have a minimal influence on the classification.

As mentioned in Section 3.2.6, to perform the classification, the doctors must click on the corresponding category value of the features, depending on the information available about the patient. This operation is called "setting the evidence". Then, by updating the belief, the high percentage of the category value in the target feature (Disease1) determines if there is a positive case of the disease or if it is a negative case. For example, Figure 9 presents the classification obtained with the following evidence: a patient over 35 years old, with a severe vaginal infection (nupucells1) and with diabetes family history, resides in any rural location, is positive to tobacco use, is positive to hypertension personal episodes, has a secondary education level, has gravidity equal to 3, and has parity equal to 2. In this case, the patient with these values has a high probability (86%) of suffering preeclampsia according to the trained TAN_{cl} classifier in Figure 5. In this case, the risk of developing preeclampsia may be due to hypertension personal history and tobacco use.

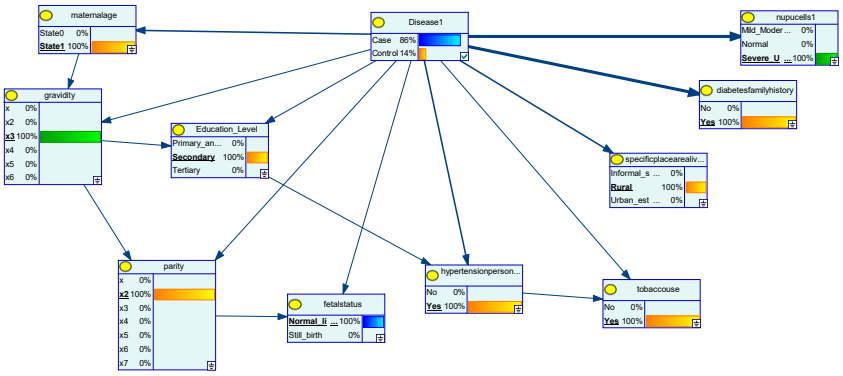

Figure 9. Setting the evidence in the BNC trained with TAN_{cl} and the selected features from Table 3.

All in all, patients who are at high risk of developing preeclampsia are those whose age is above 35 years, have a severe vaginal infection (nupucells1), live in a rural area, use tobacco, have a family history of diabetes, and have a personal history of hypertension. These features were extracted by setting the "Disease1" feature to "Case" and updating the evidence, as presented in Figure 10.

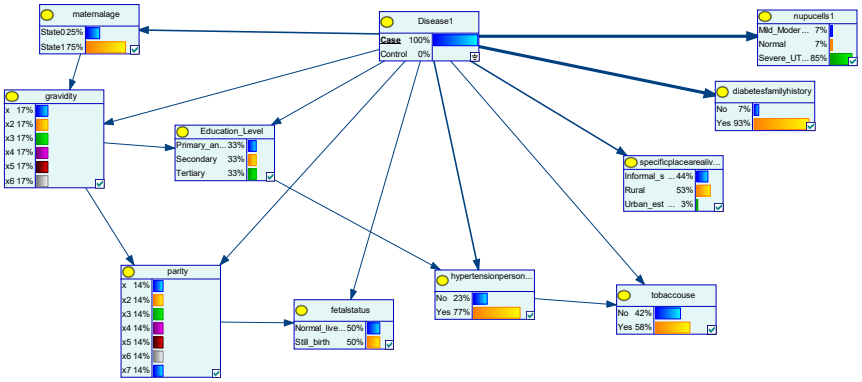

Figure 10. Patients with a high risk of suffering preeclampsia in the BNC trained with TAN_{cl} and the selected features from Table 3.

5. Conclusions

This study assesses the application of some algorithms to train Bayesian network classifiers (BNCs) that predict the risk of suffering preeclampsia in patients who are treated at the "IESS Los Ceibos" hospital in Guayaquil, Ecuador. In this work, a non-redundant feature selection (NoReFS) approach is proposed for handling the elimination of redundant features, which is not assured when using only a filter or a wrapper FSS approach.

Ten medical features were selected by the NoReFS approach. Features such as hypertension personal history, parity, gravidity, tobacco use, diabetes family history, maternal age, and education level, have also been described as important clinical features to consider for predicting preeclampsia in the medical literature. This knowledge allows us to conclude that the features selected by the NoReFS approach are coherent with regard to the medical literature.

BNCs were built with the selected features, improving the performance obtained when not applying the NoReFS task before the model training. Naïve Bayes (NB), Chow–Liu Tree-Augmented Naïve Bayes (TAN_{cl}), and Semi Naïve Bayes (FSSJ) algorithms were considered for building BNCs. The model trained with the TAN_{cl} algorithm and the NoReFS task was the best of them, achieving an accuracy close to 90%. In addition, the medical interpretation of the causal influence relationships in the classifying models has been carried out, usually agreeing with what the medical literature has mentioned to date.

Due to the high performance achieved by the model trained with the TAN_{cl} algorithm and the NoReFS task, it was selected to be deployed with Bayesfusion to carry out the classification of the risk of preeclampsia and interpretation of the results by the medical personnel working in the department of gynecology at the hospital "IESS Los Ceibos" in Guayaquil, Ecuador. From the deployed model, we can infer that the patients with the highest risk of suffering from preeclampsia are those whose age is above 35 years, have a severe vaginal infection, live in a rural area, use tobacco, have a family history of diabetes, and have a personal history of hypertension.

As future work, it remains to contrast the clinical–computational findings of the classification models presented in this work with other classification models built from other clinical data available in different hospitals in Ecuador. Furthermore, since this study has addressed the use of an unbalanced dataset, we may consider some data augmentation techniques such as the Synthetic Minority Oversampling Technique (SMOTE) and some ensemble methods such as boosting or bagging combined with BNCs to improve the robustness and accuracy of the model, especially in the presence of dataset imbalance. Finally, in the future, it is proposed to collect other clinical information from patients that may be associated with the risk of preeclampsia, such as body mass index and physical activity, among others.

Author Contributions: All the authors have contributed to the work presented in this paper. Conceptualization, F.P.-B.; methodology, F.P.-B.; validation, F.P.-B., R.C.-Q., E.R.-L. and J.B.-M.; investigation, F.P.-B.; writing—original draft preparation, F.P.-B.; writing—review and editing, F.P.-B., R.C.-Q., E.R.-L. and J.B.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to the corresponding author.

Acknowledgments: The authors also want to express their gratitude to the Service of Gynecology at the Hospital "IESS Los Ceibos" in Guayaquil, Ecuador, whose help has been crucial for this work—in particular, Freddy Villón-López and Priscilla Alcócer-Cordero.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ukah, U.V.; Payne, B.; Hutcheon, J.A.; Ansermino, J.M.; Ganzevoort, W.; Thangaratinam, S.; Magee, L.A.; Von Dadelszen, P. Assessment of the fullPIERS risk prediction model in women with early-onset preeclampsia. *Hypertension* **2018**, *71*, 659–665. [CrossRef]
2. Parrales-Bravo, F.; Saltos-Cedeño, J.; Tomalá-Esparza, J.; Barzola-Monteses, J. Clustering-based Approach for Characterization of Patients with Preeclampsia using a Non-Redundant Feature Selection. In Proceedings of the 2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Tenerife, Spain, 19–21 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–6.
3. Koulouraki, S.; Paschos, V.; Pervanidou, P.; Christopoulos, P.; Gerede, A.; Eleftheriades, M. Short- and Long-Term Outcomes of Preeclampsia in Offspring: Review of the Literature. *Children* **2023**, *10*, 826. [CrossRef] [PubMed]
4. Muldoon, K.A.; McLean, C.; El-Chaár, D.; Corsi, D.J.; Rybak, N.; Dagvadorj, A.; Guo, Y.; White, R.R.; Dingwall-Harvey, A.L.J.; Gaudet, L.M.; et al. Persisting risk factors for preeclampsia among high-risk pregnancies already using prophylactic aspirin: a multi-country retrospective investigation. *J. Matern.-Fetal Neonatal Med.* **2023**, *36*, 2200879. [CrossRef] [PubMed]
5. Moreira, M.W.; Rodrigues, J.J.; Oliveira, A.M.; Ramos, R.F.; Saleem, K. A preeclampsia diagnosis approach using Bayesian networks. In Proceedings of the 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, 22–27 May 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–5.
6. Bisson, C.; Dautel, S.; Patel, E.; Suresh, S.; Dauer, P.; Rana, S. Preeclampsia pathophysiology and adverse outcomes during pregnancy and postpartum. *Front. Med.* **2023**, *10*, 1144170. [CrossRef] [PubMed]
7. ACOG. Gestational Hypertension and Preeclampsia: ACOG Practice Bulletin. *Obstet. Gynecol.* **2020**, *135*, e237–e260. [CrossRef] [PubMed]
8. Chang, K.J.; Seow, K.M.; Chen, K.H. Preeclampsia: Recent Advances in Predicting, Preventing, and Managing the Maternal and Fetal Life-Threatening Condition. *Int. J. Environ. Res. Public Health* **2023**, *20*, 2994. [CrossRef] [PubMed]
9. Ministerio de Salud Pública del Ecuador. Gaceta de Muerte Materna SE14. **2020**. Available online: https://bit.ly/3Poz79o (accessed on 28 March 2022).
10. De Kat, A.C.; Hirst, J.; Woodward, M.; Kennedy, S.; Peters, S.A. Prediction models for preeclampsia: A systematic review. *Pregnancy Hypertens.* **2019**, *16*, 48–66. [CrossRef] [PubMed]
11. Rambaldi, M.P.; Weiner, E.; Mecacci, F.; Bar, J.; Petraglia, F. Immunomodulation and preeclampsia. *Best Pract. Res. Clin. Obstet. Gynaecol.* **2019**, *60*, 87–96. [CrossRef] [PubMed]
12. Rolnik, D.L.; Nicolaides, K.H.; Poon, L.C. Prevention of preeclampsia with aspirin. *Am. J. Obstet. Gynecol.* **2020**, *226*, S1108–S1119. [CrossRef]
13. Marić, I.; Tsur, A.; Aghaeepour, N.; Montanari, A.; Stevenson, D.K.; Shaw, G.M.; Winn, V.D. Early prediction of preeclampsia via machine learning. *Am. J. Obstet. Gynecol. MFM* **2020**, *2*, 100100. [CrossRef]
14. Parrales-Bravo, F.; Torres-Urresto, J.; Avila-Maldonado, D.; Barzola-Monteses, J. Relevant and Non-Redundant Feature Subset Selection Applied to the Detection of Malware in a Network. In Proceedings of the 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM), Cuenca, Ecuador, 12–15 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
15. Gopika, N.; ME, A.M.K. Correlation based feature selection algorithm for machine learning. In Proceedings of the 2018 3rd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 15–16 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 692–695.
16. Venkatesh, B.; Anuradha, J. A review of feature selection and its methods. *Cybern. Inf. Technol.* **2019**, *19*, 3–26. [CrossRef]
17. Aljameel, S.S.; Alzahrani, M.; Almusharraf, R.; Altukhais, M.; Alshaia, S.; Sahlouli, H.; Aslam, N.; Khan, I.U.; Alabbad, D.A.; Alsumayt, A. Prediction of preeclampsia using machine learning and deep learning models: A review. *Big Data Cogn. Comput.* **2023**, *7*, 32. [CrossRef]
18. Mihaljević, B.; Bielza, C.; Larrañaga, P. Bayesian networks for interpretable machine learning and optimization. *Neurocomputing* **2021**, *456*, 648–665. [CrossRef]
19. Kyrimi, E.; McLachlan, S.; Dube, K.; Neves, M.R.; Fahmi, A.; Fenton, N. A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future. *Artif. Intell. Med.* **2021**, *117*, 102108. [CrossRef]
20. McLachlan, S.; Daley, B.; Saidi, S.; Kyrimi, E.; Dube, K.; Grossan, C.; Neil, M.; Rose, L.; Fenton, N. Approach and Method for Bayesian Network Modelling: A Case Study in Pregnancy Outcomes for England and Wales. *medRxiv* **2024**.
21. Amiri, M.; Rostami, M.; Sheidaei, A.; Fallahzadeh, A.; Ramezani Tehrani, F. Mode of delivery and maternal vitamin D deficiency: An optimized intelligent Bayesian network algorithm analysis of a stratified randomized controlled field trial. *Sci. Rep.* **2023**, *13*, 8682. [CrossRef]
22. Moreira, M.W.; Rodrigues, J.J.; Oliveira, A.M.; Saleem, K. Smart mobile system for pregnancy care using body sensors. In Proceedings of the 2016 International Conference on Selected Topics in Mobile & Wireless Networking (MoWNeT), Cairo, Egypt, 11–13 April 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–4.
23. van Meurs, A.; Velikova, M.; van der Hout, B.; Vermeulen-Giovagnoli, B.; Oei, S. Prediction of pre-eclampsia by maternal characteristics: A case-controlled validation study of a Bayesian network model for risk identification of pre-eclampsia. *J. Matern. Fetal Neonatal Med.* **2014**, *27*, 351–352.
24. Velikova, M.; Van Scheltinga, J.T.; Lucas, P.J.; Spaanderman, M. Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare. *Int. J. Approx. Reason.* **2014**, *55*, 59–73. [CrossRef]

25. Velikova, M.; Lucas, P.J.; Spaanderman, M. A predictive Bayesian network model for home management of preeclampsia. In Proceedings of the Conference on Artificial Intelligence in Medicine in Europe, Bled, Slovenia, 2–6 July 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 179–183.
26. Mihaljevic, B.; Bielza Lozoya, M.C.; Larrañaga Múgica, P.M. bnclassify: Learning Bayesian network classifiers. *R J.* **2018**, *10*, 455–468. [CrossRef]
27. Park, H.; Hwang, B.S. The performance of Bayesian network classifiers for predicting discrete data. *Korean J. Appl. Stat.* **2020**, *33*, 309–320.
28. Fauziyyah, N.; Abdullah, S.; Nurrohmah, S. Reviewing the consistency of the Naïve Bayes Classifier's performance in medical diagnosis and prognosis problems. In Proceedings of the 5th International Symposium on Current Progress in Mathematics and Sciences (ISCPMS2019), Depok, Indonesia, 9–10 July 2019; AIP Publishing LLC: Melville, NY, USA, 2020; Volume 2242, p. 030019.
29. Wickramasinghe, I.; Kalutarage, H. Naive Bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation. *Soft Comput.* **2021**, *25*, 2277–2293. [CrossRef]
30. Rivas, J.J.; Orihuela-Espina, F.; Sucar, L.E. Recognition of affective states in virtual rehabilitation using late fusion with Semi-Naive Bayesian classifier. In Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare, Trento Italy, 20–23 May 2019; pp. 308–313.
31. Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian network classifiers. *Mach. Learn.* **1997**, *29*, 131–163. [CrossRef]
32. Pazzani, M.J. Constructive induction of Cartesian product attributes. In *Feature Extraction, Construction and Selection: A Data Mining Perspective*; Springer Science & Business Media: Berlin, Germany, 1998; pp. 341–354.
33. Spasova Dimitrova, R. Desarrollo y evaluación de métodos de selección de características para la predicción de eventos adversos en pacientes polimedicados. Universidad Pública de Navarra. **2017**. Available online: https://hdl.handle.net/2454/24594 (accessed on 8 March 2024).
34. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 1–13. [CrossRef] [PubMed]
35. Bravo, F.P.; García, A.A.; Russo, L.; Ayala, J.L. SOFIA: Selection of Medical Features by Induced Alterations in Numeric Labels. *Electronics* **2020**, *9*, 1492. [CrossRef]
36. Bravo, F.P.; García, A.A.D.B.; Veiga, A.B.G.; De La Sacristana, M.M.G.; Piñero, M.R.; Peral, A.G.; Džeroski, S.; Ayala, J.L. SMURF: Systematic Methodology for Unveiling Relevant Factors in retrospective data on chronic disease treatments. *IEEE Access* **2019**, *7*, 92598–92614. [CrossRef]
37. Stekhoven, D.J.; Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef] [PubMed]
38. Shah, A.D.; Bartlett, J.W.; Carpenter, J.; Nicholas, O.; Hemingway, H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *Am. J. Epidemiol.* **2014**, *179*, 764–774. [CrossRef]
39. Arias-Muñoz, A.C. Propuesta y evaluación de una estrategia para la imputación múltiple y multivariada de valores faltantes en series de tiempo del campo meteorológico utilizando aprendizaje automático= Proposal and evaluation of a strategy for multiple and multivariate imputación of missing values in time series of the meteorological field using machine learning. Instituto Tecnológico de Costa Rica. **2022**. Available online: https://hdl.handle.net/2238/14060 (accessed on 8 March 2024).
40. Alkabbani, H.; Ramadan, A.; Zhu, Q.; Elkamel, A. An improved air quality index machine learning-based forecasting with multivariate data imputation approach. *Atmosphere* **2022**, *13*, 1144. [CrossRef]
41. Zhang, S.; Gong, L.; Zeng, Q.; Li, W.; Xiao, F.; Lei, J. Imputation of gps coordinate time series using missforest. *Remote Sens.* **2021**, *13*, 2312. [CrossRef]
42. Párraga-Valle, J.; García-Bermúdez, R.; Rojas, F.; Torres-Morán, C.; Simón-Cuevas, A. Evaluating mutual information and chi-square metrics in text features selection process: A study case applied to the text classification in PubMed. In Proceedings of the Bioinformatics and Biomedical Engineering: 8th International Work-Conference, IWBBIO 2020, Granada, Spain, 6–8 May 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 636–646.
43. Mukherjee, S.; Dutta, S.; Mitra, S.; Pati, S.K.; Ansari, F.; Baranwal, A. Ensemble Method of Feature Selection Using Filter and Wrapper Techniques with Evolutionary Learning. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2022, Kolkata, India, 23–25 February 2022*; Springer: Berlin/Heidelberg, Germany, 2022; Volume 2, pp. 745–755.
44. Di Mauro, M.; Galatro, G.; Fortino, G.; Liotta, A. Supervised feature selection techniques in network intrusion detection: A critical review. *Eng. Appl. Artif. Intell.* **2021**, *101*, 104216. [CrossRef]
45. Rácz, A.; Bajusz, D.; Héberger, K. Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. *Molecules* **2021**, *26*, 1111. [CrossRef] [PubMed]
46. Talukdar, J.; Gogoi, D.K.; Singh, T.P. A comparative assessment of most widely used machine learning classifiers for analysing and classifying autism spectrum disorder in toddlers and adolescents. *Healthc. Anal.* **2023**, *3*, 100178. [CrossRef]
47. Nguyen, Q.H.; Ly, H.B.; Ho, L.S.; Al-Ansari, N.; Le, H.V.; Tran, V.Q.; Prakash, I.; Pham, B.T. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Math. Probl. Eng.* **2021**, *2021*, 1–15. [CrossRef]
48. Stiawan, D.; Idris, M.Y.B.; Bamhdi, A.M.; Budiarto, R. CICIDS-2017 dataset feature analysis with information gain for anomaly detection. *IEEE Access* **2020**, *8*, 132911–132921.
49. Fallucchi, F.; Coladangelo, M.; Giuliano, R.; William De Luca, E. Predicting employee attrition using machine learning techniques. *Computers* **2020**, *9*, 86. [CrossRef]

50. BayesFusion, L. BayesFusion Modeler. User Manual. 2023. Available online: https://support.bayesfusion.com/docs/ (accessed on 16 May 2023).
51. BayesFusion, L. Welcome to BayesFusion Website. BayesFusion, LLC. **2023**. Available online: https://www.bayesfusion.com/ (accessed on 8 March 2024).
52. Singla, M.; Singla, K.B.; Mewada, B.; Bhalodia, K.; Gandhi, L.M. Risk Factors Associated with Preeclampsia: A Case Control Study. *Eur. J. Mol. Clin. Med.* **2022**, *9*, 2367–2374.
53. Maeda, Y.; Kaneko, K.; Ogawa, K.; Sago, H.; Murashima, A. The effect of parity, history of preeclampsia, and pregnancy care on the incidence of subsequent preeclampsia in multiparous women with SLE. *Mod. Rheumatol.* **2021**, *31*, 843–848. [CrossRef] [PubMed]
54. Coban, U.; Takmaz, T.; Unyeli, O.D.; Ozdemir, S. Adverse outcomes of preeclampsia in previous and subsequent pregnancies and the risk of recurrence. *Med. Bull. Sisli Etfal Hosp.* **2021**, *55*, 426. [CrossRef] [PubMed]
55. Graham, N.; Stephens, L.; Johnstone, E.D.; Heazell, A.E. Can information regarding the index stillbirth determine risk of adverse outcome in a subsequent pregnancy? Findings from a single-center cohort study. *Acta Obstet. Gynecol. Scand.* **2021**, *100*, 1326–1335. [CrossRef] [PubMed]
56. Ngwenya, S.; Jones, B.; Mwembe, D.; Nare, H.; Heazell, A.E. The prevalence of and risk factors for stillbirths in women with severe preeclampsia in a high-burden setting at Mpilo Central Hospital, Bulawayo, Zimbabwe. *J. Perinat. Med.* **2022**, *50*, 678–683. [CrossRef]
57. Holme, J.A.; Valen, H.; Brinchmann, B.C.; Vist, G.E.; Grimsrud, T.K.; Becher, R.; Holme, A.M.; Øvrevik, J.; Alexander, J. Polycyclic aromatic hydrocarbons (PAHs) may explain the paradoxical effects of cigarette use on preeclampsia (PE). *Toxicology* **2022**, *473*, 153206. [CrossRef]
58. Kay, V.R.; Wedel, N.; Smith, G.N. Family history of hypertension, cardiovascular disease, or diabetes and risk of developing preeclampsia: A systematic review. *J. Obstet. Gynaecol. Can.* **2021**, *43*, 227–236. [CrossRef] [PubMed]
59. Lin, C.Y.; Lin, C.Y.; Yeh, Y.M.; Yang, L.Y.; Lee, Y.S.; Chao, A.; Chin, C.Y.; Chao, A.S.; Yang, C.Y. Severe preeclampsia is associated with a higher relative abundance of Prevotella bivia in the vaginal microbiota. *Sci. Rep.* **2020**, *10*, 18249. [CrossRef] [PubMed]
60. Shimaoka, M.; Yo, Y.; Doh, K.; Kotani, Y.; Suzuki, A.; Tsuji, I.; Mandai, M.; Matsumura, N. Association between preterm delivery and bacterial vaginosis with or without treatment. *Sci. Rep.* **2019**, *9*, 509. [CrossRef] [PubMed]
61. Tyas, B.D.; Lestari, P.; Akbar, M.I.A. Maternal perinatal outcomes related to advanced maternal age in preeclampsia pregnant women. *J. Fam. Reprod. Health* **2019**, *13*, 191. [CrossRef]
62. Farzaneh, F.; Tavakolikia, Z.; Soleimanzadeh Mousavi, S.H. Assessment of occurrence of preeclampsia and some clinical and demographic risk factors in Zahedan city in 2017. *Clin. Exp. Hypertens.* **2019**, *41*, 583–588. [CrossRef]
63. Mattsson, K.; Juárez, S.; Malmqvist, E. Influence of socio-economic factors and region of birth on the risk of preeclampsia in Sweden. *Int. J. Environ. Res. Public Health* **2022**, *19*, 4080. [CrossRef]
64. Brosens, I.; Muter, J.; Ewington, L.; Puttemans, P.; Petraglia, F.; Brosens, J.J.; Benagiano, G. Adolescent preeclampsia: Pathological drivers and clinical prevention. *Reprod. Sci.* **2019**, *26*, 159–171. [CrossRef] [PubMed]
65. Paul, T.D.; Hastie, R.; Tong, S.; Keenan, E.; Hiscock, R.; Brownfoot, F.C. Prediction of adverse maternal outcomes in preeclampsia at term. *Pregnancy Hypertens.* **2019**, *18*, 75–81. [CrossRef]
66. Krishnamurti, T.; Davis, A.L.; Simhan, H.N. Worrying yourself sick? Association between pre-eclampsia onset and health-related worry in pregnancy. *Pregnancy Hypertens.* **2019**, *18*, 55–57. [CrossRef]
67. Ruz, G.A.; Henríquez, P.A.; Mascareño, A. Bayesian Constitutionalization: Twitter Sentiment Analysis of the Chilean Constitutional Process through Bayesian Network Classifiers. *Mathematics* **2022**, *10*, 166. [CrossRef]
68. Salman, I. Learning the Structure of the Tree and Tree Augmented Naive Bayesian from Incomplete and Imbalanced Data. In Proceedings of the 2020 21st International Arab Conference on Information Technology (ACIT), Giza, Egypt, 28–30 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–7.
69. Wester, P.; Heiding, F.; Lagerström, R. Anomaly-based intrusion detection using tree augmented naive bayes. In Proceedings of the 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), Gold Coast, Australia, 25–29 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 112–121.