Computers, Materials & Continua
DOI:10.32604/cmc.2022.019882
Article

Tech Science Press

An Ensemble Methods for Medical Insurance Costs Prediction Task

Nataliya Shakhovska[1], Nataliia Melnykova[1,*], Valentyna Chopiyak[2] and Michal Gregus ml[3]

[1]Department of Artificial Intelligence, Lviv Polytechnic National University, Lviv, 79013, Ukraine
[2]Department of Clinical Immunology and Allergology, Danylo Halytsky Lviv National Medical University, Lviv, 79010, Ukraine
[3]Faculty of Management, Comenius University, Bratislava, 814 99, Slovakia
[*]Corresponding Author: Nataliia Melnykova. Email: nataliia.i.melnykova@lpnu.ua
Received: 29 April 2021; Accepted: 15 June 2021

Abstract: The paper reports three new ensembles of supervised learning predictors for managing medical insurance costs. The open dataset is used for data analysis methods development. The usage of artificial intelligence in the management of financial risks will facilitate economic wear time and money and protect patients' health. Machine learning is associated with many expectations, but its quality is determined by choosing a good algorithm and the proper steps to plan, develop, and implement the model. The paper aims to develop three new ensembles for individual insurance costs prediction to provide high prediction accuracy. Pierson coefficient and Boruta algorithm are used for feature selection. The boosting, stacking, and bagging ensembles are built. A comparison with existing machine learning algorithms is given. Boosting modes based on regression tree and stochastic gradient descent is built. Bagged CART and Random Forest algorithms are proposed. The boosting and stacking ensembles shown better accuracy than bagging. The tuning parameters for boosting do not allow to decrease the RMSE too. So, bagging shows its weakness in generalizing the prediction. The stacking is developed using K Nearest Neighbors (KNN), Support Vector Machine (SVM), Regression Tree, Linear Regression, Stochastic Gradient Boosting. The random forest (RF) algorithm is used to combine the predictions. One hundred trees are built for RF. Root Mean Square Error (RMSE) has lifted the to 3173.213 in comparison with other predictors. The quality of the developed ensemble for Root Mean Squared Error metric is 1.47 better than for the best weak predictor (SVR).

Keywords: Healthcare; medical insurance; prediction task; machine learning; ensemble; data analysis

1 Introduction

Digital health is a sector that is growing globally. In the whole world, the number of Digital Health companies has been doubled in the last five years [1]. The governments pledged hundreds of millions of dollars to support the local digital health industry. The prediction of individual

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

health insurance costs is an essential task in the healthcare system, particularly for people with orphan diseases [2–4]: prevention and medical insurance help to decrease the cost of treatment.

Data science technologies are becoming increasingly used in various fields of human activity. Such active implementation is due to the problems that can be solved using existing solutions in this scientific field. Namely, the searching for anomalies and the analysis of atypical behavior of the client, or the search for the causes of the crime; personalized solutions—selection of advertising for customers, e-mails to users, referral solutions; quantitative forecasts—efficiency indicators, qualitative indicators of enterprises, etc. the field of medicine is an excellent platform for implementing innovative approaches to data science. Since digital health is currently a trendy field of medicine, many governments in developed countries invest heavily in its development.

Health insurance in developed countries is experiencing two critical problems, such as the rapid cost of health care and the growing number of people who are not insured. Such influence generates growing political support for broad-based reforms to address these issues.

An analysis of this problem will allow us to assess the risks to human health, namely the projected cost of treatment, the quality of life of people and their level of well-being. This applies to the cost of insurance in the lives of individuals. In addition, the results of the analysis will provide an assessment of the risks of insurance companies regarding payments. Namely, the forecast results become important. They must be accurate enough to measure or quantify the amount covered by a particular policy and the insurance costs to be paid for it. Different variables evaluate these indicators, where each of them is important. Thus, insurance is a process that reduces or eliminates the cost of losses caused by various risks and factors.

If an indicator is omitted when calculating the amounts, the insurance policy changes in general. Therefore, it is critical that these tasks be performed with high accuracy because human mistakes can happen. ML can summarize the effort or method for insurance policymaking. The model trained based on insurance data can be defined as the model's input data, then the model can correctly predict the cost of the insurance policy. This reduces human effort and resources and improves the insurance company's profitability. Thus, the accuracy can be improved with the proposed three different ensemble models, which will optimize the forecasting process.

Medical insurance is an essential part of the medical domain. However, medical costs are difficult to predict since most money comes from rare conditions of the patients. Different machine learning algorithms and deep learning technics are used for data prediction. The two parameters training time and accuracy, are analyzed. The training time of the biggest part of machine learning algorithms is not too huge. However, the accuracy of the prediction results for these methods is not so high. Deep learning models allow to find hidden patterns too, but training time does not allow to use of these models in the real time [5].

That is why the paper aims to develop new ensembles for individual insurance costs prediction to provide high prediction accuracy. The novelty of the paper is new schema of stacking ensemble base on weal predictors selection and hyperparameters choosing.

Contributions: This research possesses various contributions in the domain of cost prediction.

(1) First, two feature selection technics for the comparison of the prediction accuracy of the different machine learning algorithms were applied. The weak components for the design an ensemble models were found;

(2) Second, three different ensemble models based on boosting, bagging, and stacking approaches for solving medical insurance costs prediction task were designed;
(3) Lastly, it is experimentally established that the new stacking model based on machine learning algorithms that use Random Forest as a meta-algorithm provides higher prediction accuracy for solving the stated task.

The paper is organized as following. The literature review and methods and models for medical insurance cost prediction are given in Section 2. The Section 3 represents the dataset description and Exploratory data analysis. Weak predictors were selected. Next, the novel approach based on three ensembles of the weak predictors are developed. Section 4 represents the results of the developed ensembles and comparison with other predictors. The conclusion (Section 5) underlines the novelty of the proposed approach and prospects for further research.

2 Literature Review

Methods and systems for medical data analysis are given in [5–9]. The usage of artificial intelligence in the management of financial risks will enable economically wear time and money and save the health of patients. Machine learning is associated with many expectations, but its quality is determined by choosing a good algorithm and the proper steps to plan, develop, and implement the model. The main drawback of the RBF networks for solving this task is that they provide only a local approximation of the nonlinear response surface.

The unique features of data mining with medical data are described [10]. Artificial intelligence works effectively in the initial stages of risk assessment, starting from collecting and analyzing information and ending the development of control algorithms.

Medical data has a multilevel structure with hidden dependencies [11]. There is very important to find patterns and use various methods of analysis together. That is why different ensembles of machine learning (ML) models are used for medical data analysis. The model for nested data is developed in [11]. However, this is a limitation for non-nested dataset.

In paper [12], the ensemble of random forests (RF) and support vector machine (SVM) is used to predict the modulus of elasticity of recycled aggregate concrete. This classical ensemble allows to increase the accuracy, that is why it can be taken into account.

The ML models and their performance for different application domains are analyzed in [13]. The comparison show the different quality of the proposed algorithms.

Paper [14] is focused on SMEs' credit risk problem by forecasting with the help of random subspace and MultiBoosting ensemble. This approach combines more than one type of ensembles.

Paper [15] presents a new framework incorporating 7 supervised ML algorithms to exploit multiple variant callers' strengths, using a non-redundant set of biological and sequence features.

An ensemble of K-Nearest Neighbour (KNN) classifiers for recommendation to leverage the heterogeneity of different groups of meta-features is analyzed in [16].

In work [17], an ensemble-based machine learning model comprises RF, ID3, Adaboost, KNN, Logistic Regression has experimented on diabetic retinopathy dataset. This approach can be used only for classification task.

Paper [18] represents the solution of the forecasting problem of the direction of stock price movement. The tree-based ensemble consists of Random Forest, XGBoost, Bagging Classifier,

AdaBoost, Extra Trees combined with Voting Classifier is developed. Bagging Ensemble Classifier is used for Diabetic retinopathy in [19].

The analysis of mentioned papers shown the effectiveness of ensembles in comparison with single ML-based methods. Besides, specific models can be used for regression tasks too. For example, paper [20] presents a non-iterative model using Wiener polynomial and linear SGTM neural-like structure. Wiener polynomial provides a nonlinear input extension. The approximation properties of this polynomial give highly accurate results. Polynomial coefficients are sought using SGTM ANN, which offers high speed. In general, this method shows a significant increase in solving the medical costs prediction task.

However, large degrees of the polynomial significantly increase the learning time of this model. In addition, the method's accuracy is not satisfactory for its practical implementation in insurance companies [21].

The ensembles for text analysis are developed in [22–24]. Clustering and randomized search are combined into ensemble for text sentiment classification. Experimental analysis of classification tasks includes also software defect prediction, credit risk modeling, spam filtering, and semantic mapping. However, mentioned methods are used for classification task solving.

Papers [25–28] used different machine learning technics for medical data analysis and theirs combination for accuracy increasing.

That is why it is necessary to develop new or improve existing individual insurance costs prediction methods and tools that would provide high prediction accuracy with sufficient training speed. Authors propose to develop three different ensemble models based on boosting, bagging, and stacking and compare the prediction accuracy with well-known machine learning algorithms.

3 The Materials and Methods

3.1 The Experimental Setup

The experimental setup is organized as following:

- Exploratory data analysis (missing data imputation and feature selection);
- Weak predictors selection;
- Hyperparameters choosing based on Grid search;
- The ensemble development.

3.2 Dataset Description and Exploratory Data Analysis

The medical insurance payments dataset [29] was selected. It consists of 7 attributes and 1338 vectors. The task is to predict individual payments for health insurance.

Data preprocessing stage is described in [20]. The preprocessing for mentioned dataset consists of the following stages:

- Missing data imputation,
- Data transformation.

In the missing data imputation stage MICE algorithm [30] is used. Totally 13 instances have had missing data. For data transformation stage one-hot encoding is used for binary (sex, smoker) and categorical (region) variables.

Finally, the dataset consists of 11 features, namely:

- X1 – age,

- X2, X3 – sex (female, male as a result of one-hot encoding),
- X4 – bmi,
- X5 – children,
- X6, X7 – smoker (smoker, non-smoker as result of one-hot encoding),
- X8, X9, X10, X11 – region (as a result of one-hot encoding).

Attribute Y is target variable. Dataset consists of variable charges. Statistics is represented below (Tab. 1).

Table 1: Dataset statistic indicators

Title 2	Min	1st Qu	Median	Mean	3rd Qu	Max
X1	18	27	40	39.55	51.75	64
X2	0	0	0	0.4897	1	1
X3	0	0	1	0.5103	1	1
X4	15.96	26.41	30.50	30.78	34.94	52.58
X5	0	0	1	1.076	2	5
X6	0	0	0	0.2009	0	1
X7	0	1	1	0.7991	1	1
X8	0	0	0	0.243	0	1
X9	0	0	0	0.2748	1	1
X10	0	0	0	0.2374	0	1
X11	0	0	0	0.2449	0	1
y	1122	4740	9333	13214	16547	63770

Note: The statistic of the dataset shown that distribution is unbalanced.

The next step is feature selection. To do this, Pierson coefficient is used (Fig. 1). A significant correlation between features is absent. However, smokers (x6 and x7) correlated with the target variable y. For non-smoker patients (X7), the correlation between bmi (X4) and charges (Y) is not clear.

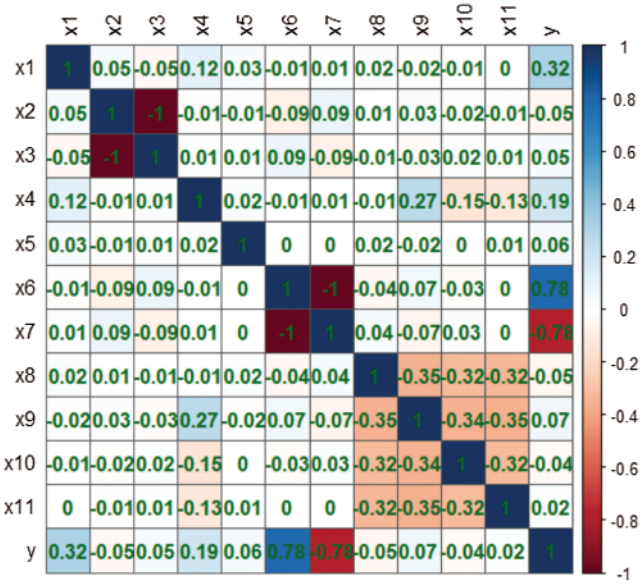

Figure 1: Correlation matrix

Next, Boruta algorithm was used for significant variables selection. Boruta is the heuristic algorithm for selecting substantial features based on the use of Random Forest. The algorithm's essence is that at each iteration, features are removed whose Z-measure is less than the maximum Z-measure among the added features. To get the Z-measure of a feature [31], it is necessary to calculate the importance of the feature, obtained using the built-in algorithm in Random Forest, and divide it by the standard deviation of the feature importance. The result of the selection is given in Tab. 2. So, age (X1), bmi (X4), smoker (X6, X7), children (X5) and region Northeast (X9) are the most important features.

X6 and X7 are chosen by two methods.

Table 2: Features selection based on the results of the Boruta algorithm

Variable	MeanIpm	Decision
X1	71.122273	confirmed
X4	60.175260	confirmed
X6	26.665077	confirmed
X7	25.300914	confirmed
X5	8.217143	confirmed
X9	1.794187	confirmed

3.3 Weak Predictors Selection

For model development, splitting the dataset into the training dataset and testing dataset is built. The general rule of thumb is 75% for split ratio, 75% train, 25% test.

The two prediction models will be built for the whole dataset and selected features, respectively. To create the ensemble, the weak predictors must be selected.

First, linear regression is built for the whole dataset. The regression coefficients and model parameters are given in Tab. 3.

Table 3: Estimate std. error t value Pr(>|t|) for linear regression

	Min	1Q Median	3Q	Max
(Intercept)	−12117.55	1116.01	−10.858	<2e-16***
X1	259.49	13.23	19.616	<2e−16***
X2	253.80	372.52	0.681	0.49583
X4	339.19	32.05	10.584	<2e−16***
X5	439.17	155.59	2.823	0.00485
X6	23653.43	465.26	50.840	<2e−16***
X8	−1245.93	531.88	−2.342	0.01934*
X9	−1119.29	531.48	−2.106	0.03544*
X10	−471.67	533.49	−0.884	0.37683

The values of residuals and statistics are the following:

- Min residuals: −10915,
- 1Q residuals: −2880,

- Median residuals: −999,
- 3Q residuals: 1305,
- Max residuals: 25377,
- Multiple R-squared: 0.7486, Adjusted R-squared: 0.7467
- F-statistic: 394.9 on 8 and 1061 DF, p-value: <2.2e−16

Second, linear regression is built for selected variables x1, x4, x5, x6, x7, x9 (Tab. 4). The median of residuals is lower than for linear regression built on whole dataset, however R-squared error is not significantly lower:

- Min residuals: −11745.4,
- 1Q residuals: −3005.1,
- Median residuals: −963.4,
- 3Q residuals: 1304.9,
- Max residuals: 26137.6,
- Multiple R-squared: 0.7472, Adjusted R-squared: 0.746
- F-statistic: 628.9 on 5 and 1064 DF, p-value: <2.2e−16

Table 4: Results of the linear regression based on the selected parameters

	Min	1Q Median	3Q	Max
(intercept)	−12382.51	1116.01	−10.858	<2e−16***
X1	259.89	13.23	19.642	<2e−16***
X4	332.57	31.94	10.413	<2e−16***
X5	435.87	155.78	2.798	0.00524**
X6	23654.16	463.96	50.983	<2e−16***
X9	−515.96	433.19	−1.191	0.23389

To sum up, there is no significant difference in R-squared error values for the whole dataset and selected features. That is why the whole dataset will be used for other predictors' development.

In the next step, a regression tree is built. 10-fold cross-validation repeated 3 times is performed. The important attributes are X6, X1, X4.

Regression tree

1) root 1070 154804600000 13214.130
2) x6 < 0.5 855 31881250000 8496.645
4) x1 < 42.5 466 10297680000 5243.559*
5) x1 >= 42.5 389 10744460000 12393.660*
3) x6 >= 0.5 215 28227340000 31974.340
6) x4 < 30.1 103 2622399000 21333.910*
7) x4 >= 30.1 112 3218960000 41759.740*

The optimal subtree is built with 3 splits, 4 terminal nodes, and a cross-validated error of 0.18 (Tab. 5).

The cross-validation error in this table represents x-error. As factors for tree pruning were used xstd, rel-error and x-error. For a description of the tree's height row was used. As a sign of

a better model's accuracy, a high number of levels in the tree could be used. Xstd is the bias of x-error. The complexity parameter (CP) controls the size of the regression tree. In addition, the selection of optimal tree size could be done with the help of CP. The stopping criteria of tree building are comparing the cost of adding another variable to the regression tree from the current node and the value of cp. If the first is higher than the second, then the building is stopped. So, CP is penalty results in a fully grown tree. Nsplit represents the number of splitting in single tree.

Table 5: The complexity table of regression tree

CP	nslit	rel-error	X-error	xstd
0.611713	0	1	1.0015	0.057191
0.144608	1	0.38829	0.3895	0.020839
0.070018	2	0.24368	0.24491	0.016339
0.01000	3	0.17366	0.17978	0.015314

In the next step, the well-known ML algorithms are analyzed for "weak" predictors choosing.

KNN, Support Vector Regression (SVR) with Radial Basis Function and perceptron with 10 neurons in the hidden layer and tangent hyperbolics (tanh) activation function, Stochastic gradient descent are used for proposed dataset analysis. The kernel trick enables the SVR to obtain a fit, and then data is charted to the initial space. The hyperpaparemeters are chosen based on Grid Search. The Cost complexity criterion is used for optimization. The hyperparameters combination was presented in grid form. In the next stage, the optimal parameters for each repressor were chosen. The comparison of weak predictors is given in Tab. 6.

Table 6: The comparison of weak predictors

Predictor	MAPE	RMSE	MAE	MSE
Linear regression	0.4169418	6030.957	4151.564	36372447
Regression tree	0.3250840	4673.567	2723.527	21215612
KNN, 3 neighbors	0.2345744	4680.489	2717.216	21164495
SVR with rbf	0.2348014	4665.074	2721.627	21206705
Perceptron with 1 hidden layer and tanh activation function	0.7998291	7867.960	8213.130	89263891
Stochastic gradient descent	0.4356261	5890.74	4115.136	34700814

So, regression tree, linear regression, KNN, SVM, and Stochastic gradient descent are selected as weak predictors.

3.4 Proposed Ensemble Development

There are three time-tested ways to make ensembles: stacking, bagging, and boosting.

- In short, the peculiarity of stacking is that we teach several different algorithms and pass their results to the input of the last, who makes the final decision. The critical difference is different algorithms because if we teach the same algorithm on the same data, it will not matter. Regression is usually used as the final algorithm.

- For bugging, we train one algorithm many times on random samples from the source data. In the end, the results are average. The most famous example of bugging is the Random Forest algorithm. It is the possibility of paralleling that gives bugging an advantage over other ensembles.
- A distinctive feature of the boosting ensemble is that we train our algorithms consistently, even though each subsequent one pays special attention to the cases in which the previous algorithm failed. We take samples from the source data in the running, but now it's not entirely random. In each new selection, we take part of the data on which the previous algorithm worked incorrectly. In fact, we are learning a new algorithm from the mistakes of the previous one. This ensemble has a very high accuracy, which is an advantage over all other ensembles. However, there is also a downside - it is difficult to parallelize. It still works faster than neural networks, but slower than bugging.

All possible ensembles are developed in the paper.

At the first stage, Boosting modes based on regression tree and stochastic gradient descent is built. Boosting is a compositional machine learning meta-algorithm, which is mainly used to reduce bias (estimation error) and variance in supervised learning also defined as a family of machine learning algorithms that transform weak learning algorithms into strong ones.

The number of folds or number of resampling iterations is equal to 10. The number of complete sets of folds to compute is equal to 3. Automatic tuning of parameters is used too. Mean absolute error (MAE), Root mean squared error (RMSE) and Rsquared error are used for model evaluation.

The results are given in Fig. 2. We can see that the Boosted Stochastic gradient descent produces a more precise model with RMSE equal to 44487.912.

```
summary.resamples(object = boosting_results)

Models: rf, gbm
Number of resamples: 30

MAE
        Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
rf  1897.465 2535.698 2637.849 2647.391 2803.202 3409.446    0
gbm 1862.611 2348.779 2483.238 2479.590 2614.298 3118.582    0

RMSE
        Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
rf  3140.132 4331.684 4638.013 4645.586 4936.347 5940.452    0
gbm 2845.302 4157.624 4468.913 4487.912 4858.473 5795.879    0

Rsquared
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
rf  0.6995033 0.8163128 0.8478365 0.8463090 0.8791907 0.9522437    0
gbm 0.7174028 0.8300307 0.8584581 0.8558444 0.8911981 0.9626013    0
```

Figure 2: Boosting results

In the next step, a new bagging machine learning algorithm is developed. Bagging includes training the same algorithm many times by applying different subsets sampled from the training dataset. The final output forecast is then averaged across the estimates of all the sub-models.

Bagged CART and Random Forest algorithms are proposed. Both algorithms include parameters that are not tuned (Fig. 3).

```
summary.resamples(object = bagging_results)

Models: treebag, rf
Number of resamples: 30

MAE
            Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
treebag 2757.035 3001.373 3120.609 3134.382 3252.172 3715.016    0
rf      1799.630 2447.891 2584.798 2583.788 2774.163 3392.942    0

RMSE
            Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
treebag 3741.864 4548.693 4893.969 4907.328 5220.188 5942.013    0
rf      3137.939 4335.779 4631.444 4651.663 4973.388 5996.950    0

Rsquared
             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
treebag 0.7058063 0.7959589 0.8362564 0.8292028 0.8606078 0.9273849    0
rf      0.6952643 0.8157238 0.8452514 0.8454689 0.8739311 0.9515307    0
```

Figure 3: Bagging results

We can see that results are worse than for Boosted Stochastic gradient descent.

The next step is the combination of multiple predictors using stacking. KNN, SVM, rtree, linear regression (lm), GBM are used for ensemble development.

Models' correlation is given in Tab. 7.

Table 7: Pairwise correlations for ensemble's 5 sub-models

Model	rtree	gbm	lm	knn	svmRadial
rtree	1.00000000	0.8500638	0.7016961	0.06054963	0.82971722
Gbm	0.85006384	1.0000000	0.7767240	0.11740359	0.84256197
Lm	0.70169614	0.7767240	1.0000000	0.22494158	0.82084279
Knn	0.06054963	0.1174036	0.2249416	1.00000000	0.09239563
svmRadial	0.82971722	0.8425620	0.8208428	0.09239563	1.00000000

Appropriative correlation (more than 0.8) is between svmRadial, rtree, Gbm and lm.

The final stacking schema is given in Fig. 4. The random forest (RF) algorithm is used to combine the predictions. 100 trees are built for RF.

We combine the predictions of the predictors using random forest. We can see that stacking model has lifted the RMSE to 3173.213 (Tab. 8).

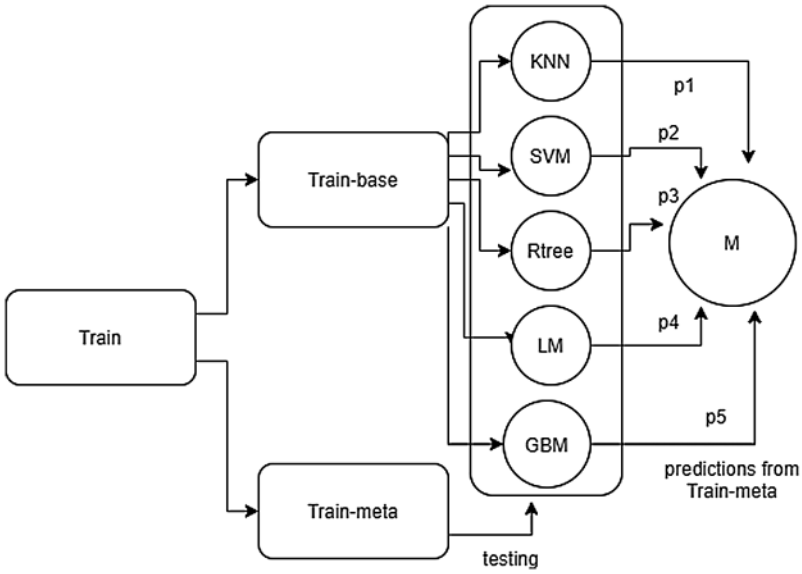

Figure 4: The used predictors stacking schema

Table 8: Stacking model results

mtry	RMSE	Rsquared	MAE
2	3733.777	0.5078975	2537.094
4	3173.213	0.5006131	2447.961

4 Results

The simulation of the proposed method was carried out using the author's software (console application). The proposed and existing methods are tested on the same hardware: Intel Core 5 Quad E6600 2.4 GHz, 16 GB RAM, HDD WD 2 TB 7200 RPM.

The comparison of ML models and proposed ensembles is shown in Fig. 5. The most significant errors in solving the stated task were obtained using classical single models (NN, Linear regression, SGD). The knn, rtree and SVR methods show slightly better results in terms of RMSE-based accuracy. However, the highest model accuracy is for stacking developed as a combination of weak predictors.

The difference between the rest two ensembles and weak predictors SVR and KNN are not significant. The tuning parameters for boosting do not allow to decrease the RMSE too. So, bagging shows its weakness in generalizing the prediction.

The quantitative indicators for evaluating the developed stacking ensemble in terms of both training and testing modes are given in Tab. 9 according to the following indicators: Mean absolute percentage error (MAPE), Sum square error (SSE), Symmetric mean absolute percentage error (SMAPE), Root mean square error (RMSE), Mean absolute error (MAE).

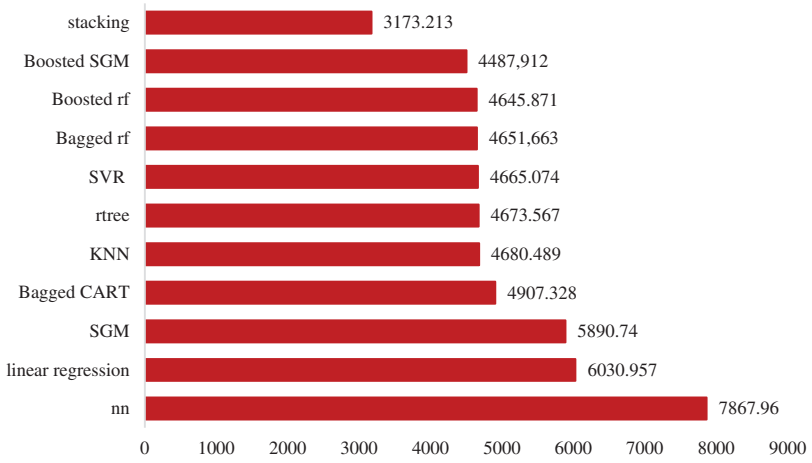

Figure 5: Comparison with other predictors based on the RMSE

The time complexity of the proposed stacking ensemble can be evaluated for distributed mode. The weak repressors should be executed in parallel and RF should summarize the results. The worth time complexity for chosen repressors is for SVR and it is equal to $O(N^3)$. The best time complexity is for KNN and it is equal to O(n). The Random forest time complexity depends on size T and maximum depth D (excluding the root), or $O(T \cdot D)$. In our case the overall time complexity is equal to $O(N^3) + O(T \cdot D)$, where $N = 1338$, $T = 2$, $D = 9$.

Table 9: Quantitative indicators for evaluating of the developed ensemble

Ensemble	MAPE	SSE	SMAPE	RMSE	MAE
	Training dataset				
Stacking	27.371	91.623	0.116	3173.213	2447.961
Boosted SGM	29.034	96.345	0.210	4487.912	34061.460
Boosted RF	29.568	98.532	0.321	4665.074	34072.569
Bagged RF	29.346	98.517	0.320	4651.663	34061.003
	Testing dataset				
Stacking	28.564	96.653	0.131	3185.423	25172.561
Boosted SGM	30.056	99.456	0.256	4527.967	34734.662
Boosted RF	30.254	99.945	0.378	4685.756	34863.614
Bagged RF	30.123	99.403	0.368	4681.322	34782.216

5 Discussion

The stacking gives the best results and it is built on chosen weak predictors. The developed method increased generalization properties.

A model averaging ensemble combines the predictions from multiple trained models. A limitation of this approach is that individual model contributes the same amount to the ensemble prediction, regardless of how well the model performed. A modification of this approach called

a weighted average ensemble weighs the contribution of each ensemble member by the trustor expected performance of the model on a holdout dataset. This allows well-performing models to contribute more and less-well-performing models to contribute less. The weighted average ensemble provides an improvement over the average model ensemble.

A further generalization of this approach is replacing the linear weighted sum (e.g., linear regression) model used to combine the predictions of the sub-models with any learning algorithm (Random Forest). In proposed stacking, an algorithm takes the outputs of sub-models as input and attempts to best combine the input predictions to better output prediction.

The simulation of the developed method for solving the medical insurance costs prediction task showed a significant increase in accuracy compared with existing approaches (regression tree, multilayer perceptron, K Nearest Neighbor, Support Vector Machine, Stochastic Gradient Descent, linear regression, etc.). The quality of developed ensemble for RMSE is 1.47 better than for the best weak predictor (SVR).

The results are presented for the whole dataset. The usage of the well-known ML methods and proposed ensembles are not significantly different.

An essential role in implementing the computational intelligence methods for solving the practical tasks of processing large data arrays is important for the duration of the training procedure. That is why the comparison of the training procedure duration for all considered methods is given too.

6 Conclusion

The paper describes three new ensembles of supervised learning predictors for managing medical insurance costs. Open dataset is used for data analysis methods development. Several weak predictors are implemented on this dataset.

As it shown, the adding new predictor can improve the predictive accuracy, because the base predictors' outputs are features for the final predictor. In this case, these 'second level' features are likely correlated because all base predictors are all trying to predict the same thing. But, they do it suboptimally. The hope is that they behave in different ways, so that the final predictor can combine the noisy predictions into a better final prediction. Loosely, then, adding new base predictors has the best chance of helping when they do a good job and behave differently than existing base classifiers, but this isn't guaranteed. If the new predictors perform at chance they can't help, and will probably hurt. The final predictor can overfit, and providing it with more base classifiers may increase its ability to do so.

Seven weal predictors were analyzed with tuned hyperparameters. The best weak predictor is SVR with RMSE equal to 4665, 074.

Four ensembles were developed in the paper, two of them are boosted ensembles. The boosting and stacking ensembles shown better accuracy than bagging. The worth accuracy is shown the bagged Random forest equal to 4651, 663. The stacking is developed using K Nearest Neighbors (KNN), Support Vector Machine (SVM), Regression Tree, Linear Regression, Stochastic Gradient Boosting. The random forest (RF) algorithm is used to combine the predictions. One hundred trees are built for RF. Root Mean Square Error (RMSE) has lifted the to 3173.213 for training dataset and to 3185.423 for testing dataset. A comparison with existing machine learning algorithms is given. The highest model accuracy is for stacking developed as a combination of weak

predictors. The quality of developed ensemble for RMSE is 1.47 better than for the best weak predictor (SVR).

The limitations of the study are the following:

- The time complexity allows to use the proposed ensemble in the real time in distributed mode only.
- The quality of the ensemble depends on the dataset. For an imbalanced dataset, the prediction accuracy will be lower;
- The modeling of charged cases should be provided together with clustering analysis. The authors plan to model each separated cluster and compare the predicted accuracy.

We also will conduct future research in designing cascades based on existing machine learning algorithms or ANN. This approach will provide the possibility of linearization of the response surface, which will significantly affect the overall accuracy of the regressor.

Author Contributions: Conceptualization, N.M. and N.S.; methodology, N.M.; software, N.S.; validation, I.I., N.S., V.C and N.M.; formal analysis, N.S.; investigation, N.M.; resources, N.S.; writing—original draft preparation, N.S. and I.I; writing—review and editing, N.M.; visualization, N.M.; supervision, N.S.. All authors have read and agreed to the published version of the manuscript.

Funding Statement: This research was funded by National Research Foundation of Ukraine.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Digital Health 150: The digital health startups transforming. The future of healthcare, [Online]. Available: https://www.cbinsights.com/research/report/digital-health-startups-redefining-healthcare/ (accessed on 9 April 2021).

[2] J. H. Lee, "Pricing and reimbursement pathways of new ophan drugs in South Korea: A longitudinal comparison. in healthcare," *Multidisciplinary Digital Publishing Institute*, vol. 9, no. 3, pp. 296, 2021.

[3] M. Czech, A. Baran-Kooiker, K. Atikeler, M. Demirtshyan, K. Gaitova *et al.*, "A review of rare disease policies and orphan drug reimbursement systems in 12 eurasian countries," *Frontiers in Public Health*, vol. 7, pp. 416, 2020.

[4] R. Spencer, C. Rossi, M. Lees, D. Peebles, P. Brocklehurst *et al.*, "Achieving orphan designation for placental insufficiency: Annual incidence estimations in Europe" *BJOG: An international journal of obstetrics & gynaecology*," *EVERREST Consortium*, vol. 126, no. 9, pp. 1157–1167, 2019.

[5] N. Shakhovska, S. Fedushko, I. Shvorob and Y. Syerov, "Development of mobile system for medical recommendations," *Procedia Computer Science*, vol. 155, pp. 43–50, 2019.

[6] S. Fedushko and T. Ustyianovych, "Medical card data imputation and patient psychological and behavioral profile construction," *Procedia Computer Science*, vol. 160, pp. 354–361, 2019.

[7] V. Yakovyna, V. Khavalko, V. Sherega, A. Boichuk and A. Barna, "Biosignal and image processing system for emotion recognition applications," *CEUR Proceedings*, pp. 181–191, 2019.

[8] M. Z. Latif, K. Shaukat, S. Luo, I. A. Hameed, F. Iqbal *et al.*, "Risk factors identification of malignant mesothelioma: A data mining based approach", in *2020 Int. Conf. on Electrical, Communication, and Computer Engineering (ICECCE)*, Kuala Lumpur, Malaysia, pp. 1–6, 2020.

[9] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, *et al.*, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, pp. 100204, 2019.

[10] K. J. Cios and G. W. Moore, "Uniqueness of medical data mining," *Artificial Intelligence in Medicine*, vol. 26, no. 1–2, pp. 1–24, 2002.
[11] H. Goldstein, W. Browne and J. Rasbash, "Multilevel modelling of medical data," *Statistics in Medicine*, vol. 21, no. 21, pp. 3291–3315, 2002.
[12] T. Han, A. Siddique, K. Khayat, J. Huang and A. Kumar, "An ensemble machine learning approach for prediction and optimization of modulus of elasticity of recycled aggregate concrete," *Construction and Building Materials*, vol. 244, pp. 118–271, 2020.
[13] S. Ardabili, A. Mosavi and A. Várkonyi-Kóczy, "Advances in Machine Learning Modeling Reviewing Hybrid and Ensemble Methods," in *Int. Conf. on Global Research and Education,* Springer, Cham, pp. 215–227, 2019.
[14] Y. Zhu, L. Zhou, C. Xie, G. Wang and T. Nguyen, "Forecasting SMEs' credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach," *International Journal of Production Economics*, vol. 211, pp. 22–33, 2019.
[15] I. Anzaru, A. Sverchkova, R. Stratford and T. Clancy, "Neomutate: An ensemble machine learning framework for the prediction of somatic mutations in cancer," *BMC Medical Genomics*, vol. 12, no. 1, pp. 1–14, 2019.
[16] X. Zhu, C. Ying, J. Wang, J. Li, X. Lai *et al.*, "Ensemble of ML-kNN for classification algorithm recommendation," *Knowledge-Based Systems*, vol. 106, pp. 933, 2021.
[17] G. Reddy, S. Bhattacharya, S. Ramakrishnan, C. L. Chowdhary, S. Hakak *et al.*, "An ensemble based machine learning model for diabetic retinopathy classification," in *2020 Int. Conf. on Emerging Trends in Information Technology and Engineering, IC-ETITE*, VIT Vellore, IEEE, pp. 1–6, 2020.
[18] E. K. Ampomah, Z. Qin and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement," *Information*, vol. 11, no. 6, pp. 332, 2020.
[19] S. K. Somasundaram and P. Alli, "A machine learning ensemble classifier for early prediction of diabetic retinopathy," *Journal of Medical Systems*, vol. 41, no. 12, pp. 1–12, 2017.
[20] R. Tkachenko, I. Izonin, P. Vitynskyi, N. Lotoshynska and O. Pavlyuk, "Development of the non-iterative supervised learning predictor based on the ito decomposition and SGTM neural-like structure for managing medical insurance costs," *Data*, vol. 3, pp. 46, 2018.
[21] R. Caruana, "An empirical comparison of supervised learning algorithms," in *Proc. of the 23rd Int. Conf. on Machine Learning*, Pittsburgh, Pennsylvania, pp. 161–168, 2006.
[22] A. Onan, M. A. Toçoğlu, "A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification," *IEEE Access*, vol. 9, no. 7, pp. 701–7722, 2021.
[23] A. Onan, S. Korukoğlu, H. Bulut, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," *Expert Systems with Applications*, vol. 62, pp. 1–16, 2016.
[24] A. Onan, S. Korukoğlu, H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Information Processing & Management*, vol. 53, no. 4, pp. 814–833, 2017.
[25] T. M. Alam, M. M. A. Khan, M. A. Iqbal, W. Abdul and M. Mushtaq, "Cervical cancer prediction through different screening methods using data mining," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, pp. 9, 2019.
[26] M. U. Ghani, T. M. Alam, and F. H., Jaskani, "Comparison of classification models for early prediction of breast cancer," in *2019 Int. Conf. on Innovative Computing (ICIC)*, Ostrava, Czech Republic, pp. 1–6, 2019.
[27] K. Shaukat, F. Iqbal, T. M. Alam, G. K. Aujla, I. Devnath *et al.*, "The impact of artificial intelligence and robotics on the future employment opportunities," *Trends in Computer Science and Information Technology*, vol. 5, no. 1, pp. 50–54, 2020.
[28] X. Yang, M. Khushi, and K. Shaukat, "Biomarker CA125 feature engineering and class imbalance learning improves ovarian cancer prediction," in *2020 IEEE Asia-Pacific Conf. on Computer Science and Data Engineering (CSDE)*, Gold Coast, Australia, pp. 1–6, 2020.

[29] Dataset license: Open database, Dataset: https://www.kaggle.com/mirichoi0218/insurance.
[30] S. V. Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, vol. 45, no. 1, pp. 1–68, 2010.
[31] A. Botchkarev, "Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology," arXiv preprint arXiv:1809.03006. 2018. [Online]. Available: https://arxiv.org/abs/1809.03006 (Accessed on 9 September 2018).