

Academia de Studii Economice din București

Facultatea de Cibernetică, Statistică și Informatică Economică

Specializarea: Analiza afacerilor și controlul performanței întreprinderii



## **PROIECT DE CERCETARE LA ȘTIINȚA DATELOR ÎN AFACERI**

### **Frauda în domeniul asigurărilor de autovehicule**

**Studenți:**

Almăjanu Jeni-Maria, 1116

Barbu Grațiela-Elena, 1116

Bulei Carmen-Ștefania, 1116

Bratu Alexandru-Florin, 1116

**Profesor coordonator:**

Bizovi Mihai

**București 2023**

## Cuprins

<b>Introducere .....</b>	<b>2</b>
<b>A. Problema identificată.....</b>	<b>2</b>
<b>B. Importanța problemei .....</b>	<b>2</b>
<b>1. Studiile efectuate anterior în domeniul asigurărilor autovehiculelor .....</b>	<b>3</b>
<b>2. Metodologie.....</b>	<b>3</b>
<b>3. Aplicații practice. Soluții implementate.....</b>	<b>4</b>
<b>3.1. Preprocesarea datelor .....</b>	<b>4</b>
<b>3.2. Arbori de decizie, Random Forest, AdaBoost și XGBoost .....</b>	<b>7</b>
<b>3.3. Analiza componentelor principale.....</b>	<b>8</b>
<b>3.3.1. Descrierea datelor .....</b>	<b>8</b>
<b>3.3.2. Extragerea componentelor principale.....</b>	<b>9</b>
<b>4. Concluzii .....</b>	<b>10</b>

## **Introducere**

### **A. Problema identificată**

Proiectul nostru, intitulat „Frauda în domeniul asigurărilor de autovehicule”, are drept scop identificarea unui model optim care să fie capabil să detecteze tentativele de fraudă din domeniul asigurărilor autovehiculelor.

Industria asigurărilor este unul dintre cele mai expuse domenii la tentativele de fraudă, înscenările fiind printre cele mai cunoscute cazuri. Fraudele sunt lipsite de etică și reprezintă pierderi pentru companiile de asigurări, segmentul auto părând să fie cel mai vulnerabil, iar schemele de fraudare sunt din ce în ce mai complexe și derulate de indivizi tot mai bine organizați. În domeniul asigurărilor de autovehicule, fraudele sunt tot mai des întâlnite și implică conspirații ce constau în pretenții false sau exagerate, în cazul daunelor morale sau a vătămări de persoane rezultate în urma întâmplării unui accident.

Frauda din asigurări este o înșelăciune deliberată comisă împotriva sau de către o companie sau un agent de asigurări în scopul obținerii de câștiguri financiare. Frauda poate fi comisă în diferite momente ale tranzacției de către solicitanți, deținători de poliță, reclamanți terți sau profesioniști care furnizează servicii solicitanților. Agenții de asigurări și angajații companiei pot, de asemenea, să comită fraude de asigurare.

### **B. Importanța problemei**

Încă din anii 1990 au fost efectuate diferite studii cu scopul de a încerca să se înțeleagă și definească fraudă în domeniul asigurărilor auto, să se identifice principalii factori și cauzele primordiale ale constituirii cazurilor frauduloase și să se identifice soluții pentru detectarea acestor tipuri de ilegalități. În ciuda acestor eforturi, nu s-a reușit, nici până în zilele noastre, identificarea unui plan general valid referitor la cauzele și factorii care conduc la apariția fraudelor.

Relevanța și importanța subiectului este evidențiată de multiplele rapoarte publicate pe subiect, care afirmă, că fraudă în asigurări afectează 10-20% din totalul contractelor încheiate chiar și în cele mai dezvoltate țări. Așadar, fraudă în asigurări este a doua cea mai răspândită fraudă financiară, după fraudă fiscală.

Prin construirea unui model care poate clasifica fraudă, se pot reduce pierderile pentru compania de asigurări. Provocarea din spatele detectării fraudelor în învățarea automată este că fraudele sunt mult mai puțin frecvente în comparație cu daunele de asigurare legitime. Acest tip de problemă este cunoscută sub numele de clasificare dezechilibrată a clasei.

## 1. Studiile efectuate anterior în domeniul asigurărilor autovehiculelor

În acest punct al proiectului vom încerca să analizăm fraudă din domeniul asigurărilor auto și metodele deja existente de detectare a acesteia. În urma unor căutări amănunțite referitoare la tema abordată am identificat numeroase articole științifice și, chiar, publicații din volumele de conferință, cele mai multe dintre acestea regăsindu-se pe site-ul „Web of Science”. O parte din aceste studii au abordat domeniul fraudei de asigurări auto teoretic ( spre exemplu: Derrig -2002; Warren & Schweitzer -2018), iar celelalte oferă un model sau o abordare specifică de identificare a fraudei ( spre exemplu: Wilson -2009; Hassan & Abraham -2016 ;Zelenkov -2019).

De asemenea, în urma unui raport publicat de FBI au fost obținute următoare informații: o familie obișnuită în SUA are \$400-\$700 cheltuieli în plus în fiecare an datorită fraudei de asigurări; de altfel în SUA și Europa de Vest, 7-10% din polițele sunt afectate, iar în Europa Centrală și de Est, această cifră ajunge la 10-20%. Deși în România nu există statistici oficiale privind fraudă în asigurarea auto, la Forumul Internațional Asigurări-Reasigurări din anul 2017 a fost susținută ideea că cel puțin 15% din asigurările auto sunt afectate de fraudă, însumând valoarea de 400 de milioane de euro anual.

## 2. Metodologie

Metodologia folosită în cadrul proiectului este constituită din următoarele tehnici:

- O primă tehnică de clasificare foarte des întâlnită este clasificare prin intermediul **arborilor decizionali**. Aceștia sunt utilizați pentru a predicționa apartenența unor indivizi/obiecte la anumite categorii/clase. Arbore de decizie reprezintă un arbore în care nodurile interne conțin condiții referitoare la variabilele predictorii iar cele frunză/terminale sunt informații privind variabila predictor.
- **Random Forest** este, de asemenea, un clasificator care conține un număr de arbori de decizie pe diferite subseturi ale setului de date dat și utilizează media pentru a îmbunătăți acuratețea predictivă a acelui set de date.
- **AdaBoost** este un meta-estimator care începe prin ajustarea unui clasificator pe setul de date original și apoi încadrează copii suplimentare ale clasificatorului pe același set de date, dar în care ponderile instanțelor clasificate incorect sunt ajustate astfel încât clasificatorii ulterioare să se concentreze mai mult pe cazuri dificile.
- **XGBoost** este un algoritm de învățare automată, bazat pe arborele de decizie, care utilizează un cadru de creștere a gradientului pentru a prezice o variabilă țintă.
- Metoda **Analizei Componentelor Principale** presupune reducerea dimensiunii setului de date prin explicarea variabilității datelor cu un număr cât mai mic de variabile, comparativ cu setul inițial de date. Pentru a cuantifica informațiile în mai puține variabile, în vederea reducerii dimensionalității și eliminarea redundanței informaționale, vom forma componente principale ce reprezintă combinații liniare ale variabilelor inițiale.

### 3. Aplicații practice. Soluții implementate

#### 3.1. Preprocesarea datelor

Datele pe care le vom folosi în analizele construite sunt:

Nr. crt	Numele variabilei	Tipul variabilei	Descrierea variabilei
1	months_as_customer	int64	perioada de timp, exprimată în luni, în care persoana implicată în accident a fost clientul firmei de asigurări;
2	insured_age	int64	vârsta persoanei implicate în accident;
3	policy_number	int64	numărul poliței de asigurare;
4	policy_bind_year	int64	anul contractării poliței de asigurare;
5	policy_bind_month	int64	luna contractării poliței de asigurare;
6	policy_bind_day	int64	ziua contractării poliței de asigurare;
7	policy_deductable	int64	suma deductibilă a poliței de asigurare;
8	insured_cod	int64	codul poliței de asigurare;
9	insured_sex	object	genul persoanei implicată în accident și posesorul poliței de asigurare;
10	insured_education_level	object	nivelul de educație al posesorului poliței de asigurare;
11	incident_year	int64	anul în care a avut loc incidentul;
12	incident_month	int64	luna în care a avut loc incidentul;
13	incident_day	int64	ziua în care a avut loc incidentul;
14	incident_type	object	tipul accidentului;
15	incident_severity	object	severitatea accidentului;
16	authorities_contacted	object	autoritățile contactate după întâmplarea evenimentului;
17	incident_state	object	statul unde a avut loc incidentul;
18	incident_city	object	orașul unde a avut loc incidentul;
19	number_of_vehicles_involved	int64	numărul de vehicule implicate în accident;
20	bodily_injuries	int64	numărul de victime implicate în accident;
21	witnesses	int64	numărul de martori;
22	vehicle_price	int64	prețul autovehiculului asigurat;
23	injury_claim	int64	valoarea pagubei produse;
24	auto_make	object	producătorul autovehiculului asigurat;
25	auto_year	int64	anul de fabricație a autovehiculului asigurat;
26	fraud_reported	int64	stabilește dacă a fost înregistrată fraudă sau nu, este o variabilă binară, pentru care valoarea 1 indică prezența fraudei și 0 inexistența acesteia.

Principalul obiectiv al acestei lucrări este construirea unui model care să fie capabil să detecteze, în proporții cât mai mari, cazurile în care a fost sesizată existența unei fraude.

Informațiile pe baza cărora se construiește modelul au fost prezentate anterior, și au fost extrase de pe site-ul Kaagle sub forma unui fișier Excel, cu extensia csv.

Principalul pas în procesul de modelare a setului de date este reprezentat de instalarea pachetelor și importarea librărilor cele mai comun folosite în Data Science, dintre care amintim: pandas, numpy, matplotlib, seaborn, time, sklearn, xgboost, tabulate, etc.

Pentru importarea datelor am utilizat librăria pandas. Fișierul de date folosit este alcătuit din 26 de coloane și 1000 de observații.

```
The data contains 1000 observations.
```

	months_as_customer	insured_age	policy_number	policy_bind_year
0	328	48	521585	2014
1	228	42	342868	2006
2	134	29	687698	2000
3	256	41	227811	1990
4	228	44	367455	2014

Preprocesarea datelor este următoarea etapă ce trebuie realizată în vederea identificării unui model de date. În cadrul acestei etape vom vizualiza tipul de date al variabilelor, anume object sau int64, object indică prezența unor variabile categoriale (insured\_sex, insured\_education\_level, incident\_type, incident\_severity, authorities\_contacted, incident\_state, incident\_city și auto\_make), iar int 64 fiind specific variabilelor cantitative (restul variabilelor), aceste rezultate fiind obținute prin utilizarea sintaxei „df.types”. Pentru afișarea valorilor unice și identificarea astfel a existenței valorilor nule vom folosi comanda „unique()”.

Rezultatele funcției „unique()” au evidențiat prezența unor valori nule, NA, pe coloanele insured\_education\_level, vehicle\_price, auto\_year și number\_of\_vehicles\_involved. Pentru variabila vehicle\_price este înregistrată o singură valoare lipsă, corespunzătoare rândului 13, și am decis să completăm acest rând cu valoarea medie a prețului autovehiculelor, însumând valoarea de 52863 u.m.. Valorile lipsă pentru celelalte trei variabile se regăsesc pentru două dintre observații, rândurile 916 și 990, și de vreme ce coincid pentru toate variabilele am decis să le eliminăm, creând un data frame care conține 998 de observații, curățat de valori nule.

Ulterior, am verificat dacă în cadrul setului de date sunt existente dubluri, și conform outputului de mai jos acest lucru este infirmat.

```
#Verificam daca exista dubluri
print(len(df2.drop_duplicates())==len(df2))

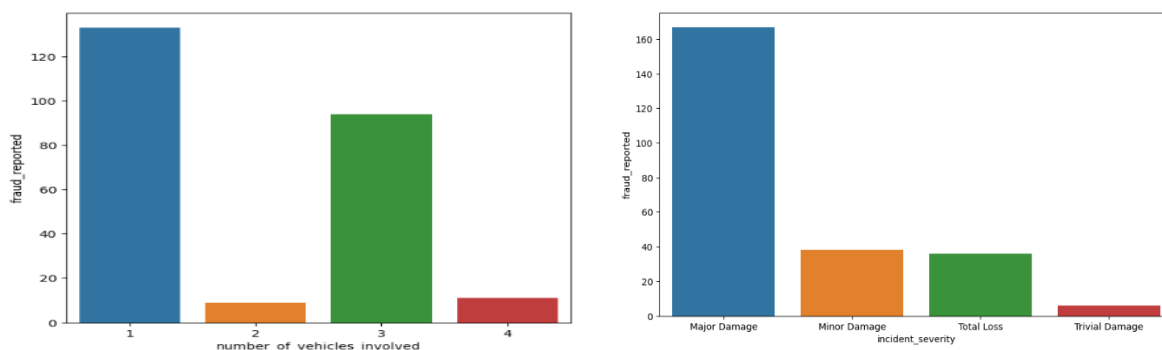
True
```

Odată stabilit conținutul final al setului de date ce stă la baza modelului am vizualizat câteva caracteristici primare ale variabilelor, numite statistici descriptive, folosind comanda „describe()”. Pentru variabilele cantitative sunt afișate numărul de observații, care în cazul nostru este 998, media, abaterea standard, cele trei cuartile, valorile minime și maxime. Statisticile descriptive specifice variabilelor categoriale sunt numărul de observații, de valori unice, modul și numărul de apariții pentru valoarea cea mai frecvent întâlnită.

Aplicăm testul Chi2 în vederea vizualizării relației dintre variabilele setului de date și variabila considerată dependentă, `fraud_reported`, ce urmează a fi previzionată. Variabilele pentru care rezultatele testului sunt relevante, p-value este mai mic decât 0.05, sunt: `fraud_reported`, `incident_severity`, `incident_type`, `authorities_contacted` și `incident_state`.

În continuare vom investiga legătura dintre variabila dependentă, `fraud_reported`, și variabilele independente prin intermediul afișărilor grafice. Am realizat o mulțime de grafice, grupând variabilele astfel: `fraud_reported` & `incident_severity`, `fraud_reported` & `incident_state` & `incident_city`, `fraud_reported` & `number_of_vehicles_involved` & `bodily_injuries` & `witnesses`, `fraud_reported` & `auto_make` & `auto_year`, `fraud_reported` & `insured_age`, `fraud_reported` & `policy_bind_year` & `policy_deductable` & `auto_year`, `fraud_reported` & `incident_type`, `fraud_reported` & `incident_severity` & `authorities_contacted`, `fraud_reported` & `insured_sex`, `insured_education_level` & `insured_age` și `fraud_reported` & `authorities_contacted` & `bodily_injuries` & `auto_make`, `fraud_reported`. Rezultatele complete pot fi vizualizate în cadrul anexei, urmând a prezenta doar cazurile ce prezintă însemnătate.

Observăm prin graficul ce prezintă distribuția cazurilor frauduloase în funcție de numărul de autovehicule implicate că în majoritatea accidentelor a fost implicat un singur autovehicul. În graficul alăturat acestuia este de notat faptul că accidentele „Major Damage” sunt frauduloase.



Ulterior etapei de înțelegere a datelor utilizate, vom iniția procesul de modelare. Pentru început se realizează formatarea setului de date prin eliminarea variabilei dependente și transformarea variabilelor categoriale în variabile numerice. `insured_sex` este transformată în variabilă binară, în vreme ce restul variabilelor calitative, care au mai multe variante de răspuns, vor fi transformate din categorii text în numere.

După efectuarea transformărilor asupra variabilelor categoriale am repetat testul Chi2 și, în plus, am calculat coeficientul de corelație Pearson. Rezultatul testului Chi2 nu se modifică semnificativ, pentru aceleași variabile înregistrându-se valori mai mici decât pragul de semnificație. Coeficientul Pearson evidențiază legături slabe și foarte slabe, cea mai puternică fiind legătura de intensitate medie (-0,41) a `fraud_reported` cu `incident_severity`.

### 3.2. Arbori de decizie, Random Forest, AdaBoost și XGBoost

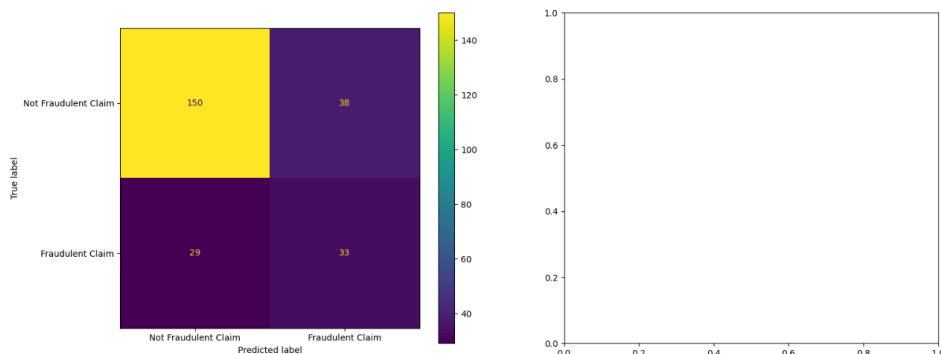
Înainte de modelarea efectivă, vom împărți setul de date în set de antrenare și de testare, folosind sintaxa „train\_test\_split”.

Primul model aplicat setului de date este arborele de decizie, menit să prezică valoarea variabilei fraud\_reported prin învățarea unor reguli deduse din caracteristicile datelor.

Gini stats	
Accuracy	0.72
balanced_accuracy	0.6650652
brier_score_loss	0.268
f1_score	0.4962061
recall_score	0.5322581
precision_score	0.4647887
roc_auc_score	0.6650652

Potrivit valorilor prezentate în tabelul alăturat arborele de decizie modelat reușește în proporție de 0.5322581 să identifice valorile True Positive, o valoare care nu satisface condițiile de semnificativitate și succes ale modelului. Vizualizând rezultatele grafice, mai jos afișate, afirmăm că previzionarea variabilei dependente folosind

arbori de decizie oferă rezultate de slabă performanță, acest model nereprezentând varianta optimă, și tocmai din această cauză vom utiliza și alte modele pentru previzionarea variabilei.



Modelul optim pentru previzionarea cazurilor frauduloase va fi identificat prin compararea statisticilor specifice celor patru modele de învățare supervizată, anume: Decision Trees, Random Forest, AdaBoost și XGBoost. Rezultatele aplicării modelelor de învățare asupra setului de date au fost sumarizate în tabelul de mai jos.

Model	Run_time	avg_accy	avg_recall	avg_precision	avg_f1	avg_matthew_corcoef	avg_roc_auc
Decision Trees	0.01	0.750667	0.473016	0.394921	0.476181	0.330237	0.657609
Random Forest	0.09	0.770667	0.292063	0.391703	0.358202	0.286802	0.609628
AdaBoost	0.05	0.774667	0.487302	0.437498	0.499539	0.37446	0.678056
XGBoost	0.15	0.794667	0.503175	0.445319	0.535499	0.428807	0.696665

În funcție de acuratețea modelului, coloana avg\_accy, modelul optim rezultă din folosirea XGBoost Classifier, oferind o acuratețe a previzionării de aproximativ 79.47%. Raportul dintre cazurilor pozitive identificate corect și numărul total de cazuri previzionate corect, avg\_recall,



indică prin valoarea sa că modelul optim este XGBoost. Coloana avg\_precision descrie numeric numărul de cazuri pozitive identificate corect raportate la numărul total de cazuri pozitive previzionate, și conform acestuia modelul ce redă cele mai bune rezultate este XGBoost. Valoarea coeficienților avg\_matthew\_corcoef și avg\_roc\_auc indică faptul că modelul XGBoost înregistrează cele mai bune performanțe, întrucât el contorizează valorile maxime. Așadar, conform celor șase statistici determinate modelul XGBoost este optim.

Pentru a confirma optimalitatea modelului XGBoost sau, dimpotrivă, pentru a stabili alt model optim vom aplica tehnica SMOTE (Synthetic Minority Oversampling Technique), tehnică care îmbunătățește performanța atunci când este suprașantionată clasa minoritară sau cea majoritară, sau ambele clase. Potrivit rezultatelor afișate în tabelul de mai jos, valorile statisticilor celor patru modele nu se modifică semnificativ, fiind confirmate astfel performanțele superioare ale acestui algoritm pentru previzionarea variabilei XGBoost.

Model	Run_time	avg_accy	avg_recall	avg_precision	avg_f1	avg_matthew_corcoef	avg_roc_auc
<b>Decision Trees</b>	0.0	0.765333	0.499206	0.537148	0.504818	0.362529	0.676114
<b>Random Forest</b>	0.07	0.777333	0.320635	0.639444	0.392484	0.32485	0.624109
<b>AdaBoost</b>	0.04	0.774667	0.487302	0.558003	0.499539	0.37446	0.678056
<b>XGBoost</b>	0.11	0.794667	0.503175	0.634087	0.535499	0.428807	0.696665

### 3.3. Analiza componentelor principale

Metoda ACP explică cea mai mare parte a variabilității datelor cu un număr cât mai mic de variabile, comparativ cu setul inițial de date, astfel, are la bază reducerea celor n variabile inițiale, la un număr de p variabile, numite componente principale. Pentru această analiză am folosit variabilele cantitative din setul de date prezentat inițial.

#### 3.3.1. Descrierea datelor

Statisticile descriptive se regăsesc în partea de preprocesare a datelor și nu diferă. Valorile coeficienților de asimetrie indică asimetrie la stânga pentru variabilele policy\_bind\_month, vehicle\_price și auto\_year, semnificând faptul că valorile sunt mai mari decât media, iar pentru restul variabilelor se înregistrează valori pozitive ale coeficientului skewness, asimetrie la dreapta, predominând valorile mici. Coeficientul de aplatizare furnizează informații referitoare la repartiția datelor, comparându-se cu valoarea de referință 3. Pentru toate variabilele analizate valoarea coeficientului kurtosis se situează sub valoarea de referință, seriile având repartiție platicurtică.

Un alt mod prin care poate fi analizată distribuția este prin intermediul histogramelor (hist) și a boxploturilor care ne oferă o interpretare vizuală a datelor prin indicarea numărului de

puncte de date care se află într-un interval de valori. Histogramele confirmă informațiile obținute anterior prin valorile coeficienților de asimetrie și aplatizare.

Vom analiza posibilele legături dintre variabile, apelând atât la coeficientul de corelație liniară Pearson, ce ne va caracteriza sensul și intensitatea legăturii dintre variabile, cât și la reprezentarea grafică a corelogramei pentru o vizualizare direct (corrplot). Corelograma indică prezența unor legături directe, de intensitate slabă și medie, remarcându-se însă două legături puternice între variabilele `insured_age&months_as_customer` și `injury_claim&vehicle_price`.

Metoda Analizei Componentelor Principale presupune reducerea dimensiunii setului de date. Primul pas este standardizarea datelor, deoarece avem indicatori ce se măsoară în unități de măsură diferite. Pentru a testa corectitudinea standardizării am calculat abaterea standard a fiecărei variabile din noul set de date și am observat că valorile înregistrate de acestea sunt egale cu 1. De asemenea, noile medii înregistrate de variabile sunt foarte apropiate de 0, ceea ce este normal și ne asigură că standardizarea a fost realizată cu succes.

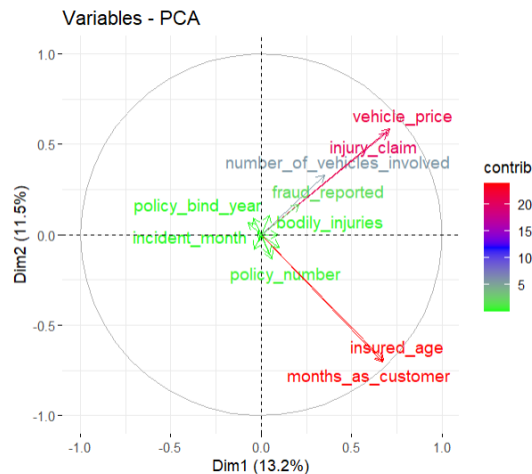
### **3.3.2. Extragerea componentelor principale**

Mai departe, vom forma componentele principale ce reprezintă combinații liniare ale variabilelor inițiale pentru a reuși să reducem dimensiunea și redundanța informațională. Pentru a stabili numărul de componente principale vom folosi criteriul lui Kaiser, criteriul procentului de acoperire și criteriul pantei. Conform tuturor celor trei criterii, amintite anterior, primele opt componente inițiale sunt considerate componente principale.

Având la dispoziție componentele principale, putem obține astfel și scorurile principale, care pot fi vizualizate în documentul unde este prezentat codul R.

Matricea factor folosește în clasificarea valorilor în componente, selectând pentru fiecare componentă valoarea cea mai mare luată în modul, considerând că face parte din cadrul componentei al cărei coloane aparține, astfel:

- Componenta 1: `months_as_customer`, `insured_age`, `number_of_vehicles_involved`, `vehicle_price` și `injury_claim`;
- Componenta 2: `insured_age`, `number_of_vehicles_involved`, `vehicle_price`, `injury_claim`;
- Componenta 3: `policy_number`, `policy_bind_month`, `incident_month`, `incident_day`;
- Componenta 4: `policy_bind_year`, `policy_bind_month`, `policy_deductable`, `witnesses`, `auto_year` și `fraud_reported`;
- Componenta 5: `policy_bind_month`, `policy_bind_day`, `bodily_injuries`;
- Componenta 6: `policy_bind_day`, `policy_deductable`, `incident_day`;
- Componenta 7: `incident_day` și `fraud_reported`;
- Componenta 8: `number_of_vehicles_involved`, `witnesses`, `auto_year`.



În graficul de mai sus este prezentată contribuția fiecărei variabile din setul de date la construcția componentelor principale, în cazul nostru, componente 1 și 2. Săgețile colorate în nuanțe de roșu închină semnifică o contribuție mai mare în alcătuirea componentelor principale, în vreme ce variabilele reprezentate de săgeți verzi contribuie într-o măsură mai mică, sub 5%.

#### 4. Concluzii

În concluzie, scopul inteligenței artificiale în domeniul fraudei din asigurări este de a facilita agenților umani să găsească și să investigheze daune și tranzacții frauduloase, mai eficient decât să verifice tone de daune într-un mod obositor și consumator de timp.

Esențialul oricărui studiu este redat de construirea unui model de clasificare competent cu ajutorul studiilor realizate anterior prin intermediul unor tehnici folosite în Data Mining și Machine Learning. Am revizuit articolele științifice existente, scopul nostru fiind să oferim informații valoroase pentru cercetători și practicieni în vederea utilizării eficiente tehnicilor de detectare a fraudei în domeniul asigurărilor auto.

Cel de-al treilea capitol al studiului este reprezentat de aplicarea tehnicilor amintite în punctul 2 al lucrării și măsurarea performanței metodelor de detectare a fraudei. Metodele de clasificare, utilizate în cadrul lucrării reprezintă un instrument statistic puternic în realizarea clasificărilor, predicțiilor și manipulării datelor. Implementând cele patru metode de clasificare am obținut diverse rezultate, iar prin compararea indicatorilor de performanță a rezultat că modelul optim este XGBoost. Însă cu toate acestea, valoarea coeficientului  $AVG\_roc\_auc$  de doar 0.6966 indică o clasificare destul de slabă, existând 69.66% șanse ca modelul să distingă cazurile unde se înregistrează fraudă și unde nu.

Prin realizarea Analizei Componentelor Principale am reușit să reducem dimensionalitatea setului de date, astfel că cele optsprezece variabile au fost concentrate în opt componente principale. Componentele principale creionate reușesc să preia 74.6% din procentul informațional total al setului de date inițial.

Așadar, în urma eforturilor noastre de a descoperi cel mai bun model în identificarea fraudelor din domeniul asigurărilor de autovehicule nu am reușit să obținem un model destul de satisfăcător care să furnizeze rezultate cu un grad ridicat de precizie care să reflecte în proporții de 100% realitatea. Una din cauzele pentru lipsa preciziei rezultatelor ar fi strâns legată de mulți alți indicatori care lipsesc din cadrul analizei noastre, de numărul redus de observații înregistrate și de corectitudinea informațiilor furnizate.